1-2013

Detecting and Punishing Unconscious Bias

Philip E. Tetlock
University of Pennsylvania

Gregory Mitchell
University of Virginia

Jason Anastasopoulos
University of California - Berkeley

Recommended Citation

Tetlock, P. E., Mitchell, G., & Anastasopoulos, J. (2013). Detecting and Punishing Unconscious Bias. *The Journal of Legal Studies, 42* (1), 83-110. http://dx.doi.org/10.1086/668403

Detecting and Punishing Unconscious Bias

Abstract

We present experimental results demonstrating how ideology shapes evaluations of technology aimed at detecting unconscious biases: (1) liberals supported use of the technology to detect unconscious racism but not unconscious anti-Americanism, whereas conservatives showed the reverse pattern, (2) liberals and conservatives opposed punishing individuals for unconscious bias but supported punishing organizations failing to use the technology to root out, respectively, racism or anti-Americanism, (3) concerns about researcher bias and false accusations mediated the effects of ideology on support for the technology, and (4) participants taking strong initial stands were likelier than moderates to reconsider their positions. Our findings demonstrate that there is substantial concern about penalizing unconscious bias at the individual level and that it will be difficult to generate broad support for regulation of unconscious bias at even the organizational level unless the technology is a reliable detector of unconscious biases that lead to frequent or serious antisocial behaviors.

Disciplines

Management Sciences and Quantitative Methods

Detecting and Punishing Unconscious Bias

Philip E. Tetlock, Gregory Mitchell, and L. Jason Anastasopoulos

ABSTRACT

We present experimental results demonstrating how ideology shapes evaluations of technology aimed at detecting unconscious biases: (1) liberals supported use of the technology to detect unconscious racism but not unconscious anti-Americanism, whereas conservatives showed the reverse pattern, (2) liberals and conservatives opposed punishing individuals for unconscious bias but supported punishing organizations failing to use the technology to root out, respectively, racism or anti-Americanism, (3) concerns about researcher bias and false accusations mediated the effects of ideology on support for the technology, and (4) participants taking strong initial stands were likelier than moderates to reconsider their positions. Our findings demonstrate that there is substantial concern about penalizing unconscious bias at the individual level and that it will be difficult to generate broad support for regulation of unconscious bias at even the organizational level unless the technology is a reliable detector of unconscious biases that lead to frequent or serious antisocial behaviors.

1. INTRODUCTION

Advances in psychology have moved mind reading out of the realm of science fiction and into the courts. Criminal defendants have offered brain scan evidence that supposedly reveals the truth of their claims of innocence (Brown and Murphy 2010; Shen and Jones 2011). Plaintiffs in civil employment cases, to support their claims of discrimination, have offered social science research that supposedly shows that approximately

PHILIP E. TETLOCK is Leonore Annenberg University Professor in Democracy and Citizenship, University of Pennsylvania. GREGORY MITCHELL is Mortimer M. Caplin Professor of Law and Class of 1948 Professor of Scholarly Research in Law, University of Virginia School of Law. L. JASON ANASTASOPOULOS is a graduate student in the Department of Political Science, University of California, Berkeley. This research was supported by the Mitchell Endowed Chair at the Haas School of Business, University of California, Berkeley. We appreciate the research assistance of Phillip Rescober, Thomas Laursen, and Max Schorer and the input of Dena Gromet on a draft.

[*Journal of Legal Studies*, vol. 42 (January 2013)]

75 percent of all white managers harbor unconscious biases toward African Americans and act on those unconscious biases (Anthony G. Greenwald, Expert Report, *Satchell v. FedEx Express,* N.D. Cal. 2006 [Nos. C 03-2659, C 03-2878]; Lane, Kang, and Banaji 2007; Barbara F. Reskin, Declaration, *Ellis v. Costco Wholesale Corp.,* N.D. Cal. 2006 [No. C-04–3341]). Recently, a federal district judge proposed using the primary social science tool for identifying unconscious bias, the Implicit Association Test (IAT) (Greenwald, McGhee, and Schwartz 1998), to detect undisclosed biases in potential jurors (Bennett 2010) and perhaps judges as well (Kang et al. 2012).[1]

The prospects for private and public uses of mind-reading technology extend well beyond trials. The IAT, for instance, has been adapted to identify pilots in training who are likely to take unsafe risks during emergencies (Molesworth and Chang 2009), adults and youth at risk for alcohol problems or marijuana use (Ames et al. 2007; Ostafin, Marlatt, and Greenwald 2008; Thush and Wiers 2007), persons at risk for molesting children or committing other acts of violence (Nock and Banaji 2007a, 2007b; Snowden et al. 2004; Steffens, Yundina, and Panning 2008), and persons whose explicit statements about events differ from their memories of these events (identifying those who knowingly, and even unknowingly, tell lies) (Sartori et al. 2008). Companies may take advantage of both the IAT and brain scan technology to uncover and appeal to consumers' unconscious preferences (Perkins et al. 2008; Venkatranum et al. 2012), and it is not inconceivable that brain scans will someday accompany body scans at airports.

Given their potential to affect employment, detention, and privacy, these new mind-reading tools raise important questions for legal policy makers. Most fundamental, the ability to read minds raises the specter of punishment of thought crimes and preventive incarceration of those who harbor dangerous thoughts. Such punishment and incarceration would arguably violate Mill's ([1859] 1978, p. 9) harm principle: "{T}he only purpose for which [state] power can be rightfully exercised over

[1] The Implicit Association Test compares millisecond reaction-time differences in a test taker's responses to varying combinations of stimuli. If the test taker reacts more quickly to some groups of stimuli than to others (for example, pictures of white faces paired with pleasant words versus pictures of black faces paired with the same pleasant words), then the test taker is assumed to have stronger associations with those stimuli or the stimuli are said to be more congruent with the test taker's unconscious attitudes (for example, a test taker who reacts more quickly to white faces paired with pleasant words would be said to be unconsciously biased in favor of whites and against blacks).

any member of a civilized community, against his will, is to prevent harm to others." But if anything is clear from the history of the harm principle, it is the multidimensionality of the concept of harm (Harcourt 1999). People can feel wounded in a vast array of ways: individually or collectively, cognitively or emotionally, and morally or spiritually. From the time of Mill on, battles have been fought over the proper definition of harm for regulatory purposes, and, as Harcourt (1999) notes, new scientific evidence is often used in campaigns to expand or restrict the definition of harm (for example, advocates of greater pornography regulation invoke social science studies on the psychological effects of viewing sexual and violent imagery in pornography).

The ability to detect unconscious biases presents the latest front in battles over which harms warrant state action (compare Bagenstos [2007] with Mitchell and Tetlock [2009]). The present study examines people's willingness to penalize individuals who hold unconscious biases or the organizations that employ such individuals. The research goes beyond this fundamental question to examine how different political factions react to this new mind-reading technology when it is used to address different societal problems and examines the psychological underpinnings of support for, and opposition to, the regulation of unconscious bias. We find widespread opposition to sanctions directed at unconsciously biased individuals, but we find that liberals and conservatives are willing to punish organizations that ignore the risks posed by unconscious biases among their employees depending on the nature of those risks. Different valuations of the costs associated with false accusations of bias and with failures to detect bias, and different views on the integrity of the underlying research, mediate this ideologically selective willingness to punish organizations for unconscious bias.

2. AN EXPERIMENTAL INQUIRY INTO SUPPORT FOR UNCONSCIOUS-BIAS DETECTION AS AN EMPLOYMENT SCREEN AND AS A BASIS FOR CIVIL PENALTIES

The most immediate prospect for widespread use of mind-reading technology involves employment decisions. Private employers may voluntarily use a detector of unconscious biases toward women or minorities as part of the application process, and public employers may seek to use unconscious-bias screens as an affirmative action taken to promote the interests of women and minorities (Ayres 2001). For jobs involving public safety or vulnerable populations, such as airline pilots, police officers, and child care workers, employers may use the new mind-reading tools to screen out applicants who pose unacceptable safety risks (Molesworth and Chang 2009; Steffens, Yundina, and Panning 2008). Any such use of this new technology presents an inevitable trade-off because no test can perfectly diagnose bias or predict future behavior: the benefits of screening out potential risks (of discrimination or public danger) must be weighed against the costs of falsely labeling an applicant a risk and improperly denying employment to that person. Any organization that decides that the costs of using the technology outweigh the potential benefits risks public second-guessing of that trade-off and possibly liability when the undetected risk discriminates or crashes a plane (that is, the organization faces a second-order trade-off concerning the costs and benefits associated with nonuse of the technology).

Makers of legal policy, by creating incentives or disincentives to use unconscious-bias detection tools, will influence how organizations make these trade-offs. In deciding how to design these incentives, policy makers are likely to be influenced by judgments about the reliability of the unconscious-bias detection tools (how common are type I errors of accepting false accusations of bias and type II errors of rejecting true claims of bias?) and about the costs associated with type I and type II errors (how much harm will be suffered by society or the individual when detection errors are made?).

We designed an experiment to examine the willingness of sophisticated decision makers to use unconscious bias as an employment screen and to impose penalties on companies for failing to use such a screen. Our participants were highly educated managers participating in an executive education program who had extensive experience inside large business organizations and held diverse political views. We devised the experiment to examine whether the harm principle would constrain adverse action against both the individual applicant and the organization (by imposing liability for unconscious bias or failure to prevent acts motivated by unconscious bias), and we included applications of this technology to different societal problems to examine the generality of adherence to the harm principle. In particular, we asked participants to suppose that scientists had created technologies that can reveal attitudes that people are not aware of possessing but that may influence their actions nonetheless. In the control condition, the core applications of these technologies (described as a mix of brain-scan technology and the IAT's reaction-time technology) were left unspecified. In the two treatment conditions, these technologies were to be used in ways predicted to be objectionable to either liberal or conservative observers: to screen employees for evidence of either unconscious racism (UR) against African Americans or unconscious anti-Americanism (UAA). In the former case, UR among managers posed a threat to the fair treatment of African American employees in the workplace, whereas in the latter case, UAA among workers in public safety positions posed a threat to the safe operation of the nation's airports and other vulnerable facilities.

On the basis of research into the value differences of liberals and conservatives (for example, Rokeach 1973; Schwartz 1992; Tetlock 1986), we predicted that these shifting uses of technology would provoke shifting patterns of value conflict among those who attach differential importance to civil liberties, equal employment opportunity, and national security. Absent a strong threat to a countervailing value such as equality or security, we predicted that commitment to the harm principle would constrain punitiveness and motivate opposition to the technology: it will be hard to justify a punitive stance toward people who have yet to do anything wrong—and harder still to justify such a stance toward persons portrayed less like agents endowed with free will and more like automatons enacting unconscious scripts. However, to the degree that there is a threat to a countervailing value, it should become increasingly difficult even for those who see themselves as defenders of civil liberty to justify inaction—which becomes tantamount to a stance of moral indifference to foreseeable threats to either equal employment opportunity or national security: how can anyone justify standing idly by when society would be obviously better off if preventive (albeit arguably punitive) measures were taken to stop unconscious attitudes from causing predictable harm? Thus, how defensible people deem expansion of the harm principle to cover unconscious-bias detection and sanctioning should hinge on ideological sympathies and antipathies, with liberals seeing the frequency and consequences of UR as sufficient to justify imposing costs on individuals and organizations but not in the case of UAA and with conservatives showing the opposite pattern.

In addition to the role of political values, we examined how views about the underlying science on unconscious bias, views about the prevalence of and harms associated with unconscious biases, and pressure to be consistent would affect support for use of unconscious-bias detection as an employment screen. We examined the influence of scientific views and views about unconscious bias as a social problem by gathering data on the perceived integrity of this research, the potential for misuse of the unconscious-bias detection technology, and perceptions of threats posed by unconscious bias and then used these responses as mediators in our data analyses. We examined the consistency constraint by exposing participants to an alternative use of the unconscious-bias detection technology after they had already considered its use in one setting (for example, after considering whether the technology should be used to prevent risks of UR, participants then considered whether it should be used to prevent risks of UAA). Although our participants were not political actors, they were elites whose views may mirror those of policy makers. Accordingly, understanding how our participants viewed the underlying science and potential harms and whether they could justify different outcomes in the racism domain versus the terrorism domain sheds light on how policy makers are likely to address unconscious-bias detection proposals. In sum, our study tested the following specific hypotheses.

1. As is consistent with the harm principle and psychological research on blaming (Tetlock 2002; Tetlock et al. 2007; Tetlock, Self, and Singh 2010), we predicted that few observers will deem it justifiable to take directly punitive measures against people on the sole basis of attitudes that have yet to translate into harmful acts and that people may not even be aware of possessing. However, observers will be more willing to support indirectly punitive measures that impose special compliance burdens on those who could have taken measures to prevent unconscious biases from manifesting and harming others.

2. Individual difference research on value hierarchies (for example, Rokeach 1973; Schwartz 1992; Tetlock 1986) has repeatedly found that conservatives put higher priority on the values of crime control and national security and lower priority on equality. Accordingly, we predicted that conservatives will be more willing to downplay fairness and civil libertarian qualms about invasions of privacy and false-positive labeling if they see a good chance to detect widespread unconscious attitudes linked to a tendency to harm these core values and less willing to downplay fairness and libertarian concerns on behalf of lower ranked values and more prone to mobilize counterarguments for resisting adoption of the technology, such as concerns about false-positive labeling of high scorers as racists, about an activist scientific community, and about creating an oppressive accountability regime.

3. Research on value hierarchies also indicates that liberals put higher priority on equality and remedying past collective wrongs and lower priority on crime control and national security. Accordingly, we predicted that liberals will be readier to downplay civil libertarian qualms

about invasions of privacy and false-positive labeling if they see a good chance to detect widespread unconscious attitudes predictive of a tendency to harm these core values, and less willing to downplay fairness and libertarian concerns on behalf of lower ranked values and more prone to mobilize counterarguments for resisting adoption of the technology, such as concerns about false-positive labeling of high scorers as terrorist threats, about an activist scientific community, and about creating an oppressive accountability regime.

4. Research on political attitudes indicates that many people are hard-to-classify moderates who do not fit the ideological ideal-type templates of liberals or conservatives (Kinder 1998; Sniderman and Tetlock 1986). These respondents will be more consistent in their stances toward harm-expansion arguments.

5. People need socially acceptable rationales for unfamiliar and potentially controversial decisions (Tetlock et al. 2007; Tetlock, Self, and Singh 2010), and, depending on the subculture, these rationales are likely to include ontological justifications (claims about the pervasiveness of undesirable unconscious attitudes), epistemic justifications (claims about the objectivity of the research community), and ethical justifications (claims about the relative dangers of false-positive versus false-negative classifications of people). It follows that the more one's ideological outlook predisposes one to see false-positive attributions as more serious than false-negative attributions, the more it predisposes one to see undesirable unconscious attitudes as pervasive; the more it predisposes one to be suspicious of the scientists, the more that outlook should predict opposition to societal applications.

6. Asking people questions that highlight the reputational risk of harboring double standards activates a reflective mind-set in which people balance the need to appear consistent ("I am not a hypocrite") against their affinity for one application of the technology over the other (Tetlock 2002). We predicted that people who embraced the first-presented application (strong liberals and conservatives who respectively welcomed advances in detecting UR and UAA) would feel consistent pressure to adopt the same technology in the service of a less congenial cause. They will then have three options for the reduction of value conflict: accept an unacceptable application, defend a double standard by explaining why one application is more acceptable than the other, or reconsider their support for the previously more congenial application. All three options are possible in a value-pluralism framework (Tetlock 1986), but in the special circumstances of this experiment, we predict an exception

to the generalization that those at the political extremes will be most unwilling to reconsider their positions: in the absence of readily accessible reasons for justifying a double standard, respondents on the left and right who have just accepted the first application and then confront a distasteful second application should find reconsideration of the first application the most attractive option.

3. METHOD

3.1. Participants

Ninety-five managers ($M_{\text{age}} = 34$; 64 men, 31 women) from executive or MBA programs at the University of California, Berkeley, participated voluntarily for no compensation or course credit.

3.2. Materials and Procedure

Participants first provided demographic information and placed themselves on a 9-point liberalism versus conservatism self-identification scale (1 = strongly liberal in the conventional sense of the term, 5 = moderate, and 9 = strongly conservative in the conventional sense of the term). Participants also rated their agreement with the following value statements on a 9-point scale (1 = strong disagreement, 5 = uncertainty, and 9 = strong agreement): (1) “I value social equality and support stronger measures to reduce poverty and discrimination” (egalitarianism); (2) “I value social equality but I am wary of policies that sacrifice individual rights to achieve equality” (libertarian constraint on egalitarianism); (3) “I value national security and support moving much more proactively against these threats” (national security); (4) “I value national security but I am wary of policies that sacrifice individual rights to achieve security” (libertarian constraint on national security).

3.2.1. Experimental Manipulation. Participants were then randomly assigned to one of three experimental conditions representing different intended uses of a new technology for measuring unconscious biases: participants in the control scenario reacted to a description of the new technology that mentioned no specific intended application, participants in the UR scenario judged the same technology but learned that its primary application was for detecting unconscious bias against African Americans by employers, and participants in the UAA scenario judged the same technology but learned that its primary application was for detecting UAA among employees in sensitive jobs.

The control group scenario informed participants that "[c]ognitive neuroscientists have long suspected that human behavior is much less under conscious control than many human beings think. They have now developed a new method of testing this hypothesis—and for measuring unconscious attitudes that people are not even aware of possessing." The technology was described as involving "measures based on a statistical combination of two types of data: data derived from functional MRI of the brain and from millisecond-reaction-time differentials in how rapidly people respond to stimuli flashing across computer screens." Participants were also told that in "follow-up work testing the validity of their measures, the researchers have found evidence that job-relevant unconscious attitudes (such as general dislike of employers) are widespread in the population and that scores on these measures of unconscious attitudes have the power to predict actual behavior, not just 'brain waves.'"

The UR scenario was identical to the control scenario except that the technology was described as detecting unconscious prejudicial attitudes among European Americans: "In follow-up work testing the validity of their measures, researchers have found evidence that unconscious prejudices against African Americans are widespread in the population and that scores on these measures of unconscious attitudes have the power to predict actual behavior, not just 'brain waves.'" The UAA scenario was identical to the control scenario except that the technology was described as detecting unconscious–anti-American attitudes among American Muslims: "In follow-up work testing the validity of their measures, researchers have found evidence that unconscious–anti-American attitudes are widespread among American Muslims and that scores on these measures of unconscious–anti-American attitudes have the power to predict actual behavior, not just 'brain waves.'"

3.2.2. Dependent Measures. After reading one of these three scenarios, participants indicated their level of agreement with the following statements on 9-point scales (unless otherwise noted, 1 = strong disagreement, 5 = somewhat agree, and 9 = strong agreement):

Misuse Potential. "All technologies can, of course, be abused. Do you agree that this technology has unusually serious potential to be abused?"

Scientific Value. "Do you agree that this technology has potentially great scientific value?"

Perceptions of Pervasiveness. "The researchers are probably right about the pervasiveness of unconscious prejudice against African Americans among European Americans." (In the UAA condition, participants were asked about unconscious–anti-American attitudes among American Muslims.)

Harm Principle. "Taking legal action against individuals on the sole basis of claims about their unconscious attitudes (not their behavior) would be unacceptable."

Researcher Bias. "The scientists doing this research may have a political agenda that is biasing their work."

Appropriate Use. "Society should use this technology to ensure that managers with unconscious prejudice against African Americans [unconscious anti-American attitudes] are prevented from making harmful decisions."

False-Positive versus False-Negative Attributions. "Which error do you see as more serious: an employer who concludes that someone has an unconscious prejudice against African Americans [anti-American attitude] when that person does not VERSUS an employer who fails to identify someone who really does have an unconscious prejudice against African Americans [an unconscious anti-American attitude]?" (1 = the first error is far more serious, 5 = the two errors are equally serious, and 9 = the second error is far more serious).

Failure to Use the Technology. "Imagine that a company refused to use the technology to screen its employees to ensure that they did not have high scores on the measure of unconscious prejudice against African Americans [anti-American attitudes]. As a result, a manager who would otherwise have been screened out was in a position to make flawed decisions that damaged the careers of African-American employees [was responsible for a security lapse that led indirectly to an accidental death]. How appropriate is it to increase the damage award against the company for not using the screening test?" (1 = extremely inappropriate, 5 = somewhat appropriate, and 9 = extremely appropriate).

Reflection on Initial Opinions. (*a*) Participants in the UR condition were asked if they would change their support for the technology if it were used to detect UAA among managers making sensitive national security decisions, and participants in the UAA condition were asked if they would change their level of support for the technology if it were used to detect UR against African Americans (1 = much less support, 5 = exactly the same support, and 9 = much more support). (*b*) Participants in the control condition were asked whether they would change their support if the technology were used to detect UR and if it were used to detect UAA. (*c*) Participants in all conditions were asked, "[L]ooking back at your

Table 1. Correlation Matrix for Ideology Questions

	(1)	(2)	(3)	(4)	(5)
Ideology self-report	1.00				
Egalitarianism	−.83	1.00			
Libertarianism-egalitarianism balancing	−.67	.58	1.00		
National security	.65	−.61	−.44	1.00	
Libertarianism–national security balancing	.58	−.58	−.19	.6	1.00

answers, do you think you were initially too eager to embrace or too quick to reject use of the technology?" (1 = I was too eager to embrace use of the technology, 5 = I wouldn't change any judgments, and 9 = I was too quick to reject use of the technology).

After answering the questions for dependent measures, participants were debriefed, and the experimental session ended.

4. RESULTS

The correlations in Table 1 show that self-identified conservatives (on the ideology scale) were traditional in orientation (attaching lower value to equality and higher value to national security), whereas liberals were social democratic in orientation (displaying mirror-image priorities). We subjected the ideology and five value scales to a maximum likelihood factor analysis with oblimin rotation, and the first factor accounted for 74 percent of the variance, with the following variable loadings: ideology (.91), egalitarianism (−.84), libertarian constraint on equality (−.74), national security (.64), and libertarian constraint on national security (.47). As these loadings imply, negative scores indicate conservative value priorities (higher on national security and lower on equality), whereas positive scores indicate liberal value priorities (lower on national security and higher on equality). The average and median scores on this factor were .00 and −.09. Figure 1 shows the overall distribution of scores: the center of political gravity in the sample was centrist, with roughly equal numbers of participants falling to the left and right of that cluster. Scores on this ideology factor served as the composite measure of ideology in the analyses that follow.

Table 2 presents the means and standard deviations for all of the dependent measures. We ran a set of three ordinary least squares (OLS) and ordered probit regressions for each dependent variable that tested the main-effect and ideology-by-use hypotheses while controlling for gender and age (Green 2009). We focus on the OLS results; the probit

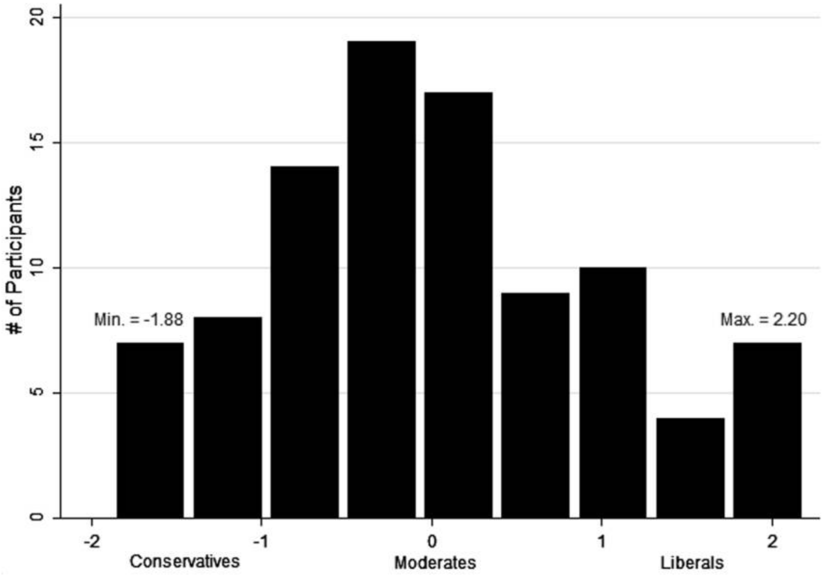

Figure 1. Liberal/conservative factor score distribution

regressions produced similar results, which demonstrate the robustness of our results across metric assumptions about the dependent variables (Cameron and Trivedi 2005). Table 3 reports the key OLS and ordered probit findings.

4.1. Test of Hypothesis 1

Consistent with the harm principle, there was near unanimity across conditions that it was unacceptable to take legal action against individuals on the sole basis of their unconscious attitudes ($M_{\text{control}} = 7.93$; $M_{\text{racism}} = 8.15$; $M_{\text{anti-Americanism}} = 8.22$; $F(2, 92) = 1.18$; $p = .31$). Thus, there was general support for the harm principle when the punitive action toward those with undesirable unconscious attitudes would be direct. There was also general opposition to imposing greater damages on companies that considered but rejected use of the technology to screen out managers with undesirable attitudes where that technology might have prevented harm ($M_{\text{control}} = 2.3$; $M_{\text{racism}} = 2.09$; $M_{\text{anti-Americanism}} = 2.25$; $F(2, 92) = .32$; $p = .73$). Participants were more accepting of the proposition that society should use the technology to seek to prevent

Table 2. Summary Statistics by Experimental Condition

	All (N = 95)		Control (n = 30)		Racism (n = 33)		Anti-Americanism (n = 32)	
Dependent Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Misuse potential	5.12	1.10	4.97	1.19	5	1.03	5.38	1.07
Scientific value	5.11	1.12	5.3	.99	5.09	.98	4.94	1.37
Perceptions of pervasiveness	5.07	1.43	4.9	1.52	5	1.52	5.31	1.26
Harm principle	8.11	.76	7.93	.91	8.15	.67	8.21	.71
Researcher bias	4.84	1.21	4.83	.99	4.76	1.32	4.94	1.32
Appropriate use	4.88	1.15	4.8	.81	4.88	1.39	4.97	1.18
False positive versus false negative	4.87	1.14	4.97	.76	4.67	1.31	5	1.24
Failure to use technology	2.21	1.08	2.3	.99	2.09	1.07	2.25	1.19
Reflection on initial opinions:								
Change support: unconscious-racism condition	4.84	.88	4.8	1	N.A.	N.A.	4.81	.69
Change support: anti-Americanism condition	5.00	.74	5.07	.64	4.94	.83	N.A.	N.A.
Too quick to embrace or reject technology	4.69	.88	4.93	.74	4.69	.95	4.47	.95

Note. N.A. = not applicable.

Table 3. Ideology by Experimental Condition Contrasts

	OLS Beta Weights	OLS Coefficients		Probit Coefficients		
	Ideology × Racism	Ideology × Anti-Americanism	Ideology × Racism	Ideology × Anti-Americanism	Ideology × Racism	Ideology × Anti-Americanism
DV1	1.30**	−.55	.54**	−.23	.61**	−.29
DV2	−.35	1.31**	−.15	.57**	−.15	.60**
DV3	−1.28**	.75*	−.70**	.42*	−.60**	.37*
DV4	.01	−.36	0	−.11	.01	−.17
DV5	.60	−1.52***	.28	−.72***	.26	−.73***
DV6	−1.19**	−.43	−.52**	−.19	−.50**	−.19
DV7	−.83*	1.49***	−.36*	.66***	−.38*	.78***
DV8	−1.11**	1.08**	−.46**	.45**	−.67**	.51**

Note. One regression was run on each dependent variable (DV) with the control as the reference group: DVn = $a + b_{1\text{ ideology × racism}} + b_{2\text{ ideology × anti-Americanism}} + b_{3\text{ race}} + b_{4\text{ anti-Americanism}} + b_{5\text{ gender}} + b_{6\text{ age}} + b_{7\text{ ideology}}$; $N = 95$ and df = 87 for each model. OLS = ordinary least squares.

*$p < .05$.

**$p < .01$.

***$p < .001$.

managers with undesirable attitudes from making harmful decisions ($M_{\text{control}} = 4.8$; $M_{\text{racism}} = 4.88$; $M_{\text{anti-Americanism}} = 4.97$; $F(2, 92) = .17$; $p = .85$).

Thus, participants' responses in the aggregate were consistent with the harm principle's constraint on direct punitive action, and this constraint seemed to extend even to indirect action punishing employers who failed to screen out managers and employees with potentially harmful unconscious attitudes. However, these group averages conceal considerable individual differences by political ideology within the different experimental conditions, as we discuss in Section 4.2.

4.2. Tests of Hypotheses 2, 3, and 4

As predicted, the correlations between ideology and support for use of the unconscious-bias detection technology shifted as a function of which political values the technology purportedly sought to protect. Using the control group as the baseline, we found that when the purported goal was to identify unconscious bias against African Americans, conservatives were more likely to see serious misuse potential ($\beta_{\text{ideology} \times \text{racism}} = 1.30$; $t(87) = 3.45$; $p < .01$), to be skeptical of researchers' claims about the pervasiveness of these negative unconscious attitudes ($\beta_{\text{ideology} \times \text{racism}} = -1.28$; $t(87) = -3.43$; $p < .01$), to view false-positive classifications of people as prejudiced as the more serious error ($\beta_{\text{ideology} \times \text{racism}} = -1.19$; $t(87) = -2.95$; $p < .01$), to oppose using the technology in routine business operations ($\beta_{\text{ideology} \times \text{racism}} = -.83$; $t(87) = -2.34$; $p < .05$), and to oppose increasing the civil liability of companies that reject using the technology, even though using the technology could have prevented harm ($\beta_{\text{ideology} \times \text{racism}} = -1.11$; $t(87) = -3.06$; $p < .01$).

By contrast, when the purported goal was to identify UAA among American Muslims, the ideology × treatment coefficients reversed signs in many instances. Although liberals were not more likely to see serious misuse potential ($\beta_{\text{ideology} \times \text{anti-Americanism}} = -.55$; $t(87) = -1.45$; $p > .05$), they were more skeptical that the technology had much scientific value ($\beta_{\text{ideology} \times \text{anti-Americanism}} = 1.31$; $t(87) = 3.30$; $p < .01$), more skeptical of researchers' claims about the pervasiveness of these negative unconscious attitudes ($\beta_{\text{ideology} \times \text{anti-Americanism}} = .75$; $t(87) = 2.01$; $p < .05$), more suspicious that the scientists have a political agenda ($\beta_{\text{ideology} \times \text{anti-Americanism}} = -1.52$; $t(87) = -4.05$; $p < .001$), more opposed to using the technology in routine business operations ($\beta_{\text{ideology} \times \text{anti-Americanism}} = 1.49$; $t(87) = 4.22$; $p < .001$), and more opposed to increasing the civil liability of companies that reject the technology, even though using it could have prevented harm ($\beta_{\text{ideology} \times \text{anti-Americanism}} = 1.08$; $t(87) = 2.97$; $p < .01$).

To test hypothesis 4, we assessed the degree to which these effects were driven by participants with strong ideological sentiments. We performed a tertile split of participants' scores on the left-right factor from the maximum likelihood analysis and then created a "supportive of the technology" index by averaging perceptions of the value of the technology and support for applications of the technology. This analysis revealed that, whereas liberals and conservatives showed full-fledged preference reversals in their support for the unconscious–mind-reading technology, moderates showed no shift in support for the technology as a function of its intended use. Liberals supported the technology when it was aimed at unconscious prejudice, but conservatives did not ($M_{\text{liberals}} = 5.75$ versus $M_{\text{conservatives}} = 4.23$; $t(16) = 4.45$; $p < .001$); conservatives supported the technology when it was aimed at anti-Americanism, but liberals did not ($M_{\text{conservatives}} = 5.9$ versus $M_{\text{liberals}} = 4.14$; $t(19) = -4.29$; $p < .001$); moderates showed moderate support for use of the technology across conditions ($M_{\text{anti-Americanism}} = 4.91$ versus $M_{\text{racism}} = 4.90$ versus $M_{\text{control}} = 5.15$; $F(2, 28) = .58$; $p = .57$).

4.3. Test of Hypothesis 5

We predicted that error aversion (preference to avoid false-positive over false-negative attributions), perceptions of the pervasiveness of bias in the general population, and perceptions of researcher bias would mediate the effects of participants' ideology on support for the technology. To test these ideas, we ran a series of OLS mediational analyses. Our dependent variable was a two-item composite indicator of support for applications of the technology: the average of the within-subject responses to questions on using the technology as an employment screen and on increasing civil liability for company failure to use the technology ($r = .44$). As can be seen in Table 4, this mediation analysis revealed that when the goal of the technology is to detect unconscious bias against African Americans, attribution error aversion preferences fully mediated the relationship between ideology and support for the technology, and perceptions of researcher bias played virtually no mediating role. Further analysis revealed that the error aversion result is due to liberal respondents (but not moderate or conservative respondents) adjusting their error aversion preferences when the goal of the technology is to detect unconscious bias against African Americans. By contrast, when the goal of the technology is to detect anti-Americanism among Muslims, the

Table 4. Mediation Analyses Using Error Aversion Preferences and Researcher Bias

	Racism Condition		Anti-Americanism Condition	
	With Mediator	Without Mediator	With Mediator	Without Mediator
Error aversion preferences as mediator:				
Degrees of freedom	57	56	56	55
R^2	.37	.48	.52	.52
Ideology × treatment	−1.20[a]**	−.67[b]	1.78[a]***	1.78[b]***
Ideology	−.11	−.15	−.08	−.08
Error aversion preferences		.38**		−.01
Researcher bias as mediator:				
Degrees of freedom	57	56	56	55
R^2	.37	.55	.52	.61
Ideology × treatment	−1.20[a]**	−.92[b]**	1.78[a]***	1.13[b]**
Ideology	−.11	−.01	−.08	−.01
Scientists conducting research are biased		−.47***		−.38**

[a] Unmediated β.

[b] Mediated β.

$^{**}p < .01$.

$^{***}p < .001$.

mediation analysis revealed less support for the mediational hypotheses: only perceptions of researcher bias emerged as a significant mediator, partially mediating the relationship between ideology and support for applying the technology.[2]

[2] All of the mediation results were confirmed by Sobel tests and 95 percent confidence intervals from bootstrapped resamplings of the indirect effect (a1 × b1). The Sobel test has become the de facto standard for mediation in social psychology, but psychometricians have warned that, although the standard errors for each coefficient are accurate as long as regression assumptions are met, the standard errors for interaction coefficients in the Sobel test are somewhat unstable, especially for smaller samples (Shrout and Bolger 2002; Preacher and Hayes 2004; Zhao, Lynch, and Chen 2010). We therefore used both tests to ensure that our findings were robust. Results of the Sobel tests and bootstrapping of the a1 × b1 interaction with 5,000 replications consistently yielded strong evidence of mediation in the unconscious-racism condition for false-negative or false-positive balancing (DV 6 in Table 3) but not for perceptions of researcher bias (DV 5) using the Sobel test (DV 6: Sobel $z = -2.35, p < .05$; DV 5: Sobel $z = -1.33, p > .10$) and using the bootstrapped a1 × b1 interaction (DV 6: 95 percent confidence interval [CI] [−.33, −.03]; DV 5: 95 percent CI [−.27, .05]). In the unconscious–anti-Americanism condition, however, strong evidence of mediation was found for perceptions of researcher bias but not false-positive or false-negative balancing using the Sobel test (DV 6: Sobel $z = .11, p > .10$; DV 5: Sobel $z = 2.80$, $p < .01$) and using the bootstrapped a1 × b1 interaction (DV 6: 95 percent CI [−.04, .07]; DV 5: 95 percent CI [.07, .37]).

To explore why error aversions were such a powerful mediator, we classified participants (on the basis of their responses on the error aversion dependent measure) as showing greater concern about false positives, approximately equal concern about both errors, or greater concern about false negatives. We then calculated mean levels of support for policy applications for these three groups across the control, UR, and UAA conditions. This analysis revealed little support in the control condition for applications of the technology across participants. But in the UR condition, participants concerned with false negatives (letting racism go undetected) were more supportive of the technology than were those in the control group ($M_{\text{racism}} = 5.25$ versus $M_{\text{control}} = 3.5$; $t(10) = 3.95$; $p < .01$), whereas those concerned with false positives (false accusations of bias) were less supportive ($M_{\text{racism}} = 2.83$ versus $M_{\text{control}} = 3.63$; $t(18) = 2.88$; $p < .05$). Those equally concerned about false positives and false negatives were statistically indistinguishable in the UR and control conditions ($M_{\text{racism}} = 3.30$ versus $M_{\text{control}} = 3.53$; $t(29) = 1.02$; $p > .10$).

Given that error aversions were a strong mediator of the ideology × racism coefficient, ideology must be correlated with false-positive or false-negative preferences. This correlation could arise in two ways: false-positive or false-negative preferences influence ideology that, in turn, influences shifting support across contexts for the technology or ideology influences false-positive or false-negative preferences. As noted above, we find strong evidence for the second explanation, and rejecting the first explanation is straightforward. If error aversions influence ideology, the following would be true: ideology and error aversion preferences would be strongly correlated within all three conditions, and error aversion preferences would mediate the ideology × condition coefficients for both the UR condition and the UAA condition. But the correlations between ideology and error aversions vary across conditions: weakly positive in the control condition ($r = .18$), strongly negative in the UR condition ($r = -.55$),[3] and weakly negative in the UAA condition ($r = -.15$). By contrast, if ideology influences context-specific error aversions, we would observe the pattern of correlations discussed earlier. These findings suggest that, in the race domain, liberals minimized cognitive dissonance by shifting other preferences to conform to their ideology, whereas conservatives were less prone to do this in the national security context, perhaps because national security was less salient to

[3] This strong negative correlation indicates that more liberal respondents were more concerned with false negatives and more conservative respondents with false positives.

this subgroup of conservatives in this context or because liberals and conservatives respond differentially to trade-offs of this sort.

4.4. Test of Hypothesis 6

We predicted that a subject's initial position would constrain later ones but that people would abandon initial positions if consistency pressures required them to embrace an application that they would strongly prefer to reject. To test this prediction, we examined reactions to new uses of the technology in four contrasts: from no specified use (control condition) to use for UR detection, from the control condition to use for UAA detection, from use for UR detection to use for UAA detection, and from use for UAA detection to use for UR detection.

In the switch from the control condition to UR detection, we find a significant liberal versus conservative crossover in which liberals offered more support for the technology on knowing its intended use (using the tertile split on the ideology factor, $M_{\text{liberals}} = 5.7$ versus $M_{\text{conservatives}} = 4.08$; $t(19) = 4.26$; $p < .001$). But when the use switched from UAA detection to UR detection, support among liberals and conservatives did not differ ($M_{\text{liberals}} = 4.44$ versus $M_{\text{conservatives}} = 4.80$; $t(14.38) = -1$; $p = .33$). Similarly, when the application switched from detecting UR to detecting UAA, support among liberals and conservatives was indistinguishable ($M_{\text{liberals}} = 4.83$ versus $M_{\text{conservatives}} = 4.83$; $t(19.6) < 1$; $p = .67$). The disappearance of a robust between-conditions effect when information from the other conditions becomes known is suggestive of an anchoring or consistency-pressure effect: initially judging a technology linked to an unpalatable application for liberals or conservatives made the technology undesirable to those groups, even when the application shifted to causes that those groups support in isolation.

To assess the impact of considering alternative applications on willingness to reconsider initial support for the technology, we ran regressions exploring the relationship between ideology and interest in reconsidering initial support. Our analysis revealed that considering potentially dissonant applications in the UR and UAA conditions caused liberals and conservatives, respectively, to reassess their views of the technology. Using the control group baseline, we found evidence that, in the UR condition, liberals were likelier to believe that they were too quick to embrace the technology and conservatives were likelier to say that they were too quick to reject it ($\beta_{\text{ideology} \times \text{race}} = .76$; $t(87) = 1.94$; $p < .10$). In the UAA condition, we found the opposite: conservatives believed that they were too quick to embrace the technology and liberals

Table 5. Too Eager to Embrace or Reject Technology by Extremity of Ideological Commitment

	Extremists			Nonextremists		
	N	Mean	SD	*N*	Mean	SD
Control	12	5.08	1	18	4.83	.5
Racism	13	4.46	1.2	20	4.85	.6
Anti-Americanism	12	3.75	1.14	20	4.9	.5

Note. Responses below 5 indicate that the participant was too eager to embrace the technology; responses above 5 indicate that the participant was too quick to reject the technology.

believed that they were too quick to reject it ($\beta_{\text{ideology} \times \text{anti-Americanism}} = -.86$; $t(87) = -2.18$; $p < .05$).

To test the ideologue hypothesis, which posits that extremists would be less open to changing their minds, we created a relative extremism dummy variable based on the distribution of ideology scores. Extremists were defined as those who scored at the right or left extremes on the ideological self-identification scale. We included this extremism dummy variable in a regression equation, included the covariates in the previous regression, and used the control group as the baseline. Contrary to the ideologue hypothesis, in both the UR and UAA conditions, extremists were likelier than nonextremists to conclude that they had been too eager to embrace the technology ($\beta_{\text{extremists} \times \text{race}} = -.44$, $t(84) = -2.28$, $p < .05$; $\beta_{\text{extremists} \times \text{anti-Americanism}} = -.55$, $t(84) = -3.79$, $p < .001$). We should, however, be careful not to overinterpret this result. A simple argument of regression toward the mean would predict on purely statistical grounds that extremists should, on second judgment, be more likely than moderates to move toward the sample average. After all, moderates already occupy that ground close to the sample average. Table 5 presents descriptive statistics for extremists and nonextremists on this measure of willingness to second-guess initial responses.

5. DISCUSSION

Our results underscore how easily a new evidence-gathering technology that could radically expand the reach of the law can become politicized. When we examined how participants in the UR-detection condition and the UAA-detection condition responded to the technology in comparison to participants in the control condition (in which no use was specified), we found strong relationships between political ideology and perceptions of the misuse potential of the technology, of the scientific significance of the technology, and of the objectivity of the scientific community linked to the technology. Liberals were consistently more open to the technology, and to punishing organizations that rejected its use, when the technology was aimed at detecting UR among company managers; conservatives were consistently more open to the technology, and to punishing organizations that rejected its use, when the technology was aimed at detecting UAA among American Muslims.

This ideologically selective willingness to apply the technology in punitive ways was fully mediated by valuations of the relative seriousness of false positives versus false negatives in the domain of discrimination and partially mediated by heightened skepticism toward the scientific community that produced the technology in the national security domain. This pattern indicates that people play favorites and draw on well-defined ideological scripts to justify that favoritism (Kunda 1990; Tetlock 2002): the justifications may take the form of ontological justifications (claims about the pervasiveness of this or that type of threat to the social order), epistemic justifications (claims about the objectivity or lack of objectivity of scientific communities), or ethical justifications (claims about the relative dangers of either false-positive or false-negative classification errors).

There were differences, however, in the mediators of technology opposition. Liberal participants were reluctant to raise concerns about researcher bias as a basis for opposition, a reluctance consistent with MacCoun and Paletz's (2009) finding that citizens tend to believe that scientists hold liberal rather than conservative political views. If scientists are expected to be liberals, then liberal participants should discount the likelihood of researchers' bias as an explanation for findings in the UAA line of research, which our liberal participants did, but conservatives should see researchers' bias as a cause for concern about the UR line of research, which our conservative participants did. Left-liberal opposition to using technology to detect unconscious–anti-American bias was grounded in concerns about the relative costs of false positives and false negatives, whereas error costs played little mediating role in conservative opposition to using the technology to detect UR. In short, conservatives worry that liberal scientists have smuggled their value judgments into a line of research that happens to advance a liberal agenda, while liberals worry that valid science may be used to advance a conservative agenda (that is, that companies or policy makers will reach a trade-off of type I and II errors different from their own).[4]

[4] We do not claim to have exhausted all possible mediators of motivated reasoning

Notwithstanding ideological differences in support for punishing organizations when an employee's UR or UAA leads to harm, we found important limits on how far both liberals and conservatives were willing to go in holding others accountable for unconscious bias. One constraint was the harm principle: virtually no one was ready to abandon that principle and endorse punishing individuals for unconscious attitudes per se—even though there was some support for indirect punishment in the form of using the technology to limit job opportunities for people with undesirable unconscious biases. Another constraint was a desire to appear principled: when directly asked, few respondents saw it as defensible to endorse the technology for one type of application but not for the other—even though there were strong signs from our experiment that differential ideological groups would do just that when not directly confronted with this potential hypocrisy. The harm principle constraint suggests widespread, albeit flexible, opposition to an excessively intrusive accountability regime that enforces laws against thought crimes and thought torts. The consistency constraint suggests widespread aversion to double standards and sensitivity to charges of hypocrisy and duplicity but only when that inconsistency is apparent.

Although most respondents were reluctant to acknowledge double standards for embracing the technology, the process of thinking about different applications encouraged a more critical second look at initial support for the technology—and those at the political extremes, who offered more initial support for the technology, had more rethinking to do when forced to consider a less palatable use of the technology. Here we have a special circumstance under which those at the extremes were more disposed than centrists to consider the possibility that they made a mistake. At first glance, this runs counter to political science and psychological research suggesting that extremists are more likely to display rigidity and intolerance of ambiguity (McClosky and Chong 1985; Tetlock 1984, 2005). The contradiction is, however, more apparent than real. As already noted, this result may be attributable simply to regression toward the mean (much more room for movement toward the mean

about science and technology. For example, Kahan, Jenkins-Smith, and Braman (2011) found that persons holding different cultural risk profiles systematically overestimated the scientific consensus in support of positions consistent with those risk profiles (for example, persons seeing climate change as a serious risk believed there was greater consensus among climate scientists than did those less concerned with climate change). Our results and those of MacCoun and Paletz (2009) suggest that liberals would be likelier than conservatives to cite scientific consensus as a basis for technology support, whereas conservatives would be likelier to dismiss the consensus as value driven instead of science driven.

from the extremists than from the moderates). And even if a purely statistical explanation is not adequate, there is a quite straightforward psychological explanation. This experiment confronted the more extreme participants with a choice between defending a double standard (explaining why one application is more acceptable) and acknowledging that they may have erred initially (reconsidering their support for the ideologically agreeable technology). Given the cognitive complexity of the task of justifying a double standard on a novel issue, it is not so surprising that those with more extreme views were more disposed to the lower effort option of simply backtracking from their initial position.

We should expect political groups to exploit the new mind-reading technology to target social ills they see as most pressing and to be myopic in doing so until confronted with the perverse effects of their advocacy. Endorsing the reliability and accepting the risks of a technology in one legal battle constrains one's ability to attack that technology in another legal battle. Our study suggests antidotes to such short-sighted advocacy: advocacy groups should include in their strategy formation moderates who are likely to have different type I and II error aversions than extremists, should encourage and reward dissent (Nemeth, Brown, and Rogers 2001), or appoint a devil's advocate whose role is to argue forcefully for the perverse effects of the contemplated strategies (Katzenstein 1996). Even the most committed advocates may reconsider their tactics when alerted to the unintended effects of those tactics.

In addition to informing debates over the use of mind-reading technology for evidential purposes (Shen and Jones 2011), our findings have relevance for the larger debate over preventive measures aimed at those who pose dangers to society and for the specific ongoing debate over the antidiscrimination laws aimed at UR (Tetlock and Mitchell 2009). All three groups in our study—liberals, conservatives, and moderates—opposed legal actions aimed directly at unconsciously biased individuals. When a danger presents itself only as a threat in the form of genetic or unconscious propensities, a rhetoric of just deserts, with an emphasis on retribution for harm, is unlikely to convince the public to support measures aimed at these individuals (Cameron, Payne, and Knobe 2010; Morse 1999). However, all participants were more accepting of societal measures aimed at preventing the harms of unconscious biases. These findings suggest that a move to a public health model of unconscious bias and its harms may be an effective strategy for advocates of legal applications of mind-reading technology.

Yet changing the public's mind-set to see discrimination, terrorism, and other potential threats of unconscious bias as public health problems will take advocates of state action to prevent future harms only so far (Morse 1999): treating unconscious bias as a disease to be managed through preventive measures against individuals will still require either a fundamental change in how we conceive of each other, from moral beings to disease vectors (as argued to justify civil commitments of persons posing a threat to themselves or others), or proof that the targets of state action can be motivated to prevent the spread of the disease and that the benefits of such measures exceed the costs (as argued to justify criminal penalties for the transmission of HIV).[5] Some who argue for applications of UR research to the law do employ the public health rhetoric of disease control (for example, Bagenstos 2007)—and our findings underscore the shrewdness of this move politically and legally—but there is no consensus on either the degree to which unconscious biases can be prevented from influencing behavior (Cameron, Payne, and Knobe 2010) or the harms actually associated with these biases (Mitchell and Tetlock 2009; Tetlock and Mitchell 2009).

Finally, our findings suggest that suspicions about tainted science may grow when scientists reporting findings challenging the conventional wisdom become involved in the political and legal debates on the policy relevance of those findings. When social scientists have become part of an explicit effort to expand antidiscrimination law and have invoked unconscious-bias research in support of that effort (Potier 2004), these public political statements inevitably raised suspicions about researcher bias, especially among conservatives. Advocates of legal applications of UR research seem to have understood the credibility-corrosive effects of this tactic and have sought to defend the scientific status of the research by dismissing doubts about the validity of this research as politically motivated backlash (Bagenstos 2007; Kang 2010; Lane, Kang, and Banaji 2007). Such counterattacks themselves may be assimilated to fit preexisting ideological viewpoints for extremists, and their effects on moderates await further study.

Most fundamental, our results raise serious questions about the role of scientists in policy debates and the dangers of crossing the traditional fact/value divide. Our participants understood that the use of even sound

[5] Note, however, with respect to a consequentialist justification for state intervention, that participants in our study were unwilling to endorse legal actions against individuals with unconscious biases even when those biases increased the likelihood of acts of terrorism. Whether denial of employment for such individuals, rather than legal action, would be seen as justified in light of the threat they may pose awaits study.

scientific technology requires value judgments. Deference to science will take scientist policy advocates only so far. Once scientists have been categorized as advocates of an issue (Pielke 2007) on a particular policy trade-off, they risk losing the deference that their linkages to the scientific community once bestowed, and the credibility of the entire scientific community may suffer.

REFERENCES

Ames, Susan L., Jerry L. Grennard, Carolien Thush, Steve Sussman, Reinout W. Wiers, and Alan W. Stacy. 2007. Comparison of Indirect Assessments of Association as Predictors of Marijuana Use among At-Risk Adolescents. *Experimental and Clinical Psychopharmacology* 15:204–18.

Ayres, Ian. 2001. *Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination*. Chicago: University of Chicago Press.

Bagenstos, Samuel R. 2007. Implicit Bias, "Science," and Antidiscrimination Law. *Harvard Law and Policy Review* 1:477–93.

Bennett, Mark W. 2010. Unraveling the Gordian Knot of Implicit Bias in Jury Selection: The Problems of Judge-Dominated Voir Dire, the Failed Promise of *Batson,* and Proposed Solutions. *Harvard Law and Policy Review* 4:149–71.

Brown, Teneille, and Emily Murphy. 2010. Through a Scanner Darkly: Functional Neuroimaging as Evidence of Criminal Defendant's Past Mental States. *Stanford Law Review* 62:1119–1208.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cameron, C. Daryl, B. Keith Payne, and Joshua Knobe. 2010. Do Theories of Implicit Race Bias Change Moral Judgments? *Social Justice Research* 23: 272–89.

Green, Donald P. 2009. Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science? Unpublished manuscript. Yale University, Institution for Social and Policy Studies, New Haven, Conn.

Greenwald, Anthony G, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 6:1464–80.

Harcourt, Bernard. 1999. The Collapse of the Harm Principle. *Journal of Criminal Law and Criminology* 90:109–94.

Kahan, Dan M., Hank Jenkins-Smith, and Donald Braman. 2011. Cultural Cognition of Scientific Consensus. *Journal of Risk Research* 14:147–74.

Kang, Jerry. 2010. Implicit Bias and Pushback from the Left. *St. Louis University Law Review* 54:1139–49.

Kang, Jerry, Mark Bennett, Devon Carbado, Pam Casey, Nilanjana Dasgupta, David Faigman, Rachel Godsil, Anthony Greenwald, Justin Levinson, and Jennifer Mnookin. 2012. Implicit Bias in the Courtroom. *UCLA Law Review* 59: 1124–86.

Katzenstein, Gary. 1996. The Debate on Structured Debate: Toward a Unified Theory. *Organizational Behavior and Human Decision Processes* 66:316–32.

Kinder, Donald R. 1998. Opinion and Action in the Realm of Politics. Pp. 778–867 in vol. 1 of *The Handbook of Social Psychology*, edited by Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. 4th ed. Boston: McGraw-Hill.

Kunda, Ziva. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108:480–98.

Lane, Kristin A., Jerry Kang, and Mahzarin R. Banaji. 2007. Implicit Social Cognition and Law. *Annual Review of Law and Social Science* 3:427–51.

MacCoun, Robert, and Susannah Paletz. 2009. Citizens' Perceptions of Ideological Bias in Research on Public Policy Controversies. *Political Psychology* 30: 43–65.

McClosky, Herbert, and Dennis Chong. 1985. Similarities and Differences between Left-Wing and Right-Wing Radicals. *British Journal of Political Science* 15: 329–63.

Mill, John Stuart. [1859] 1978. *On Liberty*. Indianapolis: Hackett.

Mitchell, Gregory, and Philip E. Tetlock. 2009. Facts Do Matter: A Reply to Bagenstos. *Hofstra Law Review* 37:737–61.

Molesworth, Brett R. C., and Betty Chang. 2009. Predicting Pilots' Risk-Taking Behavior through an Implicit Association Test. *Human Factors* 51:845–57.

Morse, Stephen J. 1999. Neither Desert nor Disease. *Legal Theory* 5:265-309.

Nemeth, Charlan, Keith Brown, and John Rogers. 2001. Devil's Advocate versus Authentic Dissent: Stimulating Quantity and Quality. *European Journal of Social Psychology* 31:707–21.

Nock, Matthew K., and Mahzarin R. Banaji. 2007a. Assessment of Self-Injurious Thoughts Using a Behavioral Test. *American Journal of Psychiatry* 164: 820–23.

———. 2007b. Prediction of Suicide Ideation and Attempts among Adolescents Using a Brief Performance-Based Test. *Journal of Consulting and Clinical Psychology* 75:707–15.

Ostafin, Brian D., G. Alan Marlatt, and Anthony G. Greenwald. 2008. Drinking without Thinking: An Implicit Measure of Alcohol Motivation Predicts Failure to Control Alcohol Use. *Behaviour Research and Therapy* 46:1210–19.

Perkins, Andrew, Mark Forehand, Anthony Greenwald, and Dominika Maison. 2008. Measuring the Nonconscious: Implicit Social Cognition in Consumer Behavior. Pp. 461–75 in *Handbook of Consumer Psychology,* edited by Curtis P. Haugtvedt, Paul M. Herr, and Frank R. Kardes. New York: Lawrence Erlbaum Associates.

Pielke, Roger A., Jr. 2007. *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge: Cambridge University Press.

Potier, Beth. 2004. Making Case for Concept of "Implicit Prejudice": Extending the Legal Definition of Discrimination. *Harvard University Gazette,* December 16. http://www.news.harvard.edu/gazette/2004/12.16/09-prejudice.html.

Preacher, Kristopher J., and Andrew F. Hayes. 2004. SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models. *Behavior Research Methods, Instruments, and Computers* 36:717–31.

Rokeach, Milton. 1973. *The Nature of Human Values.* New York: Free Press.

Sartori, Giuseppe, Sara Agosta, Cristina Zogmaister, Santo Davide Ferrara, and Umberto Castiello. 2008. How to Accurately Assess Autobiographical Events. *Psychological Science* 19:781–88.

Schwartz, Shalom H. 1992. Universals in the Content and Structure of Values. Pp. 1–65 in vol. 25 of *Advances in Experimental Social Psychology,* edited by Mark P. Zanna. New York: Academic Press.

Shen, Francis X., and Owen D. Jones. 2011. Brain Scans as Evidence: Truths, Proofs, Lies, and Lessons. *Mercer Law Review* 62:861–83.

Shrout, Patrick E., and Niall Bolger. 2002. Mediation in Experimental and Nonexperimental Studies: New Procedures and Recommendations. *Psychological Methods* 7:422–45.

Sniderman, Paul M., and Philip E. Tetlock. 1986. Symbolic Racism: Problems of Motive Attribution in Political Analysis. *Journal of Social Issues* 42:129–50.

Snowden, Robert J., Nicola S. Gray, Jennifer Smith, Mark Morris, and Malcolm J. Macculloch. 2004. Implicit Affective Associations to Violence in Psychopathic Murderers. *Journal of Forensic Psychiatry and Psychology* 15:620–41.

Steffens, Melanie Caroline, Elena Yundina, and Markus Panning. 2008. Automatic Associations with "Erotic" in Child Sexual Offenders: Identifying Those in Danger of Reoffence. *Sexual Offender Treatment* 3:1–9.

Tetlock, Philip E. 1984. Content and Structure in Political Belief Systems. Pp. 107–28 in *Foreign Policy Decision Making: Perception, Cognition, and Artificial Intelligence,* edited by Donald A. Sylvan and Steve Chan. Boulder, Colo.: Westview Press.

———. 1986. A Value Pluralism Model of Ideological Reasoning. *Journal of Personality and Social Psychology* 50:819–27.

———. 2002. Social-Functionalist Frameworks for Judgment and Choice: The Intuitive Politician, Theologian, and Prosecutor. *Psychological Review* 109: 451–72.

———. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, N.J.: Princeton University Press.

Tetlock, Philip E., and Gregory Mitchell. 2009. Implicit Bias and Accountability Systems: What Must Organizations Do to Prevent Discrimination? Pp. 3–38 in vol. 29 of *Research in Organizational Behavior,* edited by Barry M. Staw and Arthur Brief. New York: Elsevier.

Tetlock, Philip E., William T. Self, and Ramadhar Singh. 2010. The Punitiveness Paradox: When Is External Pressure Exculpatory—and When a Signal Just to Spread Blame? *Journal of Experimental Social Psychology* 46:388–95.

Tetlock, Philip E., Penny Visser, Ramadhar Singh, Mark Polifroni, Sara Beth Elson, Philip Mazzocco, and Philip Rescober. 2007. People as Intuitive Prosecutors: The Impact of Social Control Motives on Attributions of Responsibility. *Journal of Experimental Social Psychology* 43:195–209.

Thush, Carolien, and Reinout W. Wiers. 2007. Explicit and Implicit Alcohol-Related Cognitions and the Prediction of Future Drinking in Adolescents. *Addictive Behaviors* 32:1367–83.

Venkatranum, Vinod, John A. Clithero, Gavan J. Fitzsimmons, and Scott A. Huettel. 2012. New Scanner Data for Brand Marketers: How Neuroscience Can Help Better Understand Differences in Brand Preferences. *Journal of Consumer Psychology* 22:143–53.

Zhao, Xinshu, John G. Lynch, and Qimei Chen. 2010. Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research* 37:197–206.