Unconscious Prejudice and Accountability Systems:

What Must Organizations Do to Prevent Discrimination?

Philip E. Tetlock and Gregory Mitchell[*]
University of California, Berkeley, and University of Virginia

(Not final draft; forthcoming in *Research in Organizational Behavior*)

[*] We appreciate the helpful comments of Hal Arkes, Hart Blanton, David Levine and participants at a Law and Economics Workshop at the University of Pennsylvania School of Law on an earlier version of this chapter.

Table of Contents

Abstract

The positions that experts take on whether organizations do enough to ensure equal opportunity hinge on the assumptions they make about the potency of prejudice. Prominent scholars have, however, recently challenged the conventional notion that anti-discrimination norms, backed by legal sanctions, can check unconscious prejudice. The strongest form of their argument is that it is impossible to achieve equal opportunity in any society with inequality of result—impossible because objective inequalities inevitably stamp into our minds subjective associations that inevitably contaminate personnel judgments that require the exercise of discretion. We note numerous objections to this thesis but concede that the debate over what, short of quotas, organizations can do to check prejudice is currently irresolvable given the paucity of data that clashing camps jointly treat as probative. Such debates can be advanced more efficiently via adversarial collaborations in which the debaters jointly design studies that they agree, ex ante, have the potential to induce both sides to change their minds.

I. Introduction

Discussions of what employers must do to prevent discrimination have a ritualistic flavor (Dipboye & Collela, 2005). There are the textbook recitations of best practices that Human Resource managers must implement (Armstrong, 2006; Gomez-Mejia, Balkin, and Cardy, 2007); the latest court decisions that the legal department must peruse for new prophylactics (Paetzold, 2005; Kaynard & Cook, 1999); the utilization and adverse-impact analyses that the Compliance branch of Human Resources must submit to regulatory agencies (Paetzold & Willborn, 2002; Robinson et al., 2005); the validation procedures that psychometricians must apply to all selection methods (Bersoff, 1988); and the diversity guidelines that trainers must follow to instill egalitarian values in the workforce (Arthur & Doverspike, 2005; Kalev et al., 2006).

The entire topic could be—but should not be—dismissed as hopelessly atheoretic. Lurking beneath the layers of bureaucratic posturing is a volatile mix of psychological and political disputes over the potency of prejudice and where society shoud set its thresholds for judging the adequacy of defenses against discrimination—and distinguishing sham from true compliance with civil-rights laws (Bielby, 2003).

We unpack these disputes here. Our starting point is the long-running controversy over the extent to which the nation has overcome its painful history of racial prejudice. The first section of this chapter draws on Festinger's (1957) theory of cognitive dissonance and Lakatos's (1970) philosophy-of-science analysis of research programs to capture the ideological rhythms of this epochal debate, which in its starkest form pits "statist interventionists" (who stress the lingering power of racism and the need for countervailing legal pressure to level the playing

field) against "market purists" (who stress the power of competition to eliminate irrational biases). We show how each side has built up a seemingly inexhaustible reserve of auxiliary hypotheses that allow it to explain away the favorite facts of the other side.

This stalemate has been rattled by new entrants who have imported reaction-time techniques from cognitive psychology to measure unconscious, or implicit, forms of prejudice that, they insist, self-report surveys fail to detect and that bias the judgments of most people most of the time (Greenwald, Banaji & Nosek, 2004; Rudman, 2004).[1] A subset of these researchers has forcefully argued—in law journals and in the national media—that work on unconscious prejudice has now reached a level of maturity that justifies extrapolation to the real world (Kang & Banaji, 2006; Greenwald & Krieger, 2006; Kang, 2005). A further subset of scholars has aggressively promoted "social-framework" applications of unconscious-prejudice doctrines in the form of expert testimony in live legal disputes (Bielby, 2003, 2005), thereby forging cross-level links between micro-cognitive challenges to the sincerity of tolerant attitiudes espoused by individuals and macro, neo-institutionalist challenges to the sincerity of organizational efforts to check prejudice (Edelman, 1992; Edelman & Suchman, 1999; Krawiec, 2003). This convergence yields a double-barreled indictment of American workplaces: managers are far

---

[1] In this literature, "unconscious" and "implicit" are synonymous, and we use as umbrella terms "implicit prejudice" and "unconscious prejudice" to refer to research on implicit attitudes and implicit stereotypic associations linked to particular groups. Furthermore, as we discuss later, what counts as discrimination in lab studies of prejudice is often very different from what may count as discrimination in courtrooms. Courts typically require disparate treatment on tangible and fungible variables, such as differential hiring and promotion between groups (White, 1998), whereas psychological studies often focus on far subtler "micro" disparities in behavioral reactions (sometimes as subtle as eyeblinking). What should constitute *illegal* discrimination ultimately involves, of course, a value judgment about what behaviors to regulate and condemn.

more biased than they admit and the companies for which they work are far less effective at correcting bias than they claim.

The second section of our chapter explores whether the new entrants have indeed tipped the debate decisively in favor of a strong interventionist position that challenges traditional merit-based hiring (Kang & Banaji, 2006) and holds managers accountable for numerical quotas in key phases of personnel decisionmaking (McGinley, 1997, 2005). We show that the new entrants have yet to make a compelling empirical case. Specifically, we identify three persisting unknowns concerning the power of unconscious prejudice to cause workplace discrimination: (a) unresolved construct-validity disputes over whether reaction-time differentials in implicit measures, such as the Implicit Association Test (IAT), tap into unconscious prejudice as opposed to alternative, more benign constructs such as sympathy for, or unfamiliarity with, minority groups (Arkes & Tetlock, 2004); (b) unresolved psychometric disputes over what it means to score as "unconsciously prejudiced" (Blanton & Jaccard, 2006) and what must be shown to justify company-wide attributions of prejudice as explanations of disparate outcomes (Fiedler et al, 2006; Mitchell & Tetlock, 2006); (c) unresolved external validity disputes that revolve round the many differences between lab studies of implicit prejudice and early 21st century workplaces and what must be shown to allay concerns about generalizability (Mitchell & Tetlock, 2006). The net result is a mis-match between empirical accomplishments and policy prescriptions, between how little researchers know about the power of organizational checks against prejudice *and* how confidently some scholars have dismissed a host of equal employment opportunity ("EEO") efforts as Potemkin-village cloaks for discrimination (Bisom-Rapp, 1999; Edelman & Suchman, 1999).

The third section identifies two obstacles to resolving disputes over unconscious prejudice: (a) the absence of shared standards of evidence for gauging the probative value of facts; (b) the absence of shared standards of proof for balancing false-positive (holding a non-discriminator liable) and false-negative (not holding a discriminator liable) classification errors.

The absence of shared standards of evidence makes it easy for each side to neutralize dissonant findings. Even if unconscious-prejudice researchers lost all the key arguments over the validity of implicit measures, they could invoke a litany of arguably legitimate escape clauses: abandoning the IAT for "better" measures of unconscious prejudice, disparaging criterion variables that tests fail to predict as insensitive to subtle hostility, or criticizing independent variables that squelch prejudice as loaded with demand characteristics. And if critics of unconscious prejudice lost the same arguments, they could invoke their own escape clauses: faulting lab experiments that find bias for failing to capture real-world variables that check bias and faulting field studies that find bias for ignoring objective differences in group performance.

The absence of shared standards of proof further complicates hypothesis testing, making it easy for each side—when they do agree on the facts—to disagree over the policy significance of those facts. Statist interventionists see more prejudice because they rely on expansive definitions that treat a vast range of everyday behavior as evidence of prejudice, even eye-blinking. And market-purists see less prejudice because they rely on restrictive definitions that require clearly disparate treatment of almost identically-situated employees. Either way, each side is free to inflate or deflate its estimates of prejudice in response to the same facts.

We conclude that the usual methods of scientific dispute resolution are unlikely to work as long as each side can: (a) dodge troublesome data by retreating into a protective shell of

auxiliary hypotheses; (b) count on the backing of a community of co-believers dedicated to defending the hard-core tenets of its worldview. Moreover, one need not accept the strongest indeterminacy form of our argument to agree that reliance on the usual methods has proven to be too painfully slow a process for courts, regulators, and legislatures that need reasonably clear and de-politicized answers—now—to empirical questions about the drivers of inequalities (Monahan & Walker, 1991; Goldsmith & Vermuele, 2002).

The fourth section proposes an unusual approach to breaking the impasse—adversarial collaboration—that calls on the clashing camps to: (a) reflect on the dynamics of the debate; (b) identify pivotal empirical points on which they can either agree or agree to disagree; (c) specify research designs on which they can base Bayesian bets about the potency of unconscious prejudice; (d) commit themselves to threshold-of-proof statements that specify how low or high unconscious-prejudice effect sizes would have to fall or rise to induce them to change not just their psychological assessmentsl but also their policy stands.

In closing, there is room for disagreement over how to achieve the epistemic goals of adversarial collaboration. But there is no room for compromise in our philosophy of science over the goals. Any research program that rejects the Socratic soul-searching prescribed by adversarial collaboration fails the classic litmus tests for science (Merton, 1973, 1987)—and should fail the Supreme Court's *Daubert* tests for admitting scientific testimony in federal courts (Faigman & Monahan, 2005). This latter point is no fine point: social science experts have woven unconscious-prejudice arguments into their testimony in dozens of class-action cases over the last decade (Bielby, 2003), and some courts have accepted these arguments despite the many unresolved questions (Dukes v. Wal-Mart, 2006). For better or for worse, our science is

literally, not just metaphorically, on trial.[2]

II. Is prejudice as prevalent and potent as ever?

We view ideological debates over inequality (especially racial inequality) through two prisms: the psychological prism of cognitive-dissonance theory (which highlights how hard it is for individuals to admit mistakes) and the epistemological prism of Lakatos's (1970) analysis of the logic of scientific research programs (which highlights how hard it is even for ostensibly self-correcting scientific communities to admit mistakes). We also believe that, on topics as divisive as racial inequality, it is impossible to insulate scientific from political disputes (Tetlock, 1994). Partisans rarely resist the temptation to conscript convenient findings into the service of promoting favorite policy initiatives (Proctor, 1991; Suedfeld & Tetlock, 1991). We describe two polar opposite sets of causal assumptions and policy prescriptions that have dominated elite debates over racial inequality: the "statist interventionist" and "market purist" ideal types.

Orbiting, with varying tightness, round the statist-interventionist pole, we find a mix of

---

[2] We have explained elsewhere (Mitchell & Tetlock, 2006) why implicit-prejudice should *not* be admitted as a general-causation framework in class-action litigation: there is no evidence that the IAT reliably predicts class-wide discrimination on tangible and fungible outcomes in any setting, much less real workplaces with individuating information far richer, and accountability pressures far stronger, than those in any existing implicit-prejudice study. Others disagree. Blasi and Jost (2006, p. 1164) declare that "much of the science of implicit social cognition is now established well beyond the standard Daubert v. Merrell Dow Pharm., Inc. would require for its admission at trial" (Blasi & Jost, 2006, p. 1164). We worry about the potential of such open-ended proclamations to damage the long-term credibility of social science. And although we recognize our candor will not win converts to adversarial collaboration, it would be a bigger mistake to disguise the depth of the disagreements that now fissure within as well as across disciplines. One should engage in adversarial collaboration not because one finds one's adversaries, or their positions, congenial but because one recognizes: (a) the credibility of one's science hinges on the science's capacity to transcend ideology; (b) the usual methods of conducting the debate cannot break through the complex defenses each side can use to neutralize dissonant evidence.

scholars and policy advocates who are convinced that prejudice is so pervasive that it is impossible for workers who fall into protected categories (especially African-Americans) to get a fair shake in unregulated labor markets (e.g., Spann, 2005; Wang, 2006). The most ardent proponents urge the adoption of affirmative actions plans (that sunset only when psychologists no longer detect implicit biases) as the only equitable remedy in an otherwise ineradicably racist society (Kang & Banaji, 2006).

Orbiting, with varying tightness, round the market-purist pole, we find a mix of scholars and policy advocates who are convinced that market mechanisms are up to identifying and punishing employers foolish enough to hire or fire on any basis other than potential to increase productivity (e.g., Epstein, 1992; Smith & Welch, 1986, 1989; Sowell, 1994). The most ardent proponents argue that even the landmark civil-rights legislation of the 1960s and ensuing enforcement efforts had no appreciable effect on racial inequality, attributing whatever progress the nation has made toward equality to self-equilibrating labor-market forces of supply and demand (for critical reviews, see Donohue & Heckman, 1991; Heckman et al., 2000). Of course, most scholars stake out more centrist positions that acknowledge the beneficial effects of certain governmental interventions but shy away from mandating equality of result, either because they see competitive markets as powerful debiasing mechanisms, they balk at strong variants of the inevitability-of-prejudice thesis, or they fear the unintended consequences of affirmative action.

Table 1

Core Propositions of Polar Positions in American Racism Debate

	Statist Interventionist	Market Purist
Psychological (1)	Socioeconomic hierarchies imprint corresponding mental hierarchies with self-fulfilling and system-justifying behavioral	Socioeconomic hierarchies motivate less advantaged groups to work hard and invest in useful skills. It also gives these groups a labor-

	implications.	cost advantage.that destabilizes the hierarchy.
Psychological (2)	Unconscious bias pervades intergroup relations. As long as there is an element of subjectivity in personnel judgments, bias will influence judgments (true even when judges are accountable for implementing egalitarian norms).	Pervasive unconscious bias, while a possibility, remains unproven; explicit prejudice has receded in response to egalitarian norms and accountability pressures, and implicit biases will likely follow the same track.
Economic (3)	Markets cannot overcome unconscious bias in the short- or long-term.	Competitive markets can overcome bias in the short-term; less competitive ones can do so in the long-term.
Sociological-Psychological (4)	Past discrimination and continuing structural barriers to education and, health care perpetuate inequalities (a process facilitated by the fundamental attribution error).	Human capital often covaries with group membership; different subcultures encourage different tastes and skill sets; and governments will misallocate resources when they ignore these facts.
Governmental (5)	Strong government intervention, including numerical quotas, is necessary to check unconscious bias.	Government intervention disrupts market forces that would otherwise overcome group-based discrimination.
Methodological (6)	The detection of unconscious bias and its and discriminatory concomitants requires sensitive implicit measures of prejudice and controlled experiments.	Experiments cannot simulate real market conditions; only sophisticated econometric and careful field studies can assess true labor market discrimination.
Ethical (7)	Absent state intervention, the birth lottery largely determines lot in life and no lottery can provide a moral justification for social inequalities.	Regardless of the fairness of starting positions, a minimal state is the best long-term guarantee of individual freedom and welfare maximization.
Ethical (8)	It is unfair to allow advantaged groups to benefit from past discrimination; society should adopt remedies that right past wrongs.	It is unfair to impose the remedial costs of past discrimination on current employers and citizens who have not been found guilty of discrimination.

Six of the core statist-interventionist propositions are primarily empirical: (1) the psychological contention that prejudice so suffuses American life that it is impossible for managers to make color-blind decisions in any discretionary phase of personnel decision making in which protected-category status is known or can be inferred (e.g., Bertrand et al., 2005; Chugh, 2004); (2) the sociological contention that, even if prejudice plummeted to zero, structural barriers—varying access to schooling, social networks and other opportunities—would still generate big inequalities (Plous, 2003); (3) the social-psychological contention that the usual

organizational process checks on bias—reining in managerial subjectivity, holding managers accountable to EEO process norms, diversity training—are far too weak or place too much burden on victims to stop the juggernaut of psychological and sociological drivers of inequality (Bielby, 2003; Edelman, 1992; Kaiser & Miller, 2001; Krawiec, 2003); (4) the contention from economic and historical research that markets are not nearly as efficient at overcoming either individual irrationality or structural barriers as neoclassical economists suppose (Coleman, 2004; Lee, 2004); (5) the contention grounded in research on stereotype threat that conservatives exaggerate objective variation in human capital (Nisbett, 2005; Suzuki & Aronson, 2005) and that the acts of choosing how to define race and measure human capital are inevitably value laden and tainted by our cultural worldviews (Sternberg, 2005); (6) the contention from historical and sociological research that, insofar as real differences in human capital do exist across groups, we should attribute those differences to past and current prejudice and that intellectual resistance to this obvious point suggests the operation of either cognitive biases (e.g., the fundamental attribution error; Nisbett & Ross, 1980) or motivational biases (e.g., the Just-World effect of "blaming the victim," Lerner & Miller, 1978).

This network of mutually reinforcing propositons defines the Lakatosian hard-core of the statist-interventionist research program on inequality: a network that bestows thematic unity on a sprawling interdisciplinary body of work and that sympathetic researchers feel obliged to defend when key tenets come under assault. Indeed, true believers need never surrender. They can virtually always spin off one or another auxiliary hypothesis to defend the hard-core. When skeptics challenge Proposition # 1 by noting that prejudice has long been on the wane in representative national surveys, the defender can stipulate that such surveys tap into only the

most superficial forms of prejudice and that we need to crank up the microscope in pursuit of symbolic or modern racism (Sears & Henry, 2005), or aversive racism (Dovidio, 2001) or unconscious prejudice (Greenwald et al., 1998). When skeptics challenge Proposition # 2 by tracing inequalities in employment outcomes in specific firms to variation in human capital in education and skills, the defender can stipulate that the skill qualifications are arbitrary and subjective barriers to entry designed to preserve an inequitable status quo (Hart, 2005) or, failing that, the variation in human capital is itself a product of racism-tinged structural barriers to schooling and training (Brown et al., 2003; Smith, 1995). Most disconcerting is Proposition # 6: the defender has the back-pocket option of invoking arguments that verge on the ad hominem—to imply that only those sympathetic to unconscious prejudice would resist aggressive explanatory extensions of those concepts to the real world (Kang & Banaji, 2006; Sears, 2004; Banaji et al., 2004) or would look aggressively for alternative explanations that are tantamount to blaming the victim (Lerner and Lerner, 1978).[3]

From the standpoint of Popper's (1959) strict falsificationism, such lengthy recitals of auxiliary hypotheses illustrate how easily the the "Lakatosian negative heuristic"—the directive to defend the hard core—can spin out of control. But what look like fatal shortcomings to

---

[3] Meehl (1990) noted the need to rethink old standards for labeling arguments ad hominem in post-positivist philosophies of science that stress the power of psycho-social processes to distort hypothesis testing. Indeed, from a post-positivist perspective, one violates scientific norms by adopting an ostrich stance toward such distortions. Arguments that implicit-bias skeptics are animated by system-justification motives or that implicit-bias proponents are animated by a moral-cleansing quest to eliminate racism from the innermost recesses of the American psyche fit the traditional template for ad hominem, but such arguments can also be viewed as testable hypotheses. Social science has absorbed many post-positivist ideas, but it has yet to digest Meehl's insight, and we suspect the insight will prove indigestible.

skeptical outsiders are problem-solving assets (positive heuristics) to insiders who insist that research on racial attitudes has already shown that many Americans who endorse racial equality in principle should be classified as prejudiced using subtler metrics. Insiders object when skeptics dismiss searches for subtle prejudice as mere rationalizations of hard-core commitments—arguing that researchers have discovered new, not just reinterpreted old, phenomena (Banaji et al., 2004; Greenwald, et al., 2006). Similarly, insiders insist that recent research on stereotype threat provides empirical substance to Proposition # 5 by showing how standardized tests underestimate human capital among disadvantaged groups by failing to take into account the power of "stereotype threat" to depress scores (Suzuki & Aronson, 2005). And insiders argue that research on Just World theory (Lerner & Lerner, 1978; Hafer & Bègue, 2005), the fundamental attribution error (Ross, 1977), and the ultimate attribution error (Pettigrew, 1979) empirically substantiates the contention that people—majority and minority groups alike—are insensitive to the impact of structural barriers on success rates, thereby protecting Proposition # 6 against the objection that it just licenses name-calling.

As Lakatos (1970) appreciated, no clear line separates negative- from positive-heuristic science (Suppe, 1977). Falsificationism is not all or nothing: science rests on judgment calls that are susceptible to all the distortions of motivated reasoning (Kunda, 1999).

Shifting toward the market-purist pole, we discover an almost mirror-image set of causal beliefs, sometimes even based on the same data used by statist interventionists: (1) the conviction that prejudice and nasty stereotypes are truly on the wane (e.g., Epstein, 1992; Justice O'Connor's majority opinion in Grutter v. Bolllinger, 2003)); (2) the conviction that it is unlikely that biases against working with minorities or false stereotypes are causing large numbers of

employers to make suboptimal decisions because competitive markets are quick to correct such mistakes (unbiased employers will recruit and retain superior talent) (Becker, 1957; though as Heckman, 1998, notes, Becker's models of employee and customer discrimination do not imply disappearance of discrimination even in the long-run); (3) the conviction that human beings are remarkably resilient and can bounce back from historical oppression and learn to circumvent structural barriers, and that government remedies aimed at these barriers may paradoxically undercut this resiliency and motivation to achieve (S. Steele, 2006); (4) the conviction that government programs that press employers to lower standards to achieve numerical goals will only hurt the groups that the programs are designed to help (hurt by encouraging attitudes of entitlement grounded in an ideology of victimology—and by reinforcing the public perception that some groups just cannot compete with the rest of the population) (Coate & Loury, 1993); (5) the conviction that human capital is rarely evenly distributed across groups and that such variation is most parsimoniously explained by deeply rooted cultural practices and preferences (Sowell, 1994), such as minority groups penalizing members for "acting White" (Fryer, 2006), and, under the rule of "de gustibus non est disputandem," it is inappropriate to second-guess these preferences (individually or collectively, people make choices—and should live with the consequences); (6) the conviction that, insofar as intergroup inequalities of income do arise in specific workplaces, these inequalities are most parsimoniously attributed to intergroup variation in human capital (Smith & Welch, 1989).

This combination of mutually reinforcing causal beliefs defines the hard-core of the laisser-faire research program on inequality. Market purists can buffer their hard core from disconfirmation by a protective belt of auxiliary hypotheses as thick as that of the counterparts

on the left. When market purists fail in their initial multiple regressions to identify human-capital control variables that "explain away" intergroup inequalities, they retain the option of cranking up the microscope and questioning the reliability and validity of existing control variables and, failing that, looking harder for new ones that tap into dimensions of skill and performance that influence productivity but have been slighted in earlier analyses. And market purists can point to examples in which the determined pursuit of better measured human-capital variables has served not just to patch up the hard core but to advance knowledge (e.g., Carneiro et al. 2005).

Why ground our analysis of unconscious prejudice in these interminable skirmishes between statist interventionists and market purists? Our answer is two-fold. First, unconscious-prejudice researchers are the latest entrants into the long-standing debate over the causes of lingering inequalities. They themselves claim their findings tip the scales dramatically in favor of the interventionists—and demand rethinking the foundations of American employment laws and employment practices that will ultimately govern the work lives of all Americans (Kang & Banaji, 2006; Greenwald & Krieger, 2006; Potier, 2004). Second, we find Lakatosian criteria useful for evaluating the unconscious-prejudice research program—for weighing proponents' claims that such research has already demonstrated its positive-heuristic value against detractors' claims that the same research has already revealed itself to be a negative-heuristic exercise. Indeed, we have only one major reservation about our framework: it is too politically deterministic. As skeptics ourselves, we know that one need not be a market purist to doubt key claims of implicit bias theorists—and, notwithstanding the advocacy roles taken by some leading implicit bias researchers, we suspect that at least some adherents to implicit bias theory are not statist interventionists.

Ideological framing to the side, the hard-core assumption at scientific stake in this article is Proposition # 1: the claim that prejudice "thoroughly suffuses American life." On its face, this claim would appear to have been undermined by five decades of representative-sample surveys documenting greater racial tolerance among Americans. Since the landmark civil rights legislation of 1964, overt White hostility toward African-Americans has been in decline and by the turn of the century, it had reached historic lows (Schuman, Steeh, Bobo, & Kyrsan, 1997; Sniderman & Carmines, 1997; Dovidio, 2001). Whereas Americans were once deeply divided over whether Blacks and Whites should be allowed to drink from the same fountains, sleep in the same hotel rooms, attend the same schools, live in the same neighborhoods, or intermarry, there is now near unanimity that all forms of disparate treatment are unacceptable (Quillian, 2006). And whereas either pluralities or majorities of Americans once endorsed sweeping stereotypes of African-Americans as violent or dumb or lazy or hyper-sexual, strong majorities now deem such sentiments unacceptable (Sniderman et al, 1991). These profound shifts in public sentiment have seeped into employment practices: early 21st century American managers risk censure or worse if they endorse once commonplace stereotypes or if they tell once run-of-the-mill jokes targeting women or minorities (Ash v. Tyson Foods, 2006; Russell v. City of Kansas City, 2005).

Defenders of the statist-interventionist program have an array of options for defanging the most dissonant implication of the survey results: the possibility that prejudice is fading into an historical curiosity even though objective inequalites are persisting. And dissonance theory—as well as its modern incarnations depicting belief systems as self-equilibrating and coherence-seeking (Vallacher et al., 2002)—tell us which options will prove most attractive: those that reduce the most dissonance at the least cost in compromising other beliefs. For statist

interventionists, that means trivializing survey data as "empty talk" that masks a shadow universe of unconscious prejudice inaccessible to pollsters. For market purists, we shall soon see it means trivializing lab findings as "hopelessly artificial."

Viewed in this light, unconscious-prejudice research is the latest in a long line of efforts to develop ever subtler techniques for tapping into ever better camouflaged mutations of prejudice.[4] Its precursors are symbolic and modern racism theory, which implied that, although few Americans now endorse crude stereotypes or de jure segregation, they do endorse covertly racist sentiments, such as opposition to busing or affirmative action (Sears & Henry, 2005). And some researchers add that this resistance is far greater than one would expect if old-fashioned prejudice were the sole driver (Dovidio, 2001).

This first wave of efforts suffered from methodological shortcomings that unconscious prejudice researchers have consciously sought to avoid. Critiques of this first wave came from diverse directions. One critique—from political science—noted that self-report scales of covert

---

[4] Our argument does not require that unconscious-prejudice researchers see their studies as conceptual pawns in the American racism debate—no more than system-justification theorists require that high scoreres on their scales see themselves as tools of capitalist oppression (Blasi & Jost, 2006; Jost et al., 2004) and no more than psychologists require that experimental subjects possess introspective access to their cognitive processes (Greenwald, 1992). Indeed, it matters naught if all unconscious-prejudice researchers sincerely see their work as apolitical extensions of micro-cognitive studies of associative memory. Intentionality is not at issue. What matters is the power of unarticulated moral assumptions to shape research agenda. Haidt's (2007) five-component model of moral intuition offers a useful framework for understanding such processes. Imagine a research program that, instead of targeting implicit racial animus, targeted implicit animus toward groups at the top of the annoyance list for conservatives: implicit traitors (who reveal too positive associations to flag burners or too negative associations to anti-terrorism policies) or implicit Marxists (who score as unconsciously soft on communism). We urge both proponents and skeptics to adopt the following metric of politicization: would they apply the same standards of proof to implicit measures across all possible targets of assessment?

racism often equated conservatism with racism by operational fiat at the scale-item level, a practice that makes it virtually impossible to eliminate conservative beliefs and values as alternative explanations (Sniderman & Tetlock, 1986). Another critique—from cognitive psychology—noted that these indirect measures still required public reports of attitudes with racial overtones, leading these critics to wonder whether such measures could accurately tap into underlying racial hostilities—either because of impression management or lack of conscious access to the thought processes that may drive intergroup hostility (Brauer et al., 2000).

From these critical perspectives, unconscious-prejudice research has three big advantages over the first-wave efforts: (1) it does not rest on tendentious classifications of conservative opinions as racist; (2) it reduces the problems of social-desirability contamination (Nosek et al., 2006; but see Czellar, 2006; Fiedler et al., 2006) and unreliable introspection; (3) it uses tools of proven usefulness in cognitive psychology (Fazio & Olson, 2003). For these reasons, we agree with unconscious-prejudice researchers that their approach is a better long-term bet for expanding the explanatory reach of the statist-interventionist program, not just the negative-heuristic function of buffering the hard core from dissonant data.[5]

The most popular methods of measuring unconscious prejudice are affective priming and implicit association tests, with the latter now the dominant method (Hofmann et al., 2005). Both methods purport to tap into unconscious prejudice by examining associations between groups and valenced terms and purport to tap into unconscious stereotypes by examining associations

---

[5] Of course, "better bet" does not necessarily mean "good bet"—a point to which we return.

between groups and traditionally stereotypic terms (Fazio and Olson, 2003).[6]

In the affective-priming paradigm, subjects are exposed to an attitude object (the "prime," say, pictures of faces) and then to evaluative adjectives (the "target," say, words such as "nasty" or "kind") that they must categorize as good or bad as quickly as possible. Researchers use the time taken to perform the categorization ("reaction time" or "latency") to gauge the strength of association between prime and target. For instance, slower reaction times in labeling positive adjectives after an African American face-prime are taken as evidence of implicit negative associations toward African Americans—and often as evidence of negative attitudes; so too are faster reaction times in labeling negative traits.

In the Implicit Association Test, or "IAT," subjects must pair target concepts (e.g., African-American, European-American) with evaluatively-charged terms (e.g., Good, Bad, Brilliant, Stupid) and then place the stimuli (e.g., pictures of African-American and European-American faces) into one of the competing categories as quickly as possible. As with the priming approach, response latencies gauge strength of association between these paired concepts. The guiding idea is that people can respond faster to "compatible" associations (e.g., White + Good, Black + Bad) than to "incompatible" ones (e.g., Black + Bad, White + Good). In the race IAT,

---

[6] Both methods are ambitiously reductionist efforts to isolate the foundational building blocks of prejudice. We do not object to reductionism per se. The greatest successes of the physical and biological sciences are grounded in reductionism. We object only to the specific reductionist assumption that millisecond reaction-time differentials will make the same foundational contributions to knowledge of prejudice as fruit flies have to genetics. Successful reductionism requires identifying simple systems governed by laws that scale up, with minimal distortions, to complex systems. We shall see that there are good reasons for suspecting that reaction-time measures of implicit prejudice are contaminated by large amounts of measurement error and that "laws" so identified are highly sensitive to minor contextual complications .

subjects view a set of stimuli with racial connotations (e.g., faces or names associated with different races) that they must categorize by race as quickly as possible. In a separate block of trials, subjects view a set of words with positive or negative connotations (e.g., sickness, death, freedom, paradise) that they must categorize according to valence (e.g., as "pleasant" or "unpleasant"). Subjects also participate in trials in which they must pair racial and attitudinal stimuli into composite categories. In the "compatible" trials, they sort stimuli into their "natural" pairings (White + pleasant, Black + unpleasant) and in the "incompatible" trials, they sort stimuli into the reversed pairings. Faster sorting in the compatible trials is taken to be evidence of stronger implicit associations between group categories and evaluative labels.

Unconscious bias reseachers have made four categories of empirical claims that, if true, would reinforce the hard core of the statist-interventionist research program:

Claim # 1—the pervasiveness of implicit prejudice. A warehouse of data now shows that most people pair advantaged groups (White, wealthy, healthy, young, heterosexual, Christian, American) more quickly with positive adjectives and pair disadvantaged groups (non-White, poor, overweight, aged, homosexual, non-Christian, foreign) more quickly with negative adjectives. Indeed, even disadvantaged groups "prefer" dominant out-groups. Some scholars believe that such data signal alarming levels of implicit prejudice, in stark contrast to the declines in explicit prejudice in opinion surveys (e.g., Bagenstos, 2006; Salgado, 2006; Wang, 2006);

Claim # 2—the insidiousness of implicit prejudice. Scores on implicit and explicit measures of prejudice often diverge, suggesting that many people who sincerely view themselves as egalitarians nevertheless harbor unconscious prejudices. This disjunction between conscious and unconscious attitudes often elicits the most discomfort and disbelief among subjects, causing

them to question the validity of the IAT. And this disjunction makes pervasive implicit prejudice so alarming from a legal perspective, because implicit prejudice is not just hidden from public view but from private introspection as well. If implicit prejudice operates outside awareness and beyond our deliberate control, then we can no longer use direct or circumstantial evidence of intentional conduct as a good indicator of prejudice, but we must instead delve into the inner workings of the mind using IAT-type measures;

Claim # 3—the uncontrollability of implicit prejudice. Banaji (2004) has argued that "[i]mplicit attitudes . . . are disengaged from conscious thought and are unlikely to shift in response to the willful call for change (p. 135)." Such rigidity suggests that mere exhortations to managers to be fair-minded will not be enough to check unconscious biases (Hart, 2005);[7]

Claim # 4—the potency of implicit prejudice. A meta-analysis of IAT studies claims that IAT scores predict discriminatory conduct (Poehlman et al., 2006)—potentially transporting us from the realm of hypothetical torts into the realm of legally actionable employment practices.

If these four claims withstand scrutiny—and hold up in workplaces that have institutionalized hitherto legally defensible checks against prejudice—statist interventionists would have a strong case for reversing the burden of proof—from the common-sense default stance that, absent evidence to the contrary, self-reports of tolerance should be viewed as valid, to the new stance that, absent evidence to the contrary, reaction-time indicators of animus should

---

[7] The question here is one of intentional control over implicit prejudice, not contextual variation in the expression of implicit prejudice. Implicit-prejudice investigators recognize that the intensity of measured implicit biases varies by context (e.g., Banaji, 2004)—which, we believe, renders the research program vulnerable to several external validity challenges discussed later.

be viewed as valid. Statist interventionists would also have a stronger argument in courts, regulatory agencies and legislatures that, absent strict accountability pressures to minimize subjectivity and to satisfy numerical goals in personnel decisions, the default assumption should be that unconscious biases are a plausible cause whenever adverse impact materializes.

But extraordinary claims require extraordinary evidence—and we should be all the warier when the claims dovetail conveniently with the mostly statist-interventionist sympathies of the research community (Redding, 2001). We shall show in the next section that the evidence does not remotely begin to tip the scales in the American racism debate in favor of radical interventionism. However, the empirical issues in the third section are not settled science, and the value issues at stake in the fourth section (which deals with translating empirical findings into policy prescriptions) need to be formally addressed. In brief, no one is yet well-positioned to claim victory–which is why we devote the final section of this chapter to working out the conceptual preconditions that must be satisfied to move the debate forward.

III. Challenges to the unconscious prejudice research program

Debates over unconscious prejudice cover the full span of types of scientific validity: internal, external, statistical, and construct validity (Cook & Campbell, 1979). We focus on three disputes on which will turn the scientific fate of the unconscious-prejudice research program:

(1) Construct-validity debates: do we know enough about the causes and correlates of scores on implicit measures of prejudice to grant that these measures are tapping into the target construct of " unconscious prejudice" rather than into alternative constructs?;

(2) Psychometric debates: assuming that implicit measures of prejudice do tap into their target construct to some degree, can the statistical links sustain inferences of the form “people

who display differential reaction times of a given magnitude are some practically and statistically significant percentage more likely to discriminate against protected groups"?;

(3) External-validity debates: do the claimed links between implicit measures of prejudice and criterion variables hold up in the face of the institutionalized lines of defense against discrimination that many organizations have created in the post-civil-rights-era?

For each question, our answer is "probably not." It is critical though to note that, unless one opts for a circular positivism (implicit prejudice is whatever implicit measures of prejudice measure), our answer—indeed anyone's answer—must rest on a certain theoretical stance toward unconscious prejudice and an accompanying set of political-value thresholds for labeling mental states as prejudice and behavior as discriminatory.

A. Construct-validity debates: Do implicit measures of prejudice tap "prejudice"?

Psychological constructs have a shaky ontological status; not observable, we must infer preference from choice, intelligence from IQ tests, and, for the IAT, implicit prejudice from reaction times on speeded binary classification tasks. Since Cronbach and Meehl (1955), the standard solution for closing the inferential gap has been construct validation: we can conclude that a test taps into an unobservable construct to the degree that construct, and only that construct, can explain the correlations between the test and other indicators. Psychological constructs are thus defined by their location in a nomological network (Westen and Rosenthal, 2003). Unfortunately, there is no explicit psychometric model of implicit prejudice that maps the universe of related constructs (Borsboom, 2006). And this omission is especially unfortunate in the hyper-charged American racism debate: it creates ample room for partisans to spin ambiguous results as showing the glass as either mostly full or nearly empty. With that caveat,

we divide construct-validity controversies into two: those focused on the cognitive-affective causes of IAT scores and those focused on the behavioral consequences of IAT scores.

(1) Disputes over causation: Are IAT scores distorted by irrelevant processes?

Prejudice is but one of many possible explanations for differential reaction times on the IAT. The list of potential confounds is long—and the debate consequential, for media interest hinges on whether the test is packaged as a measure of implicit prejudice as opposed to familiarity or sympathy or test anxiety or tacit awareness of depressing covariations in society at large. The power to label persons or organizations as prejudiced is nontrivial; it is the power to de-legitimize the opinions or practices of those who, for whatever reason, trigger one's ire.[8]

(i) Familiarity, not Antipathy. This counter-interpretation maintains that IAT effects do not carry the morally loaded meaning assigned by IAT proponents because the effects can be better explained by confounding variation in the salience of test stimuli. The argument, backed by several experiments (Rothermund & Wentrua, 2004; Rothermund et al, 2005), can be condensed into three claims: Minority-group stimuli are rarer, thus more distinctive and prone to perceptually "pop out" than are majority-group stimuli; bad adjectives are more threatening and thus more salient than good adjectives; it is easier to respond to paired stimuli that are similar in salience than it is to mixed-salience pairs (Kinoshita & Peek-O'Leary, 2005);

(ii) Fear of Being Labeled a Bigot, not Bigotry. Implicit measures of prejudice are subtler than explicit measures. But as anyone who has taken the race-IAT appreciates, it soon

---

[8] Statist interventionists note that the power to label cuts the opposite way, with firms that show a surface commitment to diversity supposedly receiving legal deference (Edelman et al., 2001).

becomes obvious that the researchers are curious about how one thinks about race (De Houwer, 2005). Frantz et al. (2004) have shown that the more threatened participants felt by the IAT—the fear of being labeled racists—the "worse" their IAT scores. We would add that the worse one fears one has done on the IAT, the more self-conscious one is likely to be in follow-up interracial encounters—an alternative explanation for why the IAT sometimes predicts awkwardness in such encounters (see section III.A.2).

(iii) Sympathy, not Antipathy. In this view, the IAT measures associations, but the "negative" associations are rooted more in compassion or guilt than in hostility. Arkes and Tetlock (2004) captured this idea in their parable of the two Jesses. One Jesse (Jackson)is a statist interventionist who declares discrimination to be an ongoing, not just historical problem—many Whites still resent African-Americans. The other Jesse (Helms) is a market purist who believes that the big causes of racial inequality in America are now internal to the African-American community, especially the erosion of responsibility in inner cities.

Some day, someone may offer a compelling reason to expect these two individuals to exhibit different reaction times on the IAT. But no one has yet. And promoters of the IAT have presented no evidence that their reaction-time indices can distinguish qualitatively different cognitive-emotional states, such as frustration, sorrow and anger, rooted in competing appraisals of the political scene. Indeed, the evidence is to the contrary. Uhlmann et al (2005) experimentally created fictitious groups, portrayed some of them as victims of oppression similar to that suffered by African-Americans, and found that subjects had more "negative" IAT scores toward these groups—groups that elicited sympathy, not contempt, from observers.

(iv) Cognitive Dexterity, not Antipathy. This interpretation highlights a curious

implication of the IAT scoring procedure: rapid responding on compatible trials contributes as much to one's score as sluggish responding on incompatible trials. Recognizing the potential confound, IAT researchers have recommended data-transformations to minimize the influence of individual differences in processing speed. These data transformations cannot however eliminate the possibility that an individual-difference variable other than prejudice influences performance on all speeded-classification tasks that require responding flexibly, left or right, to shifting combinations of cues. Research on cognitive styles suggests that such individual differences do exist and are grounded in variation in fluid intelligence (Fiedler et al., 2006). And Mierke and Klauer (2003) have found support for this notion by experimentally inducing associations between target stimuli (blue versus red) and attribute stimuli (large versus small) and then using IAT procedures to measure these arbitrary associations. Scores on this "geometric association" IAT correlated with the original "flower-insect" IAT that is the procedural template for all IAT work on prejudice. Such consistency across IATs points to a general cognitive-procesing factor that, independently of prejudice, shapes performance on all speeded binary classification tasks.

(2) Validity disputes: Which IAT correlates support the prejudice interpretation?

(i) Test-retest reliability and convergent validity. There is no theory that tells us how stable over time implicit prejudice should be. But if we posit implicit prejudice to be a stable property of individual observers (as we must if implicit prejudice is to serve as a full functional substitute for old-fashioned prejudice in the hard core of the statist-interventionist research program), we should expect high test-retest correlations. These coefficients for the IAT and affective priming are, however, at the low end of the accepted range for measures of stable individual-difference constructs (between 0.5 and .6) (Fazio & Olson, 2003).

There is also no theory that tells us to what extent different reaction-time tasks tap into similar or different types of implicit prejudice. The two most popular tests—the IAT and affective priming—are weakly correlated (Fazio and Olson, 2003) and such multi-method divergence leaves dangling questions: Are discrete pockets of unconscious prejudice in the mind accessed by the IAT and affective priming? Should we accept the measurement-error defense that, after correcting for unreliability, the modest correlations look respectable? Or are processes unrelated to unconscious prejudice and to each other shaping responses to the two tests?

(ii) Correlations between implicit and explicit measures. Again, there is no theory that tells us what correlation to expect. Should we interpret zero correlations as evidence of discriminant validity (as IAT supporters sometimes have; Banaji, 2001) or interpret low-positive correlations as evidence of convergent validity (as IAT supporters sometimes have; Greenwald et al., 2003), or—for that matter—interpret low-negative correlations as evidence for the old psychodynamic pattern of reaction-formation: mask unconscious racism by embracing fiercely egalitarian attitudes (see Hofmann et al. 2005 for an effort to organize the disparate findings).

(iii) Correlations between implicit measures and interpersonal behavior. Beyond allusions to the automatic-controlled distinction (Chugh, 2004), the literature gives little guidance on when to expect unconscious prejudice to predict cross-racial interactions. Will African-Americans detect subtle indicators of hostility among high IAT scorers–and react more negatively (as some have argued; McConnell & Leibold, 2001)? Or will African-Americans interpret the more relaxed stance of low scorers as disengagement and react more positively (as others have argued; Shelton et al., 2005)? Or perhaps there is no connection: what Blacks see as unfriendly avoidance of eye contact is really Whites trying to follow race-blind norms (Norton et

al., 2006), with Whites who fear they have failed a test of implicit bias trying the hardest.

(3) Synthesizing the construct-validity arguments.

What proportion of variance in IAT scores should we attribute to unconscious prejudice after we subtract the variance attributable to the various counter-interpretations? No one knows precisely. But our best guess is that the answer will prove closer to 0% than to the "100% prejudice" position of Banaji et al. (2004).[9] Our guess is grounded in test-retest reliability studies that reveal half or more of IAT variance to be error variance and validity studies that reveal the need to apportion chunks of the remaining variance to the alternatives laid out here.

B. Psychometric debates: When do we have warrant to label people who "fail" tests of "implicit prejudice" as prejudiced? As prone to discriminate?

Here we confront two analytic challenges: (1) the arbitrary metric problem: IAT scores, by themselves, say nothing about propensity to discriminate; (2) the diagnosticity problem: the existence of a correlation in a lab study of sophomore conscripts does not justify inferences about the propensity of the entire sample, less still the American population, to discriminate.

(1) Arbitrary metrics.

Researchers sometimes imply that the IAT is a face-valid measure of bias because any deviation from zero in reaction times to compatible versus incompatible trials reveals a

---

[9] We cannot, however, assume that definitions stand still. Banaji et al. (2004) have already warned against nostalgically clining to Allport's (1954) old-fashioned definition of prejudice: irrational animus toward a group that one is unwilling to give up in the face of disconfirming evidence. Reputational bets in adversarial collaborations require shared definitions of key constructs.

predisposition to respond more favorably toward one or another group (e.g., Ziegert & Hanges, 2006). But as Blanton and Jaccard (2006) note, for many samples and settings, a reaction-time differential of 200 milliseconds around the IAT zero point of perfect color-blindness may have no behavioral implications whatsoever. Quite simply, no one knows: the true functional relationships between reaction times and behavior have yet to be mapped.

The problem is that, although reaction time resembles an objective ratio-scale metric, of the sort common in physical science, it is an arbitrary metric for its intended purpose in the IAT: assessing the psychological construct of implicit prejudice. As Blanton and Jaccard (2006, p. 27-28) note, "it is not known where a given [reaction time] locates an individual on the underlying psychological dimension or how a one-unit change [in observed reaction time] reflects the magnitude of change on the underlying dimension. . . . " The IAT research program has tacitly assumed that different reaction times across compatible and incompatible trials reflect relative preferences for groups. Yet "no published study has shown that the zero-point used to diagnose attitudinal preferences is the true dividing line between preference for Blacks versus Whites" (Blanton & Jaccard, 2006, p. 34).

(2) Logical Lapses.

Critics have noted at least two related flaws in IAT analyses: (1) even if we assume the IAT is a highly sensitive tool with respect to the detection of implicit bias, the modest correlation of implicit attitudes and discriminatory behavior means that many persons who "fail" the IAT will be mislabeled as being unconsciously prejudiced, if "being unconsciously prejudiced" is assumed to have behavioral connotations; (2) the use of general language about a linkage between IAT scores and behavior (e.g., Kang & Banaji, 2006) obscures just how few people who

show implicit bias on the IAT show any evidence of discriminatory behavior in experimental tests of the IAT-discrimination linkage and just how poor the IAT performs as a predictive instrument. If "any psychological tool is only as good as its ability to predict human behavior (McConnell & Leibold, 2001, p. 440)," these flaws present serious challenges to the IAT research program.

(i) Small correlation coefficients can lead to big classification errors. A seductive but specious syllogism underlies the implicit-prejudice argument that many scholars import into employment law. The major premise is that most people harbor the morally corrosive construct "implicit prejudice." (After all, 78% show implicit bias on the IAT, with 85% of Whites biased against Blacks (Greenwald & Krieger, 2006).) The minor premise is that implicit measures of prejudice have low but positive correlations with behaviors broadly construed as discriminatory (roughly .25 in Poehlman et al.'s (2006) meta-analysis). Thus, we can safely conclude that most people will engage in racial discrimination..

The conclusion does not follow from the premises. Indeed, the opposite is closer to the truth: the fact that so many fail the IAT plus the fact that IAT scores have low positive correlations with behavior expansively defined as discriminatory guarantees that many non-discriminators will be labeled prejudiced (Fiedler et al. 2006), without having done anything other than show differential reaction times on the IAT. We can estimate how common false accusations will be if we make assumptions about three parameters: (a) how often people fail the IAT (let's stipulate a "failure" rate between 70% and 90%); (b) the probability that a truly prejudiced person will "fail" the IAT (let's stipulate hit rates between 70% and 90%, which we view as generous given the numerous confounds distorting IAT scores); (c) the base rate of

"true" prejudice in the population (drawing on survey research as a starting point for analytical purposes, let's stipulate base rates between 10% and 30%).

Bayesian analysis reveals that using IAT scores to predict discrimination will yield false-accusation rates between 60% (when the IAT failure rate is 70%, the base rate of true-prejudice is 30%, and the IAT hit rate is 90%) and 90% (when the IAT failure rate is 90%, the base rate of true prejudice is 10% and the IAT hit rate is 70%).[10] Whether one deems such false-accusation rates excessive is a matter of political values, not scientific fact.

IAT defenders could counter that much hinges on assumptions about the population base rate for true prejudice. But defenders flirt with tautology if they argue that this base rate must be in the 70% to 90% range because those percentages fail the IAT. That argument carries weight only if we grant that the IAT is a pure measure of prejudice despite the arbitrary nature of the IAT metric and its lack of any external behavioral referents.

The data on IAT-criterion variable correlations point to a path out of this morass. We can deduce the correlations that must hold between the IAT and criterion variables capturing true (discrimination-inducing) prejudice when we make sufficiently specific assumptions about the p(prejudice | IAT failure), the p(IAT failure), and the base rate of true prejudice, p(prejudice).[11]

---

10 The estimates of 61% to 92% are derived from the most basic definitions of probability, as follows: p(not prejudiced |failure on the IAT) = 1 – p(prejudiced | failure on the IAT) = 1 – (p(failure on IAT | true prejudice) * p(true prejudice)/p(failure on the IAT)). Bayes' theorem provides the proofs of these results (Price, 1763).

11 "IAT failure" denotes a score supposedly indicative of implicit bias toward a group. The estimates in the text are derived from the logic of correlation coefficients. If a = p(prejudice and IAT failure), b = p(no prejudice and IAT failure), c = p(prejudice and IAT passing) and d = p(no

The results tell us that the population base rate of true prejudice must be as low as 20% when 70% "fail" the test, the p(IAT failure| prejudice) is 90%, and the correlation of the IAT with criterion variables is .22 (this latter value corresponds to the average correlation in Hofmann et al.'s (2005) meta-analysis). Put another way, the data cited by IAT proponents to support their test dovetails nicely with survey research estimates of the population base rate of true prejudice.

Such analyses, by themselves, neutralize much of the purported relevance of implicit prejudice research to the broader statist–interventionist reformist agenda. But even these estimates may still be too flattering to the IAT because they treat IAT correlations with criterion variables as perfect proxies for the IAT correlation with true prejudice—a dubious assumption in view of the plethora of alternative explanations noted earlier.

(ii) Correlation coefficients are sensitive to outliers, insensitive to absolute scores, and imperfect indicators of diagnostic utility. Blanton et al. (2007) reanalyzed the data from Ziegert and Hanges (2006), the only published study reporting a correlation between IAT scores and discrimination in hiring decisions.[12] Blanton et al. (2007) found that Ziegert & Hanges' main finding of a correlation between IAT scores and discriminatory ratings of hypothetical Black job candidates was not robust either to the deletion of outliers or to changes in regression

---

prejudice and IAT passing), the correlation is (ad – bc)/(a*b*c*d)^.5. If we make assumptions about p(IAT failure), p(prejudice), and p(IAT failure | prejudice), we can solve for the correlation between IAT and prejudice. We can then restrict that range from .2 to .25, find the associated values of p(prejudice | IAT failure), and convert those values to p(no prejudice | IAT failure).

[12] Technically, this study is not an IAT study in the Greenwaldian sense, because Ziegert and Hanges employed their own unique scoring algorithm for the IAT and did not report results using the latest version of the IAT scoring algorithm, which was designed to minimize unwanted influences, such as differences in cognitive processing speed, on the IAT score (Greenwald et al., 2003). We requested the raw IAT data so that the effects of alternative scoring methods could be examined, but Ziegert reports a loss of these data.

techniques. In one analysis, the main finding was eliminated by removal of a single outlier who exerted extreme influences on the key regression parameter.[13] To address the question of the IAT's predictive utility, Blanton et al. calculated prediction intervals for the value of every IAT score and found that the interval spanned negative to positive hiring ratings and included the value of zero for every IAT score in the sample. For example, for the participant with the highest IAT score in Ziegert and Hanges' sample, the 95% prediction interval was -0.2 to 3.7 and, for the lowest observed IAT score, the prediction interval was -4.6 to 0.9. These data indicate that the predictive utility of the IAT is limited, even in a setting such as Ziegert and Hanges (2005), in which participants were instructed to behave in a racist manner. Furthermore, a regression model that included the IAT as a predictor performed almost identically to a model that did not (error variance of 0.76 in model with IAT versus error variance of .78 in model without). These results undercut the view that most Americans seize on any pretext to treat Blacks unfairly but quite compatible with the view that, with scattered exceptions difficult to identify with any measure of prejudice used in the study, most people treated Blacks fairly.[14]

---

[13] Ziegert and Hanges (2006) found a significant correlation only in the condition in which subjects were told that the president of the hypothetical company wanted to hire a white person. Reactions to this result again illustrate each camp's ability to assimilate findings to its favorite framework. Skeptics see the result as irrelevant to antidiscrimination law (even ignoring the outlier problem) because discrimination occurred only under conditions that blatantly violated existing law and because the IAT did not predict discrimination under more realistic conditions (Mitchell & Tetlock, 2006). Proponents still see the result as revealing: "This manipulation [in Ziegert & Hanges (2006)] seems unrealistic because such preferences are no longer written down; that said, the outlandishness of the request should have worked against finding any behavioral correlation." (Kang & Banaji, 2006, p. 1074, n. 59).

[14] Blanton et al. (2007) also re-analyzed McConnell and Leibold's (2001) data, which are often cited as evidence that IAT scores predict unfriendly nonverbal behavior toward minorities. These data were also not robust against outliers, and the correlations revealed that subjects with anti-Black bias on the IAT were rated as friendlier to a Black than to a White experimenter (see also Richeson and Shelton, 2005).

C. External-validity debates: Does implicit prejudice cause discrimination in workplaces?

Even if the IAT research had escaped unscathed from earlier critiques, there are daunting obstacles to generalizing from stylized, low-stakes lab experiments to complex, high-stakes workplaces that have often erected institutional barriers – training and accountability procedures – against unlawful discrimination. Superficially, the former settings seem ideal for documenting the primacy of automatic associative functioning (what Kahneman and Frederick (2002) have termed System 1 processing), whereas the latter settings seem ideal for documenting the primacy of higher-order reflective thinking (System 2 processing in Kahneman and Frederick's model).

But it is a mistake to reduce external-validity issues to the tension between automaton versus volitional models of human nature. Although the staunchest defenders of the IAT seem wedded to quite strong versions of System 1 dominance (Bazerman & Banaji, 2004; Chugh, 2004), there is no necessary linkage: challenges to external validity can take either System 1 or System 2 forms. For instance, skeptics could embrace an old-fashioned automaton of human nature, with habit-family hierarchies (Hull, 1934), or a contemporary view, with automatic activation of associations (Bargh & Chartrand, 1999), but invoke countervailing automatic correction tendencies (Glaser & Kihlstrom, 2005). Or they could balk at the notion that the associations activated in workplaces are predominantly racial—and point to a vast array of workplace cues that could prime a vast array of associations, such as positive team spirit, norms of reciprocity, team-based in-group definition, deference to authority, and over-learned solutions to shared problems. Several lines of research reinforce such reasoning:

(a) work on affective priming shows that when racial stimuli are embedded in positive

social contexts (such as churches or workplaces), the effects of "implicit prejudice" fall close to zero (Barden et al, 2004). Such manipulations are pallid imitations of the lengths to which some companies go in engineering egalitarian, can-do cultural atmospherics (Arthur & Doverspike, 2005);

(b) work on the impact of individuating information shows that stereotype effects recede as people learn more about each other, with individuating information often overwhelming stereotype information (Kunda & Spencer, 2003), perhaps even at the implicit level (though this question has been neglected; see Govan & Williams, 2004);

(c) work on the intergroup-contact hypothesis shows that intergroup contact, even absent Allport's (1954) optimal conditions for prejudice reduction—equal status, common goals, intergroup cooperation, and institutional support—are capable of reducing prejudice (Pettigrew & Tropp, 2006), including at the implicit level (Aberson et al., 2004; Olsson et al., 2005)

The problem here is foundational. Although it is impressive when trivial primes influence overt behavior—such as reminders of Florida slowing how fast NYU students walk to an elevator (Bargh et al, 1996) —we know little about how such effects hold up when placed in competition with the cacophony of competing cues in real life. As Bargh (2006) has himself noted, it would arguably be impossible to pursue any long-term goals if we did not have sturdy executive-override capacities to block associative distractions of the moment. Implicit-prejudice researchers can counter that racial or gender cues have trump status in controlling behavior (Reskin, 2000), but that claim remains, for now, just a claim (and even the long-standing claim about the automaticity of race encoding has been qualified recently; see Cosmides et al., 2003).

Skeptics thus have solid System 1 lines of defense. And they also have recourse to

System 2 lines of defenses which deny that people are associative automatons. People are capable of stepping back and reflecting on their conduct, and many features of modern workplaces—such as accountability for implementing best-practices in performance appraisals and coaching—are well-designed to bring out the self-reflective potential in human nature.[15]

No one should be faulted for not working out a comprehensive theory of when to expect System 1 versus System 2 dominance. But IAT proponents should inform policy audiences of how little is known about where to set boundary conditions on their claims. Lacking such guidance, the next best thing is to look to the empirical literature for clues. Organizational psychologists and sociologists have developed elaborate lists of precautions that organizations should take to squelch discrimination—from micro (e.g., specificity in performance rating scales) to macro (e.g., powerful Human Resource Management operations and greater regulatory oversight (e.g., Brief et al , 2005; Kalev et al., 2006). Table 2 organizes these prophylactics into the phases of personnel decisionmaking that each targets for "de-biasing."

Table 2

Illustrative Debiasing Procedures Proposed by Academic Researchers and Expert Witnesses

	Use Evidence-Based Procedures	**Hold Managers Accountable for Applying Procedures**	**Help Managers Internalize Moral and Business Rationales for Procedures**
Recruitment, Screening and Hiring	Eliminate stereotypic statements from recruiting materials and base job descriptions on actual tasks	Professional HR staff monitor all processes and outcomes and hold managers responsible for lapses from	Cultivate awareness of potential biases in informal network methods of recruiting and in informal methods of screening

[15] Personality dimensions—such as motivation to control prejudice—may also influence how often people rein in prejudicial impulses, at both the System 1 and 2 levels (Amodio et al., 2003; Devine et al., 2002; Glaser & Knowles, in press; Hausmann & Ryan, 2004).

	performed rather than characteristics of persons traditionally employed in the position; cloak data on race, sex, age and other irrelevant attributes of applicants to the extent possible;use structured interviews focused on knowledge, skills and abilities derived from professional job analyses and/or objectively validated tests; use structured recall procedures following interviews	best practices and scrutinize significant group deviations in outcomes; consequences should reliably follow from observing or not observing rules; make managers feel accountable to superiors who are known to endorse EEO norms and have a commitment to building a diverse workforce	and selecting applicants; emphasize benefits of commitment to level playing field, of a diverse workforce, and of holding managers accountable to audience that endorses equal opportunity; emphasize the need for diversity measures to overcome past barriers; distinguish identity-conscious measures focused on ensuring equal opportunity from quota systems
Performance Evaluations and Management of Existing Employees	Use objective performance measures to the extent possible; when subjective measures are used, train managers in to apply them consistently across groups based on clear guidelines for their application, and monitor for disparities in their application; use behaviorally anchored rating scales that tap performance dimensions derived from job analyses; use structured recall before recording performance ratings; employ diverse teams in cooperative tasks and emphasize outcome interdependence when possible.	360 degree accountability in which managers must answer not only to their supervisors and HR but also to employees; monitor possible adverse impact in these processes; use group discussion and consensus among raters to the extent possible, preferably with women and minorities in the rating groups; ensure authority figures support intergroup contact; assign responsibility for EEO compliance and diversity gains to high-level officer	Cultivate awareness of recall and rating biases (e.g., halo effects and leniency bias) and self-fulfilling stereotypic prophecies and their possible adverse effects on efficiency; explain benefits of intergroup contact, shared goals, and common group identity in the reduction of stereotypes and prejudice
Promotion, compensation, and bonuses	Use structured interviews focused on job-relevant attributes plus explicit guidelines for how to integrate data on past performance, experience, seniority, and labor market conditions in making decisions; provide remedial training for groups historically under-represented in upper management	Mix of process accountability (monitoring to ensure managers make decisions in right ways) and outcome accountability (monitoring for possible adverse impact in all facets of decision making).	Alert to compounding power of cumulative advantage models of achievement; Sensitize to subtle ways in which bias can arise in cognitive, interpersonal, and institutional systems

Sources: Armstrong (2006); Arthur & Doverspike (2005); Baltes et al. (2007); Bielby (2005); Cascio & Aguinis (2005); Cleveland et al. (2000); Dobbs & Crano (2001); Equal Employment Opportunity Commission (2003); Fiske (2005); Ford et al. (2004); Gaertner et al. (2005); Kalev et al. (2006); Konrad & Linnehan (1995); Latham & Wexley (1994); Perloff & Bryant (2000); Roch (2006); Woehr & Huffcutt (1994).

Notwithstanding the confidence with which some experts express their judgments in class-action litigation critiquing the personnel practices of specific companies (e.g., Bielby, 2005), the truth is that we know little about the main effects of most of these variables on discrimination (where along each causal continuum does discrimination fall to zero?) and virtually nothing about the interactive effects of these variables (taken in complex multivariate combinations, how does the likelihood of discrimination wax or wane?). The reality could be that the "maximalists" are right: organizations can check discrimination only if they have all variables set at maximum intensity. Or it could be that, in their zeal to eliminate one form of discrimination, the maximalists have—by setting certain combinations of identity-conscious variables on "high"—triggered reverse-discrimination effects (Gullett, 2000; Matheson et al., 2000), perceptions of distributive and procedural injustice (Cropanzano et al., 2005; Hegtvedt et al., 2002; Richard & Kirby, 1998) self-doubt among the beneficiaries (though Crosby et al., 2003, see this risk as small), leniency bias in performance ratings of minority employees and job candidates (Harber, 1998, 2004), backlash reactions to specific employer diversity initiatives as well as broader opposition to remedial policies (Aberson & Haag, 2003; Harrison et al., 2006; Kidder et al., 2004;) and stereotypes about the ability of disadvantaged groups to achieve on their own (Heilman & Welle, 2006). Indeed, raising consciousness about group-based disparities in outcomes and the prospect of unconscious discrimination may lead White managers, in their efforts not to take race into consideration, to engage in the very kinds of controlled and

seemingly less friendly interpersonal behaviors that statist-interventionists contend disadvantage minorities in job interviews and employee-manager interactions (see Norton et al., 2006). Or it could be that, if all of these best practice recommendations were adopted, the measures would work at cross-purposes (e.g., group stereotypes could be weakened by shared goals and diverse work teams but group divisions could be strengthened by identity-conscious accountability practices that create de facto quota systems). Or it could be that researchers have mis-estimated conflicting effect sizes of key variables (e.g., moving from subjective to objective performance measures may lead to worse evaluations on average for certain protected groups (Roth et al., 2003) because the effects of stereotyping in subjective measures are out-weighed by pre-existing group differences in human capital (Roth et al., 2003). Finally, it could be that incremental implementation of subsets of the recommendations are sufficient to check prejudice in most early 21st century American work places.

The parametric complexity is staggering—which should give pause to erstwhile reformers. But the statist-interventionist case would be bolstered if implicit-prejudice researchers could show that unconscious biases reliably influence consequential decisions in actual workplaces that institutionalize varying degrees of protection aginst prejudice. To date, however, only a few studies have addressed this external-validity question, and these studies provide far-from-decisive evidence that implicit associations are potent causes of discrimination in personnel decisions. We have already noted the interpretive problems of the Ziegert and Hanges (2005) study in our discussion of outliers, but it is worth stressing that their problematic correlation between implicit associations and discrimination arose only when the "boss" in the simulated work environment explicitly demanded discrimination against Blacks. Rudman and Glick (2001)

found a correlation between subjective judgments of the social skills of a hypothetical women showing masculine, as opposed to feminine, traits who applied for a job requiring social sensitivity and subjects' scores on an IAT designed to measure gender stereotypes, but not between subjects' judgments of the "hireability" of this applicant and IAT scores.[16]

Given the paucity and mixed character of the evidence, a priority for would-be exporters of IAT findings should be to explore how well such effects hold up in real workplaces—a central goal of the adversarial collaborations proposed in the next sections of this chapter.[17]

IV. Coping with indeterminacy

One might suppose that each validity challenge—construct, psychometric, and external—has direct logical implications for the unconscious-prejudice component of the racism debate. If sound, each objection should presumably induce rational proponents of the statist-interventionist research program to lower their estimates of the potency of unconscious prejudice. If rebutted, each rebuttal should presumably persuade rational skeptics to raise their estimates. In this

---

[16] A few other studies examine the relation between implicit prejudice and subjective evaluations of group members, but not in work settings. Ashburn-Nardo et al. (2003) examined the relation between African-Americans' scores on the race IAT and their judgments of a Black or White person as a partner in an upcoming anagram-solving task. They found no relation between subjects' actual choices of partners and IAT scores, but they did find a correlation between IAT scores and subjects' expectations about partnering with a White or Black. Rudman and Lee (2002) found a correlation between scores on a race-stereotype IAT and subjects' ratings of the hostility and sexism evidenced by a Black target's ambiguous behavior, but this correlation did not hold for subjects' ratings of the Black target's intelligence or for ratings of a White target.

[17] However, field studies should—at minimum—control for objective differences in job performance across groups, which may reflect true differences in performance or may reveal, as a recent meta-analysis has (Roth et al., 2003), that objective measures have even greater adverse impact on minorities than subjective measures. That particular debate returns us to other aspects of the American racism debate: disputes over the size and interpretation of group differences in "human capital".

idealized scenario, the scientific method will soon compel observers from very different starting assumptions to converge on the truth, with no need for heroic interventions.

The Lakatosian analysis of research programs warns us, however, to set our expectations lower. Rationality is an elusive concept in this context (Laudan, 1996)—and there are two looming epistemic obstacles to any smooth positivist convergence on the "truth:" the absence of shared standards of evidence and of shared standards of proof. We show in the following sections that, although proponents and skeptics of unconscious prejudice both profess fealty to the scientific method, each camp assimilates new results into its own distinctive conceptual frameworks and what one side elevates to decisive fact, the other may brush off as an irrelevancy. We then lay out why we believe the usual methods of scientific dispute resolution are not up to overcoming these obstacles to scientific progress—and need to be supplemented by adversarial collaboration. Convergence will not be easy, but it need not be as hard as it now is.

A. Barriers to progress

(1) Absence of shared standards of evidence.

The neo-positivist ideal—transparent mapping of empirical facts onto theoretical propositions—is possible only in a world with well-defined correspondence rules that tightly couple theories and methods for testing theories (Suppe, 1977). This precondition is often satisfied in physical science, but rarely in unconscious prejudice research—a domain in which investigators and their critics have vast freedom to second guess operational definitions of theoretical concepts and big incentives to do so.

Imagine the worst-case scenario for statist interventionists still consistent with extant empirical knowledge. Over the next decade, this camp "loses" all of the validity arguments in

Section III. Whenever we introduce independent variables to capture realistic work conditions and dependent variables to capture tangible and fungible personnel decisions, we find no correlation between implicit prejudice and discrimination. Insofar as IAT-like measures show any predictive validity, the coefficients are confined to low-accountability lab settings in which participants have no incentives to make accurate personnel judgments, little or no individuating information about the employees being judged, and no training in how to make personnel decisions. These modest "successes" are also confined to nonverbal and paralinguistic dependent variables ("eyeblink" measures), and disappear when we remove just a few outliers.

The empirical news would be unrelentingly grim, but statist interventionists still deftly deflect the awkward laboratory findings. One marvelously adaptable defense is to complain that the independent-variable manipulations, or the background context, were so laden with demand characteristics that they made the purposes of the studies transparent and thus hopelessly obscured the effects of implicit bias. Hyper-self-conscious participants making choices in such odd settings can hardly be deemed representative of personnel decision makers in the real world. After all, who doubted that prejudiced managers under scrutiny could craftily conceal their animus behind subjective evaluation schemes packaged as business necessities?

Other defenses are more context-specifc but still effective. Recall the many methodological reasons invoked for null hypothesis or unexpected-reversal results in Section III: to argue that the IAT "does not work properly" when certain stimuli are used or certain scoring conventions violated or to maintain that high-prejudice IAT scorers do worse in cross-racial encounters, except when they do better because they are less nervous or more engaged.

When field studies are the source of the disappointing data, defenders of the research

program again have a panoply of options. They can insist that organizations that yielded no evidence of implicit bias are atypical, conjecturing that only organizations with unusually rigorous EEO compliance regimes would hazard the rigorous scrutiny of outsiders. Or they can insist that the criterion variables were insensitive to subtle forms of prejudice that preceded the personnel decisions, thereby creating the illusion of objective support for those decisions. If prejudiced managers make minorities uneasy in interviews, and there is evidence for such self-fulfilling prophecies (Sekaquaptewa et al, 2003; Word et al., 1974), we should expect that even perfectly unprejudiced third-party observers will form less favorable reactions of minorities in such interactions. If prejudiced managers cause minority employees to feel alienated from the company, we should expect that even objective metrics of performance–such as tardiness and absenteeism–will reveal shortfalls. And if the company's clientele is prejudiced, we should expect minorities to suffer under other objective metrics, such as customer complaints and sales commissions (Greenwood, 2005; Kelman, 2001). Taken together, these arguments allow statist-interventionists to dismiss virtually any objective workplace metric as hopelessly contaminated by prejudice—and to insist that we "overcontrol" when we enter for such metrics in multiple regressions designed to identify disparate treatment of "similarly situated" employees.

Only now do we arrive at the most fundamental defense which brings us closest to a tautological affirmation of faith in the hard core. Statist interventionists do not need to hitch their worldview to the validity of any particular measure of unconscious prejudice. Whether the familiarity or cognitive dexterity or prejudice camp "wins" the sub-debate over the IAT may just not matter for purposes of the wider racism debate. Indeed, there is no logical inconsistency between embracing the hard core of the statist interventionist research program and disavowing

all previous empirical efforts to capture subtle prejudice, including symbolic and modern racism scales, implicit-association tests, affective-priming indices and even fMRI scans of the human brain at work. It is never comfortable to strip off all the empirical cover from a hard-core proposition, but who would deny the amended hard-core proposition that subtle prejudice is an elusive construct and that we should expect numerous false starts.

This latter argument does, however, raise the waiting-for-Godot question: How long is it sensible to sustain faith in hard-core propositions? It is logically true that skeptics can never prove a negative existential claim (such as "unconscious prejudice exists only as a convenient figment of the liberal imagination"), but this will be faint consolation for statist interventionists wooing impatient moderates in the public square. Research programs cannot be logically falsified, only psychologically exhausted. And, although we are not close to that point now—the energy of the unconscious-prejudice program offers that much reassurance—it is a good epistemic practice to be vigilant for warning signs that programmatic quests for hard-core-validating evidence have reached points of diminishing marginal returns.

Of course, fair play requires reversing the imaginative exercise. Suppose that market purists or, more generally, skeptics of the unconscious-prejudice arguments lose all of the validity arguments—and the correlation between the IAT and indisputable discrimination is strong even when we introduce independent variables, such as accountability, that mimic realistic work conditions. Suppose also that researchers have resolved the psychometric disputes concerning where, along the IAT scale, predictive validity kicks in—and the IAT does possess the elegant interval-scale properties that its proponents claim (Greenwald et al., 2006).

If lab studies are the sources of the dissonant data, skeptics can argue that investigators

failed to capture either the full power or right configuration of independent variables in the real world. After all, it is hard to create lab forms of accountability as potent as real-world forms where careers are at stake—or to create the depth of individuating information available to managers who often have years of observations—or the cooperative team dynamics that often exist in workplaces—or the powerful incentives that managers have to put the right people in the right jobs. These interpretive maneuvers are curious mirror images of the maneuvers deployed by statist interventionists when they defended their hard-core against dissonant lab results by emphasizing the potency, not the feebleness, of lab manipulations. The ability of each side to invoke such starkly contradictory objections–and our inability, either a priori or a posteriori, to rule out either set of objections–underscores the tenuousness of the correspondence rules linking theories and methods in this sector of social science. Terms like "strong," "weak," "artificial," and "realistic" tell us what the speaker approves or disapproves of, and not much more.

If field studies are the source of the dissonant data, skeptics can no longer stress the artificiality of laboratory manipulations but they can highlight the confounding variables that come into play as we move into the real world. Even when company-wide data point to adverse impact on minority employees after controlling for objective job-relevant variation in knowledge, skills and abilities, skeptics can insist that the multivariate regressions have missed key dimensions of human capital—and that, once captured, the alleged biases will vanish (Ananda & Gilmartin, 1992;). The control variable of choice has traditionally been intelligence. Market purists as well as some eclectic thinkers have long resonated to research claims concerning the power of general cognitive ability (*g*) to predict performance across a massive array of jobs (Hunter and Schmidt, 1996; Gottfredson, 2000)—a finding that, if true, forces

statist interventionists into contortions because (*g*) often correlates disturbingly highly with membership in certain minority groups (Rushton & Jensen, 2005), and statist interventionists must then either engage in trench-warfare, job-by-job, validity-generalization challenges to ability measures (Copus , 2005) or invoke the blanket systemic-racism argument attributing variation in cognitive abilities to past and current barriers to educational opportunity and argue that a valid but weakly predictive test is being used to perpetuate inequalities (Kelman, 1991).

Ferocious though *Bell-Curve*-style exchanges over race and IQ can be (see Alderfer, 2003; Sesardic, 2000), market-purists can also choose among various lines of intellectual retreat when the need arises. They can simply tap into the vast universe of individual difference variables in psychology in search of other dimensions of objective intergroup variation (Pervin, 1998). Again, we run into the unprovabilty of negative existential statements (in this case, "the missing human-capital variables exist only as figments of conservative imaginations")—and the "waiting-for-Godot" question: How patient are we prepared to be?

The similarities in defensive maneuvers do not end here. When it becomes hard to second-guess objective gauges of knowledge, skills and abilities or of performance, market purists can—like statist interventionists—challenge the representativeness of the sample of organizations. Market purists must, however, stress how unusually bad, not good, the chosen organizations are. The standard argument is the "few bad apples." The researchers stumbled upon a few rogue organizations—or a few rogue managers within otherwise good organizations.

In sum, there is not perfect symmetry between the escape clauses available to statist interventionists and market purists but the parallels are striking. As Lakatos warned is inevitable in complex empirical disputes, each position is buffered by a thick layer of auxiliary hypotheses.

(2) Absence of shared standards of proof.

The preceding argument implies that each research program can absorb hard empirical hits and still coherently defend its hard core. At best, convergence will be painfully slow; at worst, we should expect to be trapped for eternity in a theater-of-the-absurd dialogue of the deaf.

It scarcely seems necessary then to seek additional impediments to the advancement of knowledge. But there is a big one we would be remiss not to mention: even if we witnessed, *mirabile dictu*, the instant resolution of the scientific debate over unconscious prejudice, down to its tiniest moderator-variable details, it does not follow that we would see the resolution of policy debates. Disputes over facts should not be conflated with disputes over values–a point that David Hume made compellingly two centuries ago when he warned of the is-ought fallacy, the failure to recognize the impossibility of deducing an "ought" conclusion from purely "is" or factual premises (Hume, 1739). For better or for worse (and we think for better), this pithy insight is now firmly ensconced in contemporary philosophy-of-science frameworks such as logical empiricism, which treats value judgments as non-falsifiable matters of taste (Suppe, 1977), and contemporary scientific frameworks such as signal detection theory, which distinguish perceptual acuity (*d* prime) for sorting out signals from noise from value thresholds (beta) for balancing false-positive errors (announcing a signal when there is none) and false-negative errors (failing to announce signals that are present) (Swets et al., 2000).

One implication of this fact-value distinction has special relevance here. Imagine that researchers overcome all of the Lakatosian barriers to cross-theory convergence and that the clashing camps declare in unison that, if a particular pattern of scores on an implicit measure of prejudice is found among company management, the probability of false-positive classifications

of the company will be 0.4 and the probability of false-negative classifications will also be 0.4. Moreover, imagine that they agree that if the law changes in direction x, the probability of false-positives will fall by 1% per unit change and that if the law moves in direction y, the probability of false negatives will fall by 1% per each unit.

Such an empirical convergence would be extraordinary given the noisiness and method-specificity of social science data. But it would not guarantee policy convergence. First, as in most value-laden controversies, we need to cut through layers of intellectual self-deception and public obfuscation: policy elites are notoriously reluctant to acknowledge trade-offs between false-postive and false-negative classification errors (Jervis, 1976; Tetlock, 1998). Second, assuming we can somehow tap into the trade-off functions of thoughtful observers, we will probably discover that statist interventionists are prepared to pay a steeper price in false-positive convictions of organizations in order to keep false-negative exonerations down—and that market-purists hold the flipside preferences.[18] The debate will have advanced, but onto new and perhaps more treacherous terrain—with the new focal point of contention being where policy makers should set their thresholds of proof for translating changes in probabilistic beliefs about unconscious prejudice into changes in policy preferences.

---

[18] We conducted an anonymous survey of law students at the University of Virginia to explore aversions to Type I and Type II errors as well perceptions of discrimination. Although we cannot equate liberalism and statist-interventionism or conservativism and market purism, these ideologies and regulatory positions are likely reasonably correlated (Sniderman & Carmines, 1997). Over 300 students from a student body of 1100 completed the survey, with the sample slightly tilted toward the liberal end on the scale (Mean = 3.24 on a 1-5 ideology scale, with 5 as most liberal). Liberals rated Type I errors in employment litigation (failing to hold a discriminator liable) more harmful than did conservatives, whereas conservatives rated Type II errors (holding a non-discriminator liable) more harmful than did liberals.

Figure 1 clarifies the concept of threshold of proof. The top panel lays out hypothetical probability estimates that proponents and skeptics attach to the claim that unconscious prejudice can explain 10% or more of the variance in personnel judgments in three settings: first, in companies that make no effort to hold managers accountable for upholding EEO norms; second, in companies that create EEO forms of accountability that require managers to make decisions in prescribed ways but do not manadate result; and, third, in companies that strive for equality of result by rewarding managers for achieving numerical goals. As Figure 1 shows, our hypothesis is that proponents see a greater likelihood of unconscious prejudice exerting influence in workplaces, especially in the first and second scenarios where there is more ambiguity about the power of institutional checks against prejudice. The bracketed regions capture fact-driven disagreements over effect sizes, with the largest disputes arising when there are no accountability constraints (S_A - P_A) and when there are EEO process-accountability constraints (S_B - P_B) and the smallest disputes arising when managers are under such pressure to achieve equality of result that neither proponents nor skeptics see much role for unconscious prejudice (S_C - P_C).[19]

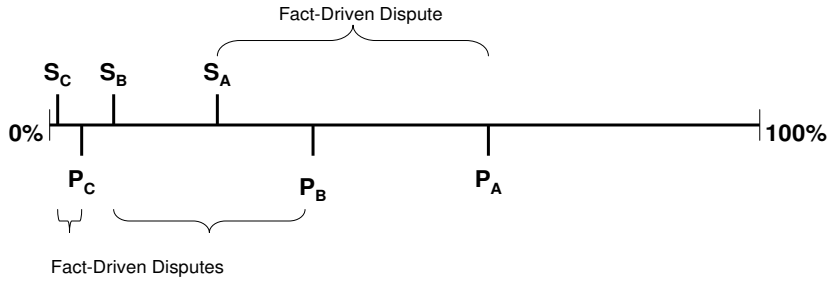
Figure 1

---

[19] Cognitive consistency theories suggest that causality is likely to be bidirectional. On one hand, experts' opinions on the potency of unconscious prejudice influence their policy stands on quota systems; on the other hand, experts' policy stands direct their search for justifications–and the bigger (or smaller) an estimate of the potency of unconscious prejudice experts can find, the more compelling becomes their case for (against) quota systems.

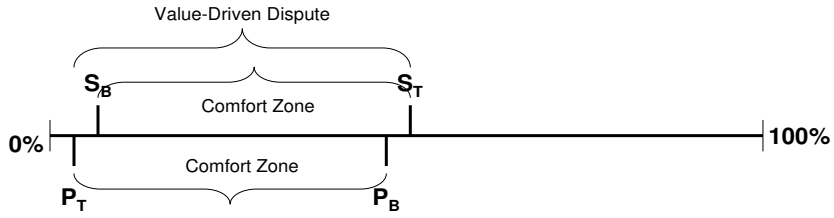
Estimates of Implicit Prejudice Explaining 10% or More of Variance in HR Evaluations

$S_{A, B, C}$ = Skeptics' estimates in three settings:
No Accountability (S_A), EEO Process Accountability (S_B),
& Equality of Result accountability (S_C)
$P_{A, B, C}$ = Proponents' estimates in same three settings



Value-Driven Policy Threshold Continuum

S_B & P_B = Skeptics' and proponents' estimates of effect sizes with
EEO Process accountability
S_T & P_T = Skeptics' and proponents' thresholds for adopting
Equality of Result accountability



The bottom panel of Figure 1 lays out the value judgments that proponents and skeptics might use in setting their thresholds of proof for classifying organizations as prone to

unconscious prejudice. Our hypothesis is tht proponents will be more tolerant of false-positive errors (condemning innocent organizations) than of false-negative errors (failing to condemn guilty organizations)–a value judgment that should lead them to set weaker standards of proof for classifying organizations as vulnerable to unconscious prejudice (we focus on organizations with EEO process accountability because such organizations are most typical of Fortune 500 companies in the early 21st century). The bracketed region between the skeptic's and proponent's thresholds of proof ($S_T - P_T$) captures disagreements over values, with skeptics requiring a higher probability than proponents of 10% effect sizes before rethinking the adequacy of process-accountability checks on prejudice—and the need to adopt quota-driven affirmative action plans.

As psychologists, we worry that so little is known about how observers of varying views trade-off the risks of false-positive and false-negative classifications of organizations. Neo-positivist injunctions to separate facts from values notwithstanding, we suspect that value judgments play a major, albeit usually covert, role in shaping the behavior of scientists—and that scientists' resistance to belief change is sharpest when that change threatens to bump probability estimates across value thresholds that require policy shifts (Hunt, 1999; Lilienfeld, 2002). Our suspicion is anchored in one of the best replicated predictions of the cognitive dissonance theory: people do not like to admit they were wrong, especially when doing so requires revising public commitments central to their self-concepts, such as "foe of racism" or "defender of meritocracy."

This argument leads to our ideological-comfort-zone hypothesis: scientific and legal observers alike try to maintain a margin of safety so that *if they lost a few big empirical contests over the validity of unconscious prejudice measures, they would not have to change their minds on basic policy issues.* In this spirit, the bottom panel in Figure 1 bestows wide comfort zones on

both proponents (the bracketed region between P_B and P_T) and skeptics (the bracketed region between S_B and S_T). Indeed, we would not be surprised to observe extra-padded comfort zones. Market purists could pad their comfort zones by setting thresholds of proof so high for adopting quotas that even IAT proponents' most generous estimates of the effect sizes for unconscious prejudice fail to pass (S_T would be set above P_B). Here the message to the other side is: "Even if your ridiculously inflated estimates of effect sizes were correct, we would still reject a policy as radical as de facto quotas." And statist interventionists might pad their comfort zones by setting their thresholds of proof so low that even skeptics' stingiest estimates of effect sizes still allow them to reach their desired policy conclusion. The message to the other side is: "Even if your ridiculously tiny effect sizes were correct, we would still embrace de facto quotas."

We see two basic ways in which each camp can preserve its comfort zones: via belief-system defenses that ensure their probabilistic beliefs never stray close to their value thresholds or via value-defenses that ensure their value thresholds never stray close to their beliefs.

To keep probabilistic beliefs in a desired range, scientists can engage in the types of epistemic double standards that cognitive researchers have long known that ordinary people practice: ratchet up the scrutiny of dissonant evidence (applying a "must-I-believe-this?" test) and switch off the scrutiny of consonant evidence (applying the vastly more lenient "can-I-believe-this?" test) (Kunda, 1999; Munro et al., 2004; Tetlock, 2005).[20] The vagueness of the

---

[20] Tetlock's (2005) work on social scientists predicting real-world events illustrates that the workings of belief-system defenses beyond the confines of lab studies of undergraduates. When the unexpected occurred, experts relied on a ingenious mix of counterarguments designed to neutralize diagnostic evidence, including the off-on-timing defense ("what I predicted hasn't

methodological ground rules means, moreover, that it will be hard to catch anyone red-handed. For any study, each side has great latitude to play semantic games. For instance, statist interventionists can portray: (a) lab manipulations as "heavy-handed" when they squelch implicit biases but as "formidable" when they fail to do so; (b) organizational accountability systems that eliminate implicit biases as atypically "rigorous" in their EEO enforcement whereas organizations that fail to check biases as "representative"; (c) correlations as "trivial" when validity generalization studies suggest that intelligence is a reliable predictor of performance across jobs (Kelman, 1991), but as "hefty" when IAT studies find support for unconscious prejudice. As always, market purists can indulge in mirror-image word play.

The aforementioned strategies for deflecting results are useful in lower-level, company-by-company, empirical skirmishes with the other side. It is appropriate, though, to single out the most versatile of all the comfort-zone maintaining strategies: the freedom to inflate or deflate our estimates of prejudice by expanding or contracting our definitions. Statist interventionists do not hide their impatience with those who cling "nostalgically" to Allport's (1954) tripartite definition

---

happened yet but it soon will"), the close-call-counterfactual defense ("what I predicted didn't happen but that was merely because this or that trivial detail derailed the prediction"), the exogenous-shock defense ("what I predicted didn't happen but that was merely because a major force–outside the domain of my theory–derailed the prediction"), and the "unlikely-events-sometimes-happen" defense ("I did attach a low probability to what happened but sometimes low-probability events do happen"). It would be wrong, of course, to dismiss all defenses as indefensible. In the dozens of contexts in which they were invoked, the defenses may often well prove defensible. But taken in the aggregate, the selectivity with which such arguments are advanced is suspicious. When the unexpected occurs, experts rarely say that they were just lucky on timing or on trivial details or on exogenous shocks and that they will therefore soon be proven wrong. Good luck tends to come into play as a tool for dismissing the other side's predictive successes; bad luck as a tool for trivializing one's own predictive failures.

of prejudice: hostility, rigidity and inaccuracy (Banaji, Greenwald and Nosek, 2004). In their view, they have identified previously hidden psychological structures that cause previously ignored forms of disparate treatment–and that the structures involved merit the label prejudice even if they involve no conscious hostility, even if they are flexibly updated in response to evidence, and even if they reflect reality distressingly accurately. Not surprisingly, skeptics are less than impressed by claims of disparate treatment grounded in millisecond differentials in response to arbitrary pairings of stimuli that flash across computer screens or grounded in differential rates of eye blinking in conversations. In their view, this expansive conception of prejudice is presumptuous, a usurpation of the political by the psychological (Arkes & Tetlock, 2004; Redding, 2004; Suedfeld, 2004). The net effect is yet another stalemate in which each side has: (a) wide definitional flexibility because there is no value-neutral definition to anchor the discussion; (b) strong incentives to promote its preferred definition. It is odd to suppose that scientists deserve the final word on what is ultimately a value-laden semantic dispute.

Taken together, these belief-system defenses often function so smoothly that there is no need to resort to the other, less popular, class of comfort-zone-preserving maneuvers: keeping our value thresholds reassuringly far from our beliefs so that, even if we are compelled by inconvenient truths to move our probabilistic beliefs to an unsettling degree, we can preserve our margin of safety by moving our value thresholds to the same degree. This strategy is distasteful because it requires tacitly confessing fallibility: that one would endorse the same policy even if one were less than 100% confident in its correctness. And this strategy becomes an even more bitter pill when, despite one's best negative-heuristic defenses, one's beliefs have taken an empirical hammering and are sinking below the 50/50 point of total agnosticism. To maintain

one's comfort zone in this scenario, one must set one's threshold below 50/50, which requires acknowledging that one is so wedded to avoiding an error of one type that even if one views the alternatives to the unconscious-prejudice explanation as more plausible—perhaps by a large degree—one will still rely on the unconscious-prejudice explanation for policy guidance.

The threshold-adjusting class of comfort-zone-preserving maneuvers also depend on motivated reasoning: this time, of a biased utilitarian sort. Consider a market-purist who wants to set a high threshold for accepting unconscious-prejudice explanations. Biased utilitarian reasoning could help that observer bolster that value priority by:

(a) highlighting the negative consequences of false-positive accusations of unconscious prejudice, such as the risks of encouraging conspiracy theories about Whites that are already disturbingly common among African-American political elites (Simmons & Parsons, 2005), fostering a culture of victimology that some observers warn is already entrenched in African-American communities (S. Steele, 2006), imposing stultifying regimes of political correctness that can trigger reverse discrimination against Whites and high-achieving minorities (Aberson & Ettlin, 2004), and abetting frivolous class-action litigation that uses unconscious prejudice research to label vast populations of managers as presumptively prejudiced;

(b) minimizing the negative consequences of drawing false-negative conclusions (i.e., exonerating guilty organizations). Such arguments might include the risks of trivializing whatever "little prejudice" remains and exaggerating the power of market forces and existing regulatory-legal mechanisms to check prejudice.

Conversely, a state-interventionist—who believes that false-negative errors should trump false-positive errors—could put the cognitive machinery into reverse by:

(a) highlighting the seriousness of the consequences of failing to identify true prejudice, such as demoralizing hard-working members of protected categories who have been mistreated and who may infer that society does not really care;

(b) trivializing false accusations by arguing that such accusations pressure implicit bigots to rein in their bigotry, that plaintiffs rarely win in employment litigation (Selmi, 2003), and that the costs of frivolous litigation in general are minimal compared to the costs of discrimination.[21]

Tetlock's (2005) work on expert judgment suggests that most experts prefer belief-system over value-threshold defenses. Experts resorted to the "I-made-the-right-mistake" defense in real-world settings mostly after they felt—rightly or wrongly—that they had lost the factual debate over the efficacy of a particular policy (such as libeal opposition to Reagan's defense policies in the 1980s or conservative support for the invasion of Iraq in 2003). This finding makes sense in a dissonance framework: it is preferable not to admit a factual mistake than it is to admit the mistake and then defend it as the result of relying on the correct value threshold for translating beliefs into actions. If the same belief-system dynamic is at play in debates over unconscious prejudice, we should expect invocations of the "I-made-the-right-mistake" defense only when one or the other side has encountered data that are particularly hard to neutralize with the usual defenses, something that we have yet to observe—perhaps because, as noted earlier, neither side is remotely close to depleting its ample stock of hard-core-protective, auxiliary hypotheses.

---

[21] MacCoun (1998) offers a simple formula for setting value thresholds: divide the disutility of false positives by the sum of the disutility of false positives and the disutility of false negatives. This formula tells us to adopt a high threshold for accepting a claim affirming the ubiquity of unconscious prejudice to the degree one values avoiding false positive accusations of prejudice more than false-negative exonerations.

B. Beyond cursing the darkness

Thus far, we have stressed the multiplicity of belief-system and value-threshold defenses to which each side has recourse. The theme has been that proponents have too much flexibility to attribute any adverse impact to unconscious prejudice, whereas opponents have too much flexibility in challenging the validity of unconscious-prejudice research.

The traditional adversarial approach to scientific dispute resolution—with its point-counterpont exchanges in journals—makes it too easy for both proponents and opponents to hunker down in impenetrable theoretical bunkers. Each side has too many degrees of freedom for devising auxiliary hypotheses to defend its respective hard-core commitments. In our view, the best hope for resolution lies in the willingness of both sides to agree to exercise restraint in using its conceptual degrees of freedom. And that requires regime change: a shift from a closed-minded adversarial system that rewards tenaciously defending extreme positions to a system that encourages open-minded exploration of the boundary conditions that moderate when implicit prejudice is likely to influence various personnel decisions.

Accordingly, we devote this final section to exploring an alternative approach to disentangling the factual and value components of debates on unconscious prejudice. We draw on the writings of four extraordinary scientists: Robert Merton (1987) who argued that sciences advance most rapidly when they acknowledge ignorance most openly; William McGuire (1983) who urged social scientists to conceive of their work as a contextualist process of discovery rather than a deductive exercise of pinpointing the correct theory; Paul Meehl (1990) who warned against the defects in hypothesis-testing procedures in psychology; and Daniel Kahneman (2003) who developed adversarial collaboration as a method of avoiding what he saw

as increasingly pointless point-counterpoint exchanges with the many critics of his high-impact research program on judgment and choice.

Below, we describe each phase of what we project to be a four-phase process.

(1) Reflecting on the cognitive dynamics of the debate.

Insofar as each side in the debate concedes some truth to bounded-rationality models of human nature that stress the power of our preconceptions to lock us into particular views of reality (Greenwald, 1980; Tetlock, 2005), each side should acknowledge the risk of scientists in opposing camps "losing perspective" and failing to see when they have deviated from rational-actor models for belief updating. This concession carries corollaries. In principle, both sides should agree that: (a) we may all like to think of ourselves as open-minded belief updaters but our self-assessments are too generous and accountability interventions—from outside our community of co-believers—may be essential for checking violations of Bayesian norms in hypothesis testing; (b) we may all recognize the power of egocentric and in-group biases but such abstract process knowledge is no sure inoculation against these biases—and, again, accountability to skeptical outsiders may be essential for correcting lapses from scientific norms, such as ad hominem arguments that, to insiders, feel undeniably on the mark (statist interventionists often look like "victimologists" to market purists, whereas market purists often look like "racist system justifiers" to statist interventionists); (c) we may all recognize the power of hindsight bias to distort our recollection of past states of ignorance (Faust, 1984) but such abstract process knowledge is again an unreliable check on each camp's ability to explain away, ex post, inconvenient results by drawing on its store of auxiliary hypotheses—and, again, accountability to skeptical outsiders may be essential to ensure that each camp states its

expectations explicitly and ex ante.

Each proposition captures a psychological threat to the capacity of research programs to live up to the Mertonian conception of science as a social system of "organized skepticism" (Merton, 1973). Taken together, these propositions define the hard-core of a new, middle-ground research program in the American racism debate, a program guided by the ideals of adversarial collaboration and informed by the psychological realities of how science is done. Recognizing the perils of politicization, each side commits itself to resolving empirical disagreements under shared methodological ground rules; each accepts the necessity of spelling out predictions before data collection; and each appreciates the need for self-restraint in a world in which it is easy for either side to impugn the motives of the other.

(2) Mapping the limits of empirical knowledge: Epistemic audits.

The temptations to claim to know more than ones knows—or more than anyone could know—take many forms, from individual career advancement to cross-disciplinary competition for prestige (Merton, 1987). And the opportunities for getting away with over-claiming are manifold. It is hard to strip away the evaluative overlay on any data language and all the harder when the issues are as politically charged, and methodological ground rules as slippery, as in work on unconscious prejudice: When do we have epistemic warrant to call our competitors for power prejudiced? How big must the reaction-time differentials be? How awkward must the cross-racial encounters become?

For these reasons, we see the need to supplement the normal-science institutions of self-policing–peer review, methodological transparency, direct and conceptual replications–with adversarial collaboration. And we see epistemic audits as an essential starting point for

adversarial collaboration in high-policy-stakes (and thus low-trust) debates. By epistemic audit, we mean any reasonably sustained effort by skeptical outsiders to gauge the correspondence between empirical evidence and theoretical assertions. The simplest epistemic audits—like their corporate counterparts—take the "books" at face value. In social science, that means mapping the data language (as laid out in Methods and Results sections of peer-refereed articles) onto the theory language (as laid out in the Introduction and Discussion sections). Unfortunately, in the current context, suspicion is likely to trump convenience. Skeptics are likely to insist on full-scale epistemic audits in pursuit of flaws that might have eluded peer review. In social science, that means reanalyzing original data and trying to replicate key claims.

Blanton et al. (2007) do not report a full-scale epistemic audit, but they do raise three foundational challenges to the unconscious-prejudice literature: (1) questionable judgment calls in published studies cited as evidence for the ubiquity of prejudice; (2) failure of researchers to share data for reanalysis; (3) doubts about the replicability of the published literature, noting how rare exact replication is in social psychology in general, noting failures to replicate in work on unconscious prejudice, and pointing out how problematic the failure to conduct exact replications is in domains in which effect sizes are small, null-hypothesis results rarely see the light of publication day, outliers play key roles in driving aggregate results, and researchers and their assistants may have strong experimenter expectations (Rosenthal, 1968). If such criticisms are correct, skeptics could conceivably disqualify large fractions of the already published "data-language" foundation of unconscious-prejudice arguments by showing that the research community has violated key Mertonian norms of science, such as objectivity, organized skepticism, and communal sharing of data.

For current purposes, however, let us assume that these foundational challenges to unconscious-prejudice research have not yet delivered a full knock-out punch and that the recent Poehlman et al (2006) meta-analysis of the literature reveals evidence of some replicable patterns. Can we find a data language for characterizing these patterns that will be objective enough to command trans-ideological consensus—that skeptics will not dismiss as over-claiming and that proponents will not dismiss as unduly modest? We suggest that the hypothetical proponent and skeptic could profitably focus on four starting points:

(i) researchers have found low-to-moderate test-retest and internal-consistency reliability coefficients for reaction-time measures of "implicit prejudice" (results consistent with an implicit-prejudice account but also consistent with alternative psychological accounts (including familiarity, egalitarian sympathy, test anxiety, cognitive dexterity, cultural knowledge, and tacit awareness of depressing societal regularities));

(ii) researchers have documented low-to-moderate correlations between their reaction-time measures and criterion variables of debatable relevance to legally actionable discrimination (results that are consistent with an "implicit-prejudice" account and with the aforementioned alternative explanations);

(iii) the low-to-moderate correlations between reaction-time measures and criterion variables could be consistent with a widely diffused propensity in the population to discriminate or equally consistent with the view that alternative psychological processes tapped by the tests are widely diffused in the general population or equally consistent with the view that the correlations are driven by small pockets of respondents who score as outliers on either "independent" or "dependent" variables;

(iv) judging by the artificiality of the paradigms used in implicit-prejudice research and the many dimensions on which these paradigms differ from actual work settings, a long list of empirical unknowns must be addressed. We do not know how rapidly the effects of unconscious prejudice (or whatever processes reaction-times do tap) dissipate as a function of how much: (a) information observers have about individual employees; (b) training observers have received in conducting performance evaluation and coaching; (c) training observers have received about rating biases; (d) training observers have received on the importance of diversity as part of the business strategy; (e) guidance observers have received about the need to justify their ratings to key constituencies in the company.

(3) Specifying research designs to identify boundary conditions on key claims.

Let's suppose that epistemic audits reveal key unknowns that, if resolved, could induce each side to change its mind to some degree. Suppose also that adversaries can agree on the rough contours of a methodological approach to transforming the unknown into the known. Working from a contextualist premise (McGuire, 1983), they agree it is easiest to focus on external-validity challenges that readily translate into boundary-condition-interaction hypotheses in which proponents expect unconscious-prejudice effects to hold up even under adverse conditions whereas skeptics doubt that such effects can be replicated even under favorable original conditions, and insist that even if these doubts prove unfounded, the effects will prove hot-house flowers that wilt in the face of the slightest resistance—whether it take the form of training or individuating information, accountability, or intrinsic or extrinsic incentives.

The next step would be to sketch research designs that manipulate one or more of the hypothesized threats to external validity, such as individuating information about the targets of

personnel judgments and accountability pressures on the manager. Given the context specificity of many effects, it would also be necessary to specify a long list of methodological background conditions that must be satisfied for maximally diagnostic hypothesis-testing: managerial experience, the assumptions participants have about the study, the incentives they have to make optimal choices, which research manipulations are between versus within subjects, …. In brief, the research designs must sketched in sufficient detail so that the adversarial collaborators can offer reasonably confident inputs to the belief updating equation.

Assuming that neither side entrenches itself in ridiculously dogmatic prior hypotheses about the correctness of each other's positions, the key input at this juncture will be the likelihood ratios: the probability that each side assigns to a result assuming the correctness of its own position divided by the probability it assigns to the result assuming the correctness of the other side's position, $p(d \mid H_S)/p(d \mid H_P)$ (see Figure 2). Research designs with likelihood ratios close to 1.0 are the least theoretically informative because they have the least potential to budge prior-odds ratio. But likelihood ratios in which skeptics assign near-zero probability to data patterns that do materialize have the potential to devastate the skeptics' positon and likelihood ratios in which proponents assign massive probabilities to data patterns that do not materialize have the potential to devastate the proponents' position. The more sharply the likelihood ratio departs from 1.0, the more promising the study as a candidate for adversarial collaboration.

Figure 2

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Likelihood Ratio}$$

$$\frac{p(H_s|d)}{p(H_p|d)} = \frac{p(H_s)}{p(H_p)} \times \frac{p(d|Hs)}{p(d|H_p)}$$

H_s = Skeptics' Hypothesis
H_p = Proponents' Hypothesis

We find it easy to imagine studies that could induce reasonable skeptics to shift their estimates of Banaji et al.'s (2004) claim that the IAT is a reliable measure of pure prejudice. Let's stipulate that: (a) a hypothetical skeptic attaches an 80% probability to the claim being false and a 20% probability to it being true (the skeptic's prior odds ratio is thus 4:1); (b) the proposed experiment will occur in a work setting, that participants will be experienced managers trained to perform personnel functions, that participants will have incentives to take the task seriously, and that the between-subjects component of the design will be a 3 X 4 factorial that manipulates the individuating information managers have about candidates for promotion (no information, three-paragraph vignettes of performance-relevant information, and six such vignettes) and the types of accountability pressures on managers (e.g., no accountability; expectation of need to justify one's decisions to the people being judged; expectation of need to justify one's decisions to a human-resources ("HR") manager charged with ensuring the efficiency and fairness of the process; and accountability to both the persons being rated and to the HR manager). The repeated-measures component of the design would manipulate the

protected-category status of 10 possible candidates for promotion (for simplicity – we randomize order of presentation and which candidate is paired with which vignettes).[22]

We focus here on the worst possible outcomes for each side. It would unsettle the skeptics most if unconscious prejudice proved a potent predictor of bias even in the condition with the most individuating information and accountability pressure. The likelihood ratio for this data pattern in just this one experiment requires the hypothetical skeptic—who started off 80% confident that unconscious prejudice could predict nothing of consequence in realistic settings—to lower his confidence to .50–or coin-toss uncertainty (0.5). If the skeptic lost a second bet with the same terms, his confidence in his original position should fall to 0.2—a dramatic about-face rarely seen in science. The worst possible outcome for proponents of unconscious prejudice would be the failure of their measures to predict anything anywhere, even in the no-individuating information and no-accountability conditions. Proponents would then incur the mirror-image belief-updating obligations.

This task of specifying odds illustrates a benefits of adversarial collaboration that could materialize before any data are collected: greater open-mindedness. Many adversarial collaborators will find themselves moderating their beliefs as the time for setting reputational

---

[22] Such a study, even if it showed discriminatory effects of unconscious prejudice on tangible and fungible outcome variables, may not convince labor economists that discrimination has big marginal effects on wages in markets. As Heckman (1998) commented on audit studies showing discrimination by individual firms, advocates of market-based solutions to discrimination do not contend that the market eliminates all discriminators; rather, they contend that the market greatly reduces or eliminates the marginal effects of discrimination. The study proposed here addresses psychological skeptics who doubt that implicit measures of prejudice reliably predict discrimination in individual firms—a question more relevant to psychological theory (and class-action law) than to economic theory.

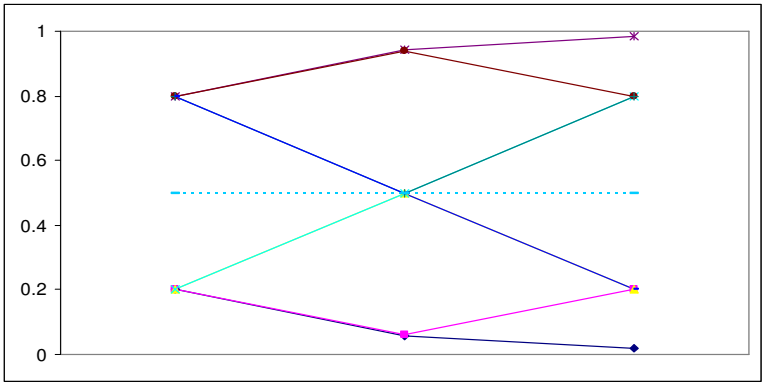
bets approaches because they know—at some visceral level—that rhetorical flourishes that play well to their community of co-believers raise the risk of losing the entire argument in a single study. But it is worth stressing that, even with more moderate wagers, scientists would be obliged to concede much more belief change under the strict Bayesian ground rules than under the lax Lakatosian ones. If the skeptics lost the bet we proposed, the norms of adversarial collaboration would preclude them from invoking such tempting post hoc auxiliary hypotheses as "the individuating-information and accountability manipulations were far weaker than in realistic work settings." Indeed, research on cognitive conservatism in belief updating suggests that, if skeptics had unrestricted access to their auxiliary hypotheses, and if skeptics were like other samples of social scientists studied, they would change less than half of the prescribed amount (Tetlock, 2005). And research on cognitive styles in belief updating suggests that if the skeptics were market-purist "hedgehogs," they might well exhibit no belief change in response to unexpected outcomes (Tetlock, 2005).

Figure 3 (below) estimates the power of adversarial collaboration to stimulate scientific progress by comparing belief change under idealized Bayesian and down-to-earth Lakatosian rules (of the sort that typically "regulate" politicized disputes). The top panel of Figure 3 illustrates the trajectories of belief change that we should expect–under Bayesian rules–given that: (a) statist interventionists and market purists either win or lose each study in a two-study adversarial collaboration; (b) each camp starts with a prior-odds ratio of 4:1 (attaching an 80% likelihood to its own hypothesis and a 20% likelihood to the other side's hypothesis) and attaches a likelihood ratio of 4:1 to their predicted patterns of data emerging given the methodology. Under Bayesian rules, the clashing camps converge rapidly—after the first study—when the data

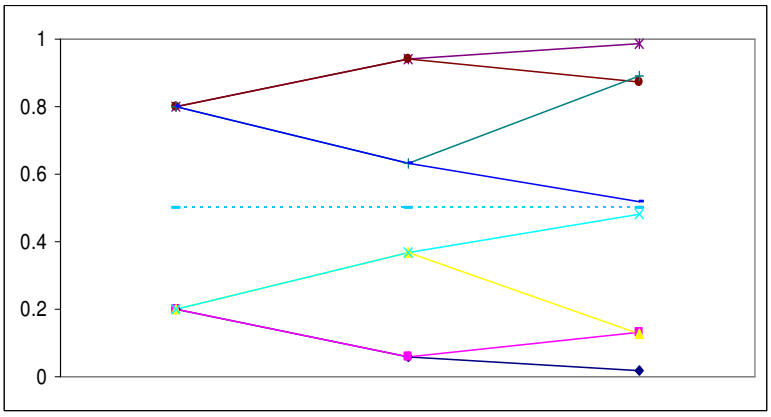
are equally dissonant for each camp. And cross-overs occur (skeptics become proponents and vice versa) if the second study yields equally dissonant data for each camp (admittedly an unlikely scenario, but logically possible if each side frames its likelihood ratios so that certain patterns of data could undercut both camps).

The bottom two panels of Figure 3 illustrate the trajectories of belief change–under Lakatosian rules–when the camps make the same predictions about the same studies and observe the same outcomes. In the middle, moderate Lakatosian panel, in which each side uses auxiliary hypotheses for protecting its hard core in a restrained fashion, participants treat confirming evidence in a Bayesian fashion but treat the disconfirming evidence in the first study as only half as diagnostic as a Bayesian would and they treat later disconfirming evidence as progressively less diagnostic ("heads we win; tails we lose by ever decreasing amounts"). We therefore see slower convergence than in the Bayesian case—and the convergence is never complete. In the extreme Lakatosian panel, each side makes full-throttle use of auxiliary hypotheses and we see no convergence (of "heads we win; tails we lose nothing"). Each side manages to maintain a substantial ideological comfort zone and to stay safely on its own side of its value-threshold line (which we have drawn here at the psychologically natural location of the midpoint of 0.5–a location that also captures legal intuitions about preponderance of proof in civil litigation).

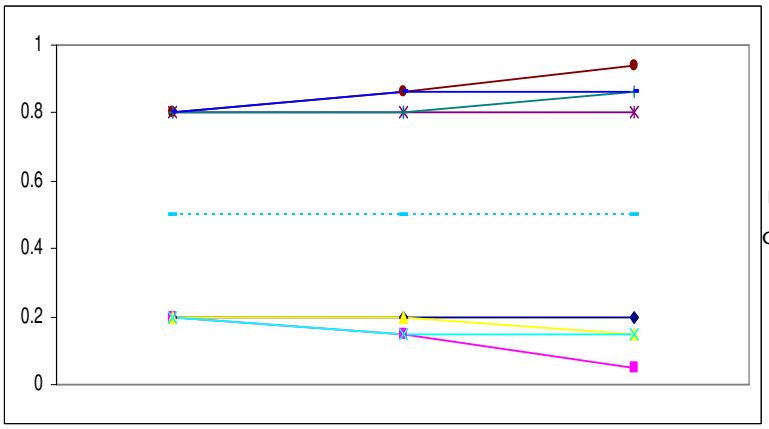
Figure 3



Bayesian - Cross-over Occurs



Moderate Lakatosian - Gradual but Incomplete Convergence



Extreme Lakatosian - No Convergence

Note: Dashed line at .5 signifies value threshold for translating belief shifts into policy shifts.

Although tedious, there are benefits from working through the diagnostic implications of the diverse data patterns that could arise in complex factorial designs. The process forces each side to be pre-emptively self-critical (which may be why Daniel Kahneman once commented that adversarial collaboration raises one's methodological IQ by a standard deviation). It is also worth noting that the data will rarely fit the picture-perfect expectations of either side. We have focused here on the starkest possibities. But nature often throws curveballs. A full-scale adversarial collaboration, involving elicitation of ex ante likelihood ratios, would require participants to think through a variety of dissonant possibilities.

(4) Setting thresholds of proof.

For adversarial collaboration to have policy as well as scientific impact, each side must make itself politically as well as scientifically vulnerable. For this reason, we see ex ante specification of value thresholds as every bit as critical as ex ante specification of prior-odds and likelihood ratios. The challenge is achieving transparency in not only the hypothesis-testing process but also in the process of translating shifts in the relative plausibility of hypotheses into policy implications. And this aspect of adversarial collaboration is especially relevant in settings such as the American racism debate, in which researchers often act as policy advocates (Kang & Banaji, 2006) and expert witnesses (Bielby, 2003) whose thresholds of proof translate directly into policy prescriptions.

The need for more transparency at the science-policy interface strikes us, frankly, as transparent in its own right. Sophisticated consumers of science know that all inductive knowledge is tentative and this is, *a fortiori*, true of social scientific knowledge (Epstein & King, 2002). Yet one rarely hears social scientists in policy debates declaring that, if a favorite

hypothesis approached or fell below a confidence threshold, they would revise particular policy stances—for statist interventionists to concede that if they lost a series of reputational bets, and their confidence in the unconscious-prejudice hypothesis fell below, say, 0.5, they would revise their blanket support for numerical racial quotas in hiring and promotion, or for market purists to concede that if their subjective probability rose above 0.5, they would modify their blanket opposition to quotas. Trade-off aversion is a well-documented cognitive and political phenomenon (Jervis, 1976; Tetlock, 1998): there are some categories of trade-offs that most of us prefer not to acknowledge either to ourselves or to others.

The usefulness of adversarial collaboration as a tool for increasing trade-off transparency strikes us as equally self-evident. By calling on participants to specify cut-off points, ex ante, where changes in their probabilistic beliefs translate into policy recommendations, adversarial collaboration brings covert trade-offs into sharp focus. The deepest of these trade-offs is the tension between partisans' desire for padded ideological comfort zones (which require setting extreme thresholds so that they are at little risk of having to change policy stands even if they suffer a string of empirical reversals) and partisans' desire to appear reasonable in the eyes of external constituencies (which requires setting moderate but empirically riskier value thresholds that show they are giving due weight to all of the competing values at stake).

Extremely safe resolutions of this canonical trade-off reveal more than most extremists want revealed—for the safer the cut-off point, the more extremists reveal their indifference to either false-positive or false-negative errors in policy debates. Imagine statist interventionists who insist on supporting quotas even if their side loses a series of reputational bets that require lowering their confidence below 0.5 for even weak forms of the unconscious-prejudice

hypothesis. Predicating a policy on an hypothesis more likely to be false than true sends a message about societal priorities: in effect, "I care so much about avoiding false-negative exonerations of organizations that I am prepared to pay a large price in false-positive accusations." Conversely, imagine market purists who continue rejecting numerical quotas even if skeptics of IAT-like measures acknowledge that losses in adversarial collaborations require raising their estimates above 0.5 for the unconscious-prejudice hypothesis. Now the message flips, "I find false-positive accusations so distasteful that I want society to pay a large price in false-negative exonerations." Indeed, as noted earlier, extremely safe resolutions of the trade-off can lead to ridiculously well padded comfort zones in which one can retain one's policy positions even if one loses all of the empirical arguments.

Of course, it is not inherently unreasonable to set extreme burdens of proof. It is reasonable if one is sure that extreme consequences are linked to a policy option. Anyone who has marched through security-clearing procedures in airports in the aftermath of 9/11 has experienced a policy that places vastly greater value on avoiding false-positive classifications of passengers as potential terrorists than on avoiding false-negatives. In the unconscious prejudice debate, however, the clashing values are more comparable in strength than momentary passenger inconvenience versus mass murder, and adversarial collaborators will risk public ridicule for their extremism (e.g., "you are willing to countenance 1000 false-positive classifications of organizations as repositories of prejudice in order to catch one guilty organization?").

That said, we know remarkably little about where participants in anti-discrimination

policy debates draw their value-threshold lines—or about how rigorously they separate their factual from their value judgments.[23] But we do know how to start specifying the parameters of our ignorance. Work on the psychophysics of subjective probabilities tells us that people often pay rather cursory attention to shifts in intermediate probabilities between 0.2 and 0.8 (Kahneman and Tversky, 1979; Tversky & Kahneman, 1991). There appears to be a large and ill-defined "maybe" wilderness, between probably false and probably true, in which motivated political reasoners can surreptitiously adjust their value thresholds so that their probabilistic beliefs never cross the line. It follows that, if we do not collectively make a special effort in adversarial collaborations to pin down these value thresholds, ex ante, true believers will have a great deal of conceptual cover for sneaky ex post threshold maneuvers that preserve their ideological comfort zones.[24]

---

[23] One telling sign of fact-value contamination is sudden spikes in resistance to disconfirming evidence whenever that evidence threatens to drag their beliefs across a value threshold.

[24] Surveying the subective-probability scale, from zero (impossible the hypothesis is correct) and inevitability (certain the hypothesis is correct), there is a vast expanse of intermediate options that imply greater tolerance for one type of error or the other. We suspect, though, that the 50/50 value represents a salient cutoff point. In adversarial collaborations, skeptics whose empirical losses compel them to assign more than 0.5 probability to unconscious-prejudice explanations cross the Rubicon into the land of proponents, whereas proponents crossing the other direction become skeptics. Here one's pet explanation changes from "more likely true" to "more likely false," the point where "we" become "them" and "they" become "us." And, in setting value thresholds, 50/50 signifies a commitment to even-handedness. Setting one's value threshold below 0.5 implies that one is prepared to predicate a policy on a belief in which one places less than even odds, setting one up for the rejoinder: how implausible does your pet hypothesis need to become before you will cease to ground public policy on it? Whatever one's choice, adversarial collaboration flushes the trade-offs into public view. One pays a price in credibility for pushing one's value threshold to extremes, but one gains the benefit of not having to change one's policy convictions as quickly if one loses a few reputational bets.

From a scientific perspective, such post hoc threshold shifting is not as troubling as post hoc Lakatosian defenses that neutralize dissonant evidence in the hypothesis-testing process. Indeed, from a positivist perspective, one might ask why one should care about scientists' values anyway? But from a policy perspective, value threshold adjustments are troublesome because they are tantamount to moving the goalposts—thereby causing policy convergence to lag behind scientific convergence . Put simply, we need sensitive threshold indicators to keep the policy debate honest. Most external observers agree that there is something odd about a research program in which repeated empirical defeats trigger value-threshold shifts but never policy shifts. It is unclear why the relative importance of avoiding false accusations of prejudice should spike when the data have just revealed that the likelihood of such errors is lower than previously supposed. Nor is it clear why the relative importance of avoiding false exonerations should spike when the data have shown that the risks of that error are lower than previously supposed. Such patterns of threshold shifting do not prove but are certainly consistent with the view that policy justification is a key epistemic driver of the research programs.[25]

Finally, we recognize how hard it will be to induce observers–be they scientists or litigators or politicians–to divulge where they set their value thresholds for false-positive and

---

[25] Value thresholds need not be qualitative breakpoints. It is more reasonable to posit a continuum in which pressure for change rises as one approaches a designated threshold—and researchers have the option of specifying these more sophisticated functional forms. Imagine that skeptics lost "big" in the adversarial-collaboration and were under a Bayesian order to increase their estimates of the other side being right by either seven-fold (from .05 to .35) or nine-fold (from .05 to .45). And imagine that only the latter magnitude of belief change pushed skeptics beyond their value threshold of .40 for opposing quotas. It would be odd if the answer were to reject any need to change for the .35 group and to insist on a complete reversal for the .45 group.

false-negative errors. In the terminology of the sacred-value-protection model (Tetlock et al. 2000), admitting that one is willing to tolerate any hint of racism or any chance of convicting the innocent feels–to most citizens–to be a transgression of a taboo (Tribe, 1971). However, explicit methods for revealing these trade-offs is a sign of rigorous science–and acceptance of the tragic necessity of balancing sacred values is a healthy sign of value pluralism (Berlin, 2002). The legal and political challenge is to reframe taboo trade-offs as tragic trade-offs.[26]

V. Closing thoughts

Our patient readers may have noticed that we never fully answered the question with which we opened: "what must organizations do to check implicit bias?" But we have not been totally delinquent. We have shown that the strong unconscious-prejudice argument—it is so potent that only de facto quota systems can keep it at bay--- has no scientific support. And we have highlighted empirical issues that, if resolved via adversarial collaborations, would tip the scales in favor of either statist interventionists (and their maximalist position that only quotas will check prejudice) or market purists (and their minimalist position that incentives to get it right—reinforced with gentle reminders to respect EEO norms—will suffice).

Venturing further opinions, prior to adversarial collaboration, would be rash. But it is worth reflecting on the pros and cons of launching adversarial collaborations at the minimalist versus maximalist ends of the debiasing continuum. We chose minimalist here. Our working

---

[26] In our philosophy of science, setting value thresholds is the job of judges and legislators, not scientists. But we do not deny that scientists have value thresholds or that these thresholds influence their work. We simply agree with Hammond and Adelman (1976): there are advantages to crisp fact-value divisions of labor in policy debates.

hypothesis was that starting minimalist—assessing the debiasing power of weak forms of accountability, individuating information, etc., and then ratcheting up—makes sense in a research community in which prejudice is widely and rightly stigmatized. Combining the psycho-logic of group polarization (no one wants to appear soft on bigotry—Myers, 1976) with that of the fundamental attribution error (observers often under-estimate the power of the situation to control behavior—Ross et al., 1977), we suspected that there may be a "rhetorical bias" among investigators to over-estimate the situational pressure needed to squelch discrimination (cf. Kaheneman & Renshon, 2007). Starting minimalist also makes sense in light of classic experiments that demonstrate the power of weak manipulations of social pressures to produce surprisingly large conformity and obedience effects (Ross et al., 1977) as well as more recent work on the power of weak debiasing manipulations–such as accountability, outcome interdependence, and the salience of egalitarian norms–to check biases (Fiske & Neuberg, 1990; Lerner & Tetlock, 1999; Tetlock, 1992). Taken together, these arguments underscore the need to establish baseline metrics for accountability and other workplace checks on prejudice and to track how effect sizes for unconscious prejudice wax and wane with these key parameters. [27]

That said, there are good but easily overlooked arguments for sometimes starting at the maximalist end. These arguments are easily overlooked because it seems trivially obvious that

---

[27] The minimalist view also garners unintended support from unlikely quarters. In numerous expert reports submitted in major lawsuits over the last decade, such as the nationwide class action against Wal-Mart for gender discrimination, Dr. William Bielby has simultaneously implied that minimalist lab forms of accountability can check prejudice (such as the work of Tetlock, 1992) and argued that many companies, including several with formal EEO procedures, fail even this minimalist test (see Bielby, 2000, 2003, 2005). These arguments could be readily tested in adversarial collaborations that start with minimalist simulations.

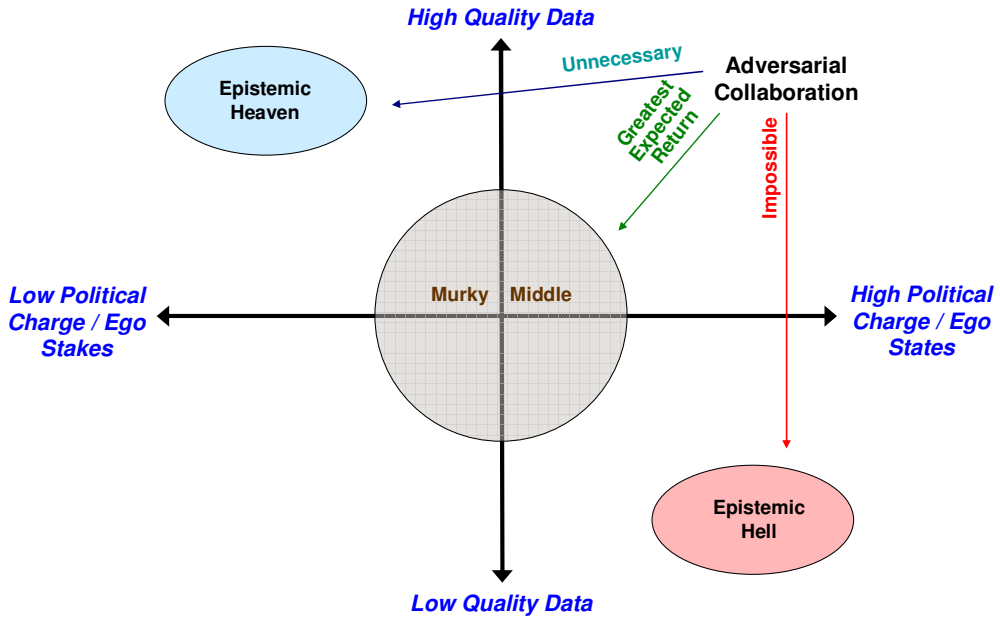
one could eliminate alleged bias against any group by simply cranking up the incentives for managers to satisfy quotas linked to labor-market availability. Even so, adversarial collaborations at the maximalist end could play a valuable role here by clarifying the price of guaranteeing equality via such dependent variables as “reverse discrimination," employee perceptions of procedural justice, and economic efficiency. There is rich reputational-bet potential here: Relative to proponents, IAT skeptics may suspect that variation in human capital is so large that quota systems raise the risk of favoring objectively less well-qualified members of protected groups--perhaps to the point of triggering resentment and impairing profitability.

It is also worth reflecting on the logically possible outcomes if adversarial collaborations were launched from both the minimalist and maximalist ends of the debiasing continuum. There are the quick knockout scenarios. Market purists win all if minimalist manipulations of accuracy incentives and accountability can check bias -- and quota forms of accountability trigger severe adverse consequences. Statist interventionists win all if minimalist manipulations fail to check bias -- and quota systems have no adverse consequences. There are also the protracted stalemate scenarios in which, at least initially, both sides lose, and adversarial collaborations settle into methodological trench warfare over how to operationalize forms of accountability that check bias at acceptable costs to other values. And, for logical completeness, there are the "both-win" scenarios. Market purists prove to be right that minimalist debiasing manipulations generally suffice and statist interventionists prove to be right that quota systems have few adverse effects.

Of course, what makes adversarial collaboration scientifically attractive—the prospect of breaking epistemic impasses—may also render it politically unattractive. Nothing will happen if either side decides that it is better off when there is less scientific clarity.

For this reason, failures to broker adversarial collaborations are profoundly informative: they signal to the policy world that the American racism debate and the sub-debate on unconscious prejudice may be politicized beyond scientific redemption. Tetlock (2006) has offered rough sociology-of-science diagnostics for judging the odds of failures of this sort. Adversarial collaboration is most feasible when least needed: when the clashing camps have advanced testable theories, subscribe to common canons for testing those theories, and disagreements are robust but respectful. And adversarial collaboration is least feasible when most needed: when the scientific community lacks clear criteria for falsifying points of view, disagrees on key methodological issues, relies on second- or third-best substitute methods for testing causality, and is fractured into opposing camps that engage in ad hominem posturing and that have intimate ties to political actors who see any concession as weakness. Tetlock (2006) calls the former community as “epistemic Heaven.” the latter "epistemic hell," and maintains—in the spirit of Figure 4—that if adversarial collaboration is indeed unnecessary in heaven and impossible in hell, we should expect the greatest expected returns in the "murky middle" in which theory-testing conditions are less than ideal but not yet hopeless.

Figure 4



The feasibility-necessity paradox raises the question of where in this two-dimensional space to locate disputes over unconscious prejudice. The signs are mixed. Sometimes we seem closer to Heaven. Each side claims to be playing by neo-positivist ground rules for judging claims. Other times, we seem closer to Hell. Each side impugns the others' motives and waves off the other side's favorite findings. Weighing the conflicting signs, we suspect that the dispute falls in the "murky middle" —which means there is roughly a 50% chance of our proposal proving quixotic. If so, we should beware of the threat to our collective credibility. Failure will reflect—and should reflect—poorly on the capacity of the social sciences to transcend ideological divisions and offer reasonably balanced advice on deeply politicized problems.

References

Aberson, C.L, & Ettlin, T.E. (2004). The aversive racism paradigm and responses favoring African Americans: Meta-analytic evidence of two types of favoritism. *Social Justice Research*, *17(1)*, 25-46.

Aberson, C.L., & Haag, S.C. (2003). Beliefs about affirmative action and diversity and their relationship to support for hiring policies. *Analyses of Social Issues and Public Policy (ASAP)*, *3*, 121-138.

Alderfer, C.P. (2003). The science and nonscience of psychologists' responses to The Bell Curve. *Professional Psychology: Research and Practice*, *34*, 287-293.

Allport, G.W. (1954). *The nature of prejudice.* Oxford: Addison-Wesley.

Ananda, S., & Gilmartin, K. (1991). Inclusion of potentially tainted variables in regression analysis for employment discrimination cases. *Industrial Relations Labor Journal*, *13*, 121 -152.

Arkes, H., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Asosciation Test?" *Psychological Inquiry, 15*(4), 257-278.

Armstrong, M. (2006). *A handbook of human resource management practice* (10th edition). London: Kogan Page.

Arthur, W., Jr., & Doverspike, D. (2005). Achieving diversity and reducing discrimination in the workplace through human resource management practices: Implications of research and theory for staffing, training, and rewarding performance. In In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 305-327). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Bagenstos, S.R. (2006). The structural turn and the limits of antidiscrimination law, *California Law Review*, *94*, 1-47.

Baltes, B.B., Bauer, C.B., & French, P.A. (2007). Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism? *Journal of Applied Psychology*, *92*, 151-164.

Banaji, M.R. (2001). Implicit attitudes can be measured. In H.L. Roediger, J.S. Nairne, I Neath & A. Suprenant (Eds.)., *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.

Banaji, M.R., Nosek, B.A., & Greenwald, A.G. (2004). No Place for Nostalgia in Science: A Response to Arkes and Tetlock. *Psychological Inquiry*, *15*, 279-310.

Bargh, J.A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, *36*, 147-168.

Bargh, J.A., & Chartrand, T.L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*, 462-479.

Barden, J., Maddux, W.W., & Petty, R.E. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology*, *87*, 5-22.

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

Baron, J. N., & Bielby, W. T. (1980). Organizational barriers to gender equality: Sex segregation of jobs and opportunities. In A. Rossi (Ed.), *Gender and the life course* (pp. 33-47). New York: Aldline.

Baron, J. N., Mittman, B. S., & Newman, A. E. (1991). Targets of opportunity: Organizational

and environmental determinants of gender integration within the California civil services. 1979-1985. *American Journal of Sociology, 96,* 1362-1401.

Bazerman, M.H., & Banaji, M.R. (2004). The social psychology of ordinary ethical failures. *Social Justice Research*, *17*, 111-115.

Becker, G.S. (1957). *The economics of discrimination*. Chicago: University of Chicago Press.

Berlin, I. (2002). *Liberty*. Oxford: Oxford University Press.

Bersoff, D.N. (1988). Should subjective employment devices be scrutinized? It's elementary, my dear Ms. Watson. *American Psychologist*, *43*, 1016-1018.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, *95*, 94-98.

Bielby, W. T. (2000). Minimizing workplace gender and racial bias. *Contemporary Sociology, 29,* 120-129.

Bielby, W.T. (2003). Can I get a witness? Challenges of using expert testimony on cognitive bias in employment discrimination litigation, *Employee Rights & Employment Policy Journal*, 7, 377.

Bielby, W.T. (2005). Applying social research on stereotyping and cognitive bias to employment discrimination litigation: The case of allegations of systematic gender bias at Wal-Mart stores. In R. L. Nelson & L. B. Nielsen (eds.), *Handbook of Employment Discrimination Research: Rights and Realities* (pp. 395-407). New York: Springer.

Bisom-Rapp, S. (1999). Bulletproofing the workplace: Symbol and substance in employment discrimination law practice. *Florida State University Law Review*, *26*, 959-1047.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*,

27-41.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P.E. (2007). Does the IAT predict discriminatory behavior? (unpublished manuscript).

Bobo, L., & Kleugel, J. R. (1993). Opposition to race-targeting: Self-interest, stratification ideology, or racial attitudes? *American Sociological Review, 58*, 443-464.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Braddock, J. H., III, & McPartland, J. M. (1987). How minorities continue to be excluded from equal employment opportunities: Research on labor market and institutional barriers. *Journal of Social Issues, 43,* 5-39.

Brauer, M., Wasel, W., & Niedenthal, P. (2000) Implicit and explicit components of prejudice. *Review of General Psychology*, *4*, 79-101.

Brief, A. P., Butz, R. M., & Deitch, E. A. (2005). Organizations as reflections of their environments: The case of race composition. In In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 119-148). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Brief, A. P., Dietz, J., Cohen, R. R., Pugh, S. D., & Vaslow, J. B. (2000). Just doing business: Modern racism and obedience to authority as explanations for employment discrimination. *Organizational Behavior and Human Decision Processes, 81,* 72-97.

Brown, (2003).

Carneiro, P., Heckman, J., & Martinov, D.V. (2005). Labor market discrimination and racial differences in pre-market factors. *Journal of Law & Economics*, *48*, 1-39.

Cascio, W.F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th

ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Chugh, D. (2004). Societal and managerial implications of implicit social cognition: Why milliseconds matter. *Social Justice Research*, *17*, 203-222.

Cleveland, J.N., Stockdale, M., Gutek, B.A., & Murphy, K.R. (2000). *Women and men in organizations: Sex and gender issues at work*. Mahwah, NJ: Lawrence Erlbaum Associates.

Coate, S., Loury, G.C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, *83*, 1220-40

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation*. Boston: Houghton Mifflin Co.

Copus, D. (2005). A lawyer's view: Avoiding junk science. In F.J. Landy (Ed.), *Employment Discrimination Litigation: Behavioral, Quantitative, And Legal Perspectives* (pp. 450-502). San Francisco: Jossey-Bass.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Science*, *7*, 173-179.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129,* 414-446.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cropanzano, R., Slaughter, J., Bachiochi, P.D. (2005). Organizational justice and black applicants' reactions to affirmative action. *Journal of Applied Psychology*, *90*, 1168-1184.

Czellar, S. (2006). Self-presentational effects in the Implicit Association Test. *Journal of Consumer Research*, *16*, 92-100.

Darley, J. M. (2001). The dynamics of authority influence in organizations and the unintended action consequences. In J. M. Darley, D. M. Messick, & T. R. Tyler (Eds.), *Social influences in ethical behavior in organizations* (pp. 37-52). Mahweh, NJ: Lawrence Erlbaum Associates.

Dipboye, R. L., & Colella, A. (Eds.). (2005). *Discrimination at work: The psychological and organizational bases*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Dobbs, M., & Crano, W.D. (2001). Outgroup accountability in the minimal group paradigm: Implications for aversive discrimination and social identity theory. *Personality and Social Psychology Bulletin*, *27*, 355-364.

Doverspike, D., Winter, J. L., Healy, M. C., & Barrett, G. V. (1996). Simulations as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance, 9,* 259-273.

Dovidio, J.F. (2001). On the nature of contemporary prejudice: The third wave. *Journal of Social Issues*, *57*, 829-849.

Edelman, L. B. (1992). Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology, 97,* 1531-1576.

Edelman, L.B., & Suchman, M. (1999). When the have's hold court: Speculations on the organizational internalization of law. *Law & Society Review*, *33*, 941-991.

Epstein, L., & King, G. (2002). The rules of inference. *University of Chicago Law Review*, *69*, 1-133.

Epstein, R.A. (1992). *Forbidden grounds: The case against employment discrimination laws*. Cambridge, MA: Harvard University Press.

Equal Employment Opportunity Commission. (Dec. 31, 2003). Best practices of private sector employers.

Faigman, D.L., & Monahan, J. (2005). Psychological evidence at the dawn of the law's scientific age. *Annual Review of Psychology*, *56*, 631-659.

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis, MN: University of Minnesota Press.

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

Festinger, L. (1957). *A theory of cognitive dissonance.* Oxford: Row, Peterson.

Fiedler, K., Messner, C., and Bluemke. M. (2006). Unresolved problems with the "I", the "A" and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*, 74-147.

Fiske, S.T. (2005). What we know about the problem of the century: Lessons from social science to the law, and back. In R. L. Nelson & L. B. Nielsen (eds.), *Handbook of Employment Discrimination Research: Rights and Realities* (pp. 59-71). New York: Springer.

Fiske, S.T., & Neuberg, S.L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol.). New York: Academic Press.

Ford, T.E., Gambino, F., Lee, H., Mayo, E., & Ferguson, M.A. (2004). The role of accountability in suppressing managers' preinterview bias against African-American sales job applicants. *Journal of Personal Selling & Sales Management*, *24*, 113-124.

Fryer, R.G., Jr. (2006). "Acting white." *Education Next*, 2006 No. 1, 53-59.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61-89). New York: Academic Press.

Gaertner, S.L., Dovidio, J.F., Nier, J., Hodson, G., & Houlette, M.A. (2005). Aversive racism: Bias without intention. In R. L. Nelson & L. B. Nielsen (eds.), *Handbook of Employment Discrimination Research: Rights and Realities* (pp. 377-393). New York: Springer.

Gelfand, M. J., Nishii, L. H., Raver, J. L., & Schneider, B. (2005). Discrimination in organizations: An organizational-level systems perspective. In In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Glaser, J., & Kihlstrom, J.F. (2005). Compensatory automaticity: Unconscious volition is not an oxymoron. In R.R. Hassin, J.S. Uleman & J.A. Bargh (Eds.), *The new unconscious* (pp. 171-195). Oxford: Oxford University Press.

Goldsmith, J., & Vermeule, A. (2002). Empirical methodology and legal scholarship. *University of Chicago Law Review*, *69*, 153-167.

Gomez-Mejia, L. R., Balkin, D. B., & Cardy, R. L. (1998). *Managing human resources*. Upper Saddle River, NJ: Prentice-Hall.

Gottfredson, L.S. (2000). Skills gaps, not tests, make racial proportionality impossible. *Psychology, Public Policy, and Law*, *6*, 129-143.

Govan, C.L., & Williams, K.D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology*, *40*, 357-365.

Greenwald, A.G. (1980). The totalitarian ego: Fabrication and revision of personal history.

American Psychologist, *35*, 603-618.

Greenwald, A.G., & Krieger, L.H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*, 945-967.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An Improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197-216.

Greenwald, A.G., Nosek, B.A., & Sriram, N. (2006). Consequential Validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist*, *61*, 56-61.

Greenwood, D.J.H. (2005). Markets and democracy: The illegitimacy of corporate law. *UMKC Law Review*, *74*, 41-104.

Gullett, C.R. (2000). Reverse discrimination and remedial affirmative action in employment: Dealing with the paradox of nondiscrimination. *Public Personnel Management*, *29*, 107-118.

Hafer, C.L., & Bègue, L. (2005). Experimental research on just-world theory: Problems, developments, and future challenges. *Psychological Bulletin*, *131*, 128-167.

Harber, K. (2004). The positive feedback bias as a response to out-group unfriendliness. *Journal of Applied Social Psychology*, 34, 2272-2297.

Harber, K. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of Personality and Social Psychology*, *74*, 622-628.

Harrison, D.A., Kravitz, D.A., & Mayer, D.M. (2006). Understanding attitudes toward affirmative action programs in employment: Summary and meta-analysis of 35 years of research. *Journal of Applied Psychology*, *91*, 1013-1036.

Hart, M. (2005). Subjective decisionmaking and unconscious discrimination. *Alabama Law Review*, *56*, 741-791.

Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives, 12,* 101-116.

Hegtvedt, K.A., Clay-Warner, J., Ferrigno, E.D. (2002). Reactions to injustice: Factors affecting workers' resentment toward family-friendly policies. *Social Psychology Quarterly*, *65*, 386-400.

Heilman, M. E., & Haynes, M. C. (2005). Combating organizational discrimination: Some unintended consequences. In In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 353-377). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Heilman, M.E., & Well, B. (2006). Disadvantaged by diversity? The effects of diversity goals on competence perceptions. *Journal of Applied Social Psychology*, *36*, 1291-1319.

Hoffman, W., Gawronski, B., & Gschwendner, T. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*,1369-1385.

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On Implicit-explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality*, *19*, 25-49.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897-913.

Hull, C.L. (1934). The concept of the habit-family hierarchy, and maze learning. Part I.

Psychological Review, *41*, 33-54.

Hunt, M. (1999). *The new know-nothings*. New Brunswick, NJ: Transaction Publishers.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*(3), 340-362.

Hunter, J.E., & Schmidt, F.L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, *2*, 447-472.

Jervis, R. (1976). *Perception and misperception in international politics*. Princeton, NJ: Princeton University Press

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33,* 1-27.

Jost, J.T., Banaji, M.R., & Nosek, B.A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25*, 881-919.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129,* 339-375.

Judge, T. A., Higgins, C. A., & Cable, D. M. (2000). The employment interview: A review of recent research and recommendations for future research. *Human Resource Management Review, 10,* 383-406.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697-720.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and*

biases: The psychology of intuitive judgment (pp. 49-81). New York: Cambridge University Press.

Kahneman, D., & Renshon, J. (2007). Why hawks win. *Foreign Policy*, Jan./Feb. 2007.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

Kahneman, D., & Tversky, A. (1991).

Kaiser, C.R., & Miller, C.T. (2001). Stop complaining! The social costs of making attributions to discrimination. *Personality and Social Psychology Bulletin*, *27*, 254-263.

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

Kang, J. (2005). Trojan horses of race. *Harvard Law Review*, *118*, 1489-1593.

Kang, J., & Banaji, M.R. (2006). Fair measures: A behavioral realist revision of "affirmative action", *California Law Review*, *94*, 1063-1118.

Kelman, M. (1991). Concepts of discrimination in “general ability” job testing. *Harvard Law Review*, *104*, 1157-1247.

Kelman, M. (2001). Market discrimination and groups. *Stanford Law Review*, *53*, 833-896.

Kidder, D.L., Lankau, M.J., & Chrobot-Mason, D. (2004). Backlash toward diversity initiatives: Examining the impact of diversity program justification, personal and group outcomes. *International Journal of Conflict Management*, *15*, 77-102.

Kinder, D. R., & Sears, D. O. (1981). Prejudice and politics: Symbolic racism versus racial

threats to the good life. *Journal of Personality and Social Psychology, 40,* 414-431.

Kinoshita, S., & Peek-O'Leary, M. (2005). Does the compatibility effect in the race Implicit Association Test reflect familiarity or affect? *Psychonomic Bulletin & Review*, *12*, 442-452.

Konrad, A.M., & Linnehan, F. (1995). Formalized HRM structures: Coordinating equal employment opportunity or concealing organizational practices? *Academy of Management Journal*, *38*, 787-820.

Krawiec, K.D. (2003). Cosmetic compliance and the failure of negotiated governance. *Washington University Law Quarterly*, *81*, 487-544.

Kunda, Z. (1999). Parallel processing of stereotypes and behaviors. In S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp. 314-322). New York: Guilford Press.

Kunda, Z., & Spencer, S.J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*, 522-544.

Latham, G. P., & Wexley, K. N. (1994). *Increasing productivity through performance appraisal* (2nd ed.). New York: Addison-Wesley.

Laudan, L. (1996). *Beyond positivism and relativism*. Boulder, CO: Westview Press.

Lerner, J., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255-275.

Lerner, M.J., & Miller, D.T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin*, *85*, 030-1051.

Lilienfeld, S.O. (2002). When worlds collide: Social science, politics, and the Rind et al. (1998) child sexual abuse meta-analysis. *American Psychologist*, *57*, 176-188.

Martin, J. (2002). *Organizational culture: Mapping the terrain.* Thousand Oaks, CA: Sage.

Matheson, K.J., Warren, K.L., & Foster, M.D. (2000). Reactions to affirmative action: Seeking the bases for resistance. *Journal of Applied Social Psychology*, *30*, 1013-1038.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-125). New York: Academic Press.

McConnell, A.R., & Leibold, J.M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435-442.

McGinley, A.C. (2005). Discrimination in our midst: Law schools' potential liability for employment practices. *UCLA Women's Law Journal*, *14*, 1-58.

McGinley, A.C. (1997). The emerging cronyism defense and affirmative action: A critical perspective on the distinction between colorblind and race-conscious decision making under Title VII. *Arizona Law Review*, *39*, 1003-1059.

McGuire, W.J. (1983). A contextualist theory of knowledge: Its implications for innovations and reform in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 16, pp. 1-47). New York: Academic Press.

Meehl, P. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108-141.

Meertens, R. W., & Pettigrew, T. F. (1997). Is subtle prejudice really prejudice? *Public Opinion Quarterly, 61,* 54-71.

Merton, R.K. (1973). *The sociology of science*. Chicago: University of Chicago Press.

Merton, R.K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance and strategic research materials. *Annual Review of Sociology*, *13*, 1-28.

Mierke, J., & Klauer, K.C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology*, *85*, 1180-1192.

Mitchell, G., & Tetlock, P.E. (2006). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal*, *67*, 1023-1121.

Monahan, J., & Walker, W.L. (1991). Empirical questions without empirical answers. *Wisconsin Law Review*, *1991*, 569-594.

Monteith, M. J., Deneen, N. E., & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and Blacks. *Basic and Applied Social Psychology, 18,* 267-288.

Munro, G.D., Leary, S.P., & Lasane, T.P. (2004). Between a rock and a hard place: Biased assimilation of scientific information in the face of commitment. *North American Journal of Psychology*, *6*, 431-444.

Nisbett, R.E. (2005). Heredity, environment, and race differences in IQ: A commentary on Rushton and Jensen (2005). *Psychology, Public Policy, and Law*, *11*, 302-310.

Nisbett, R.E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Norton, M.I., Sommers, S.R., Apfelbaum, E.P., Pura, N., & Ariely, D. (2006) Color blindness and interracial interaction: Playing the political correctness game." *Psychological Science*

17, 949-953.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2006). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265-292). Psychology Press.

Olsson, A.; Ebert, J.P., & Banaji, M.R.(2005). The role of social groups in the persistence of learned fear. *Science*, *309*, 785-787.

Paetzold, R. L. (2005). Using law and psychology to inform our knowledge of discrimination. In In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 329-351). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Paetzold, R. L., & Willborn, S. L. (2002). *The statistics of discrimination: Using statistical evidence in discrimination cases*. Rochester, NY: Thomson West.

Perkins, L. A., Thomas, K. M., & Taylor, G. A. (2000). Advertising and recruitment: Marketing to minorities. *Psychology and Marketing, 17,* 235-255.

Perloff, R., & Bryant, F.B. (2000). Identifying and measuring diversity's payoffs: Light at the end of the affirmative action tunnel. *Psychology, Public Policy, and Law*, *6*, 101-111.

Pettigrew, T.F. (1979).The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*, 461-476.

Pettigrew, T.F., & Tropp, L.R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751-783.

Plous, S. (2003). *Understanding prejudice and discrimination.* New York: McGraw-Hill.

Poehlman, T.A., Uhlmann, E.L., Greenwald, A.G., & Banaji, M.R. (2006). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Unpublished manuscript.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Potier, B. (Dec. 16, 2004). Making case for concept of 'implicit prejudice': Extending the legal definition of discrimination, *Harvard University Gazette*, available at http://www.news.harvard.edu/gazette/2004/12.16/09-prejudice.html.

Prctor, R. (1991). *Value-free science? Purity and power in modern knowledge*. Cambridge, MA: Harvard University Press.

Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S., *Philosophical Transactions of the Royal Society of London*, *53*, 370.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology* , *32*, 299-328.

Quine, W.V.O. (rev. ed. 1992). The pursuit of truth. Cambridge: Harvard University Press.

Redding, R.E. (2004). Bias on Prejudice? The politics of research on racial prejudice: Comment. *Psychological Inquiry*, *15*, 289-293.

Redding, R.E. (2001). Sociopolitical diversity in psychology: The case for pluralism. *American Psychologist*, *56*, 205-215.

Reskin, B.F. (2000). The proximate causes of employment discrimination. *Contemporary Sociology*, *29*, 319-328.

Richard, O.C., & Kirby, S.L. (1998). Women recruits' perceptions of workforce diversity program selection decisions: A procedural justice examination. *Journal of Applied Social*

Psychology, *28*, 183-188.

Riordan, C. M., Schaffer, B. S., & Stewart, M. M. (2005). Relational demography within groups: Through the lens of discrimination. In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 37-61). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Roberson, L., & Block, C. J. (2001). Racioethnicity and job performance: A review and critique of theoretical perspectives on the causes of group differences. *Research in Organizational Behavior, 23,* 247-326.

Roch, S.G. (2006). Discussion and consensus in rater groups: Implications for behavioral and rating accuracy. *Human Performance*, *19*, 91-115.

Rosenthal, R. (1968). Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. Psychological Bulletin, 70, Dec 1968. pp. 30-47.

Ross, L.D., Amabile, T.M., & Steinmutz, J.L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*, 485-494.

Roth, P.L., Huffcutt, A.I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology*, *88*, 694-706.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*, 139-165.

Rothermund, K., Wentura, D., & De Houwer, J. (2005). Validity of the salience asymmetry account of the Implicit Association Test: Reply to Greenwald, Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General, 134*, 426-430.

Rudman, L.A. (2004). Social justice in our minds, homes, and society: The nature, causes, and

consequences of implicit bias. *Social Justice Research*, *17*, 129-142.

Rudman, L.A., Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*, 743-762.

Rushton, J.P., & Jensen, A.R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235-294.

Salgado, R. (2006). Dan the xenophobe rides the A-train, or the modern , unconscious racist in "enlightened America." *American University Journal of Gender, Social Policy and the Law*, *15*, 69-110.

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial attitudes in America: Trends and interpretations, revised edition.* Cambridge, MA: Harvard University Press.

Sears, D.O. (2004). A perspective on implicit prejudice from survey research: Comment. *Psychological Inquiry*, *15*, 293-297.

Sears, D.O., & Henry, P.J. (2005). Over Thirty Years Later: A Contemporary Look At Symbolic Racism. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 95-150). San Diego, CA: Elsevier Academic Press.

Sears, D. O., Van Laar, C., Carrillo, M., & Kosterman, R. (1997). Is it really racism? The origins of White Americans' opposition to race-targeted policies. *Public Opinion Quarterly, 61,* 16-53.

Seidel, M., Polzer, J. T., & Stewart, K. J. (2000). Friends in high places: The effects of social networks on discrimination in salary negotiations. *Administrative Science Quarterly, 45,* 1-24.

Sekaquaptewa, D. Espinoza, P., & Thompson, M. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, *39*, 75-82.

Selmi, M. (2003). The price of discrimination: The nature of class action employment discrimination ligitation and its effects. *Texas Law Review*, *81*, 1249-1335.

Sesardic, N. (2000). Philosophy of science that ignores science: Race, IQ and heritability. *Philosophy of Science*, *67*, 580-.

Shelton, N., Richeson, J.A., & Salvatore, J. (2005). Ironic effects of racial bias during interracial interactions. *Psychological Science*, *16*, 397-402.

Sidanius, J., & Pratto, F. (1999). *Social dominance.* New York: Cambridge University Press.

Siegelman, P. (1999, March). Racial discrimination in "everyday" commercial transactions: What do we know, what do we need to know, and how can we find out? In M. Fix & M. A. Turner (Eds.), *A national report card on discrimination: The role of testing*. Washington, DC: Urban Institute.

Smith, J.P., & Welch, F.R. (1986). *Closing the gap: Forty years of economic progress for blacks*. Santa Monica, CA: The Rand Corp.

Smith, J.P., & Welch. F.R. (1989). Black economic progress after Myrdal. *Journal of Economic Literature*, *27*, 519-64.

Sniderman, P.M., Brody, R.A., & Tetlock, P.E. (1991). Reasoning and choice: Explorations in political psychology. Cambridge: Cambridge University Press.

Sniderman, P.M., & Carmines, E.G. (1997). *Reaching beyond race*. Cambridge, MA: Harvard University Press.

Sniderman, P. M., & Tetlock, P. E. (1986). Symbolic racism: Problems of motive attribution in political analysis: *Journal of Social Issues, 42,* 129-150.

Sniderman, P. M., & Tetlock, P. E. (1986). Reflections on American racism. *Journal of Social Issues, 42,* 173-188.

Sowell, T. (1994). *Race and culture: A world view*. NewYork: BasicBooks.

Spann, G. (2005). Neutralizing *Grutter*. *University of Pennsylvania Journal fo Constitutional Law*, *7*, 633-668.

Stanovich, K. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Earlbaum Associates.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613-629.

Steele, S. (2006). *White guilt: How blacks and whites together destroyed the promise of the civil rights era*. New York: HaperCollins Publishers.

Sternberg, R.J. (2005). There are no public-policy implicatons: A reply to Rushton and Jensen (2005). *Psychology, Public Policy, and Law*, *11*, 295-301.

Suedfed, P. (2004). Racism in the Brain; Or Is It Racism on the Brain: Comment. *Psychological Inquiry*, *15*, 298-302.

Suedfeld, P., & Tetlock, P. E. (Eds.) (1991). *Psychology and social policy.* Washington, D.C.: Hemisphere.

Suppe, F. (1977). Afterword. In F. Suppe (Ed.), *The structure of scientific theories* (2d. ed.) (pp. 617-730). Urbana, IL: University of Illinois Press.

Suzuki, L., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law*, *11*, 320-327.

Swets, J.A., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1-26.

Tetlock, P. E. (in press). Psychology and politics: The challenges of integrating levels of analysis in social science. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles.* New York: Guilford.

Tetlock, P. E., & Arkes, H. (2004). The implicit-prejudice exchange: Islands of consensus in a sea of controversy. *Psychological Inquiry, 15*(4), 311-321.

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the ideological beholder? *Administrative Science Quarterly, 45,* 293-326.

Tetlock, P. E. (1998). Social psychology and world politics. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 2) (pp. 868-914). New York: McGraw-Hill.

Tetlock, P. E. (1994). Political psychology or politicized psychology: Is the road to scientific hell paved with good moral intentions? *Political Psychology, 15,* 509-530.

Tetlock, P. E. (1994). How politicized is political psychology and is there anything we should do about it? *Political Psychology, 15,* 567-577.

Tetlock, P. E. (2006). Adversarial collaboration: Impossible when most needed? Unnecessary when most possible? Presentation to Board of directors of Russell Sage Foundation, New York City.

Tetlock, P.E., Kristel, O.V., & Elson, S.B., (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853-870.

Tetlock, P.E., & Mitchell, G. (1993). Liberal and conservative approaches to justice:

Conflicting psychological portraits. In B. Mellers & J. Baron (Eds.), *Psychological perspectives on justice*. Cambridge: Cambridge University Press.

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. In M. Zanna (Ed.), *Advances in experimental social psychology, 25* (pp. 331-376). New York: Academic Press.

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior, 7* (pp. 297-332).

Thomas, K. M. & Chrobot-Mason, D. (2005). Group-level explanations of workplace discrimination. In R. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 63-88). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tougas, F., Brown, R., Beaton, A. M., & Joly, S. (1995). Neosexism: Plus ca change, plus c'est pareil. *Personality and Social Psychology Bulletin, 21,* 842-849.

Tribe, L.H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, *84*, 1329-1393.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Econoimcs*, 1039-1061.

Van der Zee, K. I., Bakker, A. B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87,* 176-184.

Von Bergen, C. W., Soper, B., & Foster, T. (2002). Unintended negative effects of diversity management. *Public Personnel Management, 31,* 239-251.

Wang, L. (2006). *Discrimination by default*. New York: New York University Press.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naïve theories in corrections for perceived bias. *Journal of Personality and Social Psychology, 68,* 36-51.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101,* 34-52.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608-618.

Wexley, K. N., & Latham, G. P. (2002). *Developing and training human resources in organizations* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

White, R.H. (1998). De minimis discrimination. *Emory Law Journal*, *47*, 1121-1192.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Woehr, D. J., & Miller, M. J. (1997). Distributional ratings of performance: More evidence for a new rating format. *Journal of Management, 23,* 705-720.

Word, C.O., Zanna, M.P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109-120.

Uhlmann, E.L., Brescoll, V.L., & Paluck, E.L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice, *Journal of Experimental Social Psychology*, *42*, 491-499.

Vallacher, R.R., Read, S.J., & Nowak, A. (2002). The dynamical perspective in personality and social psychology. *Personality and Social Psychology Review*, *6*, 264-273.

Ziegert, J.C., & Hanges, P.J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology* , *90*, 553-562.