Philip E. Tetlock
University of Pennsylvania

Gregory Mitchell
University of Virginia

L. Jason Anastasopoulos
University of California, Berkeley

Uncovering and Punishing Unconscious Bias:

An Experimental Allegory on the Politicization of Technology

This paper is part of the Decision Making for a Social World webconference

Organized by the International Cognition and Culture Institute and
the Philosophy, Politics and Economics Program at the University of Pennsylvania

The paper is followed by a **discussion**

Uncovering and Punishing Unconscious Bias:

An Experimental Allegory on the Politicization of Technology

1. INTRODUCTION

Long considered topics more appropriate to science fiction and philosophical musing than serious science and policy debate, new technology for mapping the brain's associative networks, and predicting future thoughts and behavior from those associations, has made the prospect of mind-reading and precognition topics of immediate legal and political concern. While ethicists focused on the human genome project and advances in neuroscience, social psychologists introduced an instrument based on simple reaction-time technology that poses equally weighty ethical questions. This instrument, the Implicit Association Test ("IAT"; Greenwald et al., 1998), purportedly identifies associative networks that often operate beneath conscious awareness and that affect both thought patterns and behavioral reactions to stimuli. For instance, the "forensic IAT" is used to detect differences between implicit knowledge about one's own criminal acts and explicit statements about those acts, making it a much simpler-to-use lie detection tool than fMRI-based methods (Sartori et al., 2008).

In addition to lie detection, the IAT has already been adapted for identifying pilots-in-training likely to take unsafe risks during emergencies (Molesworth & Chang, 2009), adults and youth at risk for alcohol problems or marijuana use (Ames et al., 2007; Ostafin et al., 2008; Thush & Wiers, 2007), and persons at risk of committing acts of child molestation or other acts of violence to themselves or others (Nock & Banaji, 2007a, 207b; Snowden et al., 2004; Steffens et al., 2008). The IAT's inventors market the test as means to identify and ameliorate discrimination risks within a company (Project Implicit, 2011) and to improve product advertising and probe consumer preferences (Perkins et al., 2008). Findings from IAT research are already being used in litigation to "post-dict" the unconscious motivations of managers in defendant organizations (Greenwald, 2006; Reskin, 2006), making it perhaps just a matter of time before an IAT is administered to parties themselves, as well as jurors, to reveal their unconscious motives and biases (Ayres, 2001; Bennett, 2010).

If an IAT can reliably identify pilots likely to panic during crises in flight, then future airplane passengers will surely endorse use of the test in this domain. But what if it turns out that the risky-pilot IAT has both a high true positive rate and a high false negative rate (i.e., low sensitivity), meaning that it identifies most pilots who would panic in emergencies as well as many who would not? Does the avoidance of pilot errors with dire consequences justify the wrongful termination of many careers? And if we accept this trade-off in the air flight domain, should we accept similarly structured trade-offs in the domains of employment and terrorism prevention, where IATs could be used to identify anti-minority and anti-American biases that could lead to discrimination or acts of terrorism? Should a private organization's resolution of these trade-offs be subject to second-guessing through the legal system, allowing excluded pilots to sue for using an insensitive risky-pilot IAT or allowing the estates of those killed in a plane crash to sue an airline for failing to use the risky-pilot IAT to weed out dangerous pilots?

So far public debate about the IAT centers on whether the basic research behind the IAT justifies widespread use of IATs in real world domains (e.g., Fiedler, 2006; Tetlock & Mitchell, 2009a, 2009b), rather than the large ethical issues presented by these potential applications.

There is much debate over whether IATs works as advertised and very little over whether IATs *should* be used even if they do work as advertised. We shift attention here to the ethical issues, for whether ready for application or not, extension of the IAT beyond the laboratory is likely to continue as long as the IAT is seen to have any scientific validity as a gauge of unconscious sentiments that predict future behavior.

Specifically, we present the results of an experiment designed to examine the conditions under which liberals and conservatives will support the use of unconscious bias detection as an employment screen and will support penalties against unconsciously-biased individuals or organizations that fail to use this new technology to detect biased individuals. This experiment also allowed us to examine the psychological underpinnings of this support and how support for using the IAT and punishing organizations failing to use it shift when it becomes clear that the IAT may be exploited by one's political rivals to root out unconscious biases that are not priorities within one's own value system. Before describing our experiment and its results in more detail, we situate our study in the psychological literature on the detection and punishment of norm violators and present the theoretical framework driving our inquiry.

2. THE PSYCHOLOGY OF NORM ENFORCEMENT

Evolutionary theorists have wondered how our ancestors on the savannah plains could create intricate patterns of normative order under conditions that game theorists see as unpromising: strangers interacting in large groups in which there is no centralized authority and no guarantee of future interaction (Axelrod & Hamilton, 1981; Tooby & Cosmides, 1989). One solution to this puzzle has been to posit a moralistic streak in human nature that predisposes people to value, as an end in itself, the detection and punishment of norm violators. In this view, norms are largely upheld by censorious third-party observers—intuitive prosecutors—who are not only willing to punish cheaters, but take pleasure in doing so, even if the cheaters have not cheated them personally and punishment requires a material sacrifice (Fehr & Fischbacher, 2004; de Quervain et al., 2004).

Tetlock (2002; Tetlock et al., 2007; Tetlock, Self & Singh, 2010) proposed an intuitive-prosecutor model of how people go about judging how punitive to be toward their fellow humans. The core idea is that well-socialized citizens, by definition, internalize the normative order and adopt, to varying degrees, a stance of prosecutorial vigilance toward norm violators. The key phrase is “to varying degrees.” Although citizens are prepared to defend normative systems they see as legitimate, few want to live in an oppressive world in which they themselves are subject to false accusations and intrusive scrutiny. Accordingly, Tetlock (2002) offered a *fair-but-biased-yet-correctible* (FBC) model of the intuitive prosecutor which adds psychological complexity to the stylized prosecutorial view of human nature. The model has three key components:

(1) *Fairness integral to our self-image*. People see themselves as reasonably fair-minded and do not like to think of themselves as extremists or as biased or prone to double standards. They intuitively sense the dangers of both excessive strictness and excessive leniency—and seek equilibrium accountability solutions that penalize norm violations but allow wiggle room for considering justifications or excuses (Tooby & Cosmides, 1996; Edgerton, 1985; Scott & Lyman, 1968);

(2) *Bias toward punitiveness*. People become more concerned about the dangers of excessive leniency and shift into a prosecutorial mindset when they see others showing

contempt for widely shared values—and getting away with it. This mindset has affective indicators (people become angry, even outraged), cognitive indicators (people assign more culpability—and are more dismissive of excuses for misconduct) and behavioral indicators (people endorse harsher punishment—and also endorse punishing those who fail to punish violators). Once in this mindset, people engage in motivated forms of reasoning (Kunda, 1990) that they might otherwise deem biased or extremist—and defend judgmental tendencies such as the fundamental attribution error and the severity effect that they might otherwise dismiss as judgmental flaws (Goldberg, Lerner, & Tetlock, 1999; Lerner, Goldberg, & Tetlock, 1999; Tetlock, 2000; Tetlock et al., 2007);

(3) *Potential for self-correction*. Many people recognize, however, that they are capable of slipping into emotion-laden states of mind that can distort their judgment—and are willing, on sober second thought, to revise their opinions (Lerner et al., 1998), although they may under-correct and sometimes even over-correct (Petty, Briñol, Tormala, & Wegener, 2007). In low-threat laboratory settings, people can often be induced to disengage from the prosecutorial mindset by minimalist forms of accountability that merely pose questions reminding them that their judgments are under scrutiny. When confronted in a repeated-measures design by evidence they have given more weight than they realized to a manipulated factor (such as accidental severity of consequences), many respondents are willing to revise their judgments to bring them into line with their implicit standards of fairness, rationality and consistency (Tetlock et al., 2007).

Unfortunately, like other social-psychological approaches to attribution of responsibility (e.g., Alicke, 2000; Heider, 1958), the FBC model is vague on what counts as a “norm violation”–and falls back on the circular practice of relying on operational definitions: a norm violation is whatever members of a moral community agree counts as a violation. Among other shortcomings, this approach obscures an array of provocative research questions, such as how members of moral communities go about deciding where to set their thresholds for labeling conduct a norm violation, why different subgroups set their thresholds in different locations, and why subgroups sometimes shift their thresholds in response to new arguments and evidence.

One escape from positivist circularity is to turn to ethical or legal theory for criteria for determining what should qualify as a norm violation deserving a punitive response (e.g., Malle & Nelson, 2003; Woolfolk, Doris & Darley, 2006), and one classic philosophical solution is provided by John Stuart Mill’s (1859/1978) harm principle: “the only purpose for which [state] power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others" (p. 9). However, if anything is clear from the ensuing 150-year debate over the harm principle, it is the multidimensionality of the concept of harm. In addition to the obvious desire to be free of harm to one’s physical person and property, people can feel wounded in a vast array of symbolic ways: individually or collectively, cognitively or emotionally, and morally or spiritually.

This definitional ambiguity gives intuitive prosecutors much room to engage in motivated reasoning in choosing which forms of harm to highlight or trivialize. In effect, whoever defines harm can set prosecutorial priorities. Thus definitions of harm become part of political debates, with the conservative right defining harm expansively to neutralize threats to public order, the nuclear family, property rights and national security, and the egalitarian left defining harm expansively to neutralize threats to the dignity of women and traditionally disadvantaged minorities, the poor and even non-human entities such as ecosystems (Harcourt, 1999). From Mill on, defenders of the harm principle have worried about good-cause temptations to expand the definition of harm beyond palpable harm to persons and property to cover ever more intangible forms of harm. The crux of their concern is that when intuitive prosecutors feel the urge to crack down on conduct they find particularly irksome–be it pornography or graffiti–it is all too easy for them to tweak vague definitional boundaries of harm and blurry societal thresholds for activating punitive responses.

This study explores how people react to an extension of the harm principle into what was once the realm of science fiction through advances in psychological science: the possibility of holding people accountable not just for their deeds and conscious states of mind but also for unconscious cognitions that may increase the likelihood of harmful conduct (Gazzaniga, 2007). Specifically, this study asked participants to imagine in the near future that scientists have created technologies that can reveal unconscious attitudes that people are not aware of possessing but that may influence their actions. 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 either of unconscious prejudice against African-Americans or unconscious anti-Americanism. In the former case, unconscious prejudice among managers posed a threat to the fair treatment of African-American employees in workplace, whereas in the latter case, unconscious anti-Americanism among workers in security fields posed a threat to the safe operation of the nation's airports and other vulnerable facilities.

From the standpoint of the FBC model, these shifting uses of technology should provoke shifting patterns of value conflict among observers who attach differential importance to civil liberties, equal employment opportunity, and national security. Absent a strong threat to a deep countervailing value such as equality or security, the FBC model predicts the default response to be inaction: the harm principle will constrain punitiveness. It will be hard for observers who see themselves as fair-minded to justify a punitive stance toward human beings 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. Indeed, what legitimate rationale can there be for penalizing people who are not driven by conscious choices to flout society's values—and who, having harmed no one, can hardly be presumed to merit just-desert penalties linked to pain inflicted on others? However, to the degree there is a strong threat to a countervailing value, it should become increasingly difficult for observers 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 it is so obvious that society would be better off if preventive (yet arguably punitive) measures were taken to stop unconscious attitudes from causing predictable harm?

From this latter, utilitarian perspective, intuitive prosecutors should not just be concerned with deterring particular acts by identifiable people; they should care about precedents and defending a general rule aimed at preventing aggregate harms caused by undesirable unconscious attitudes. They can invoke as analogies criminal and civil law prohibitions on negligent and reckless conduct such as dangerous driving, which focus on the enhanced likelihood of harm because the only practical way to reduce risk is an outright ban. Taking action against only those reckless drivers who cause accidents is a profoundly suboptimal policy. Similarly, waiting for unconscious attitudes to cause discrimination or security breaches is equally suboptimal.

The motivated-reasoning postulate of the intuitive-prosecutor framework (the “biased” component of the FBC model) predicts that exactly “how defensible” we deem such harm-expansion arguments will hinge on deep ideological sympathies and antipathies. Holding facts constant, people are likely to set their thresholds for sounding their harm alarms at predictably (ordinally) different points along the risk continuum, with liberals setting their thresholds lower for unconscious prejudice than for unconscious anti-Americanism and conservatives doing the opposite. It is useful to view these harm alarms as carrying prosecutorial implications that can range from varying forms of criminalization to varying enhancements of civil liability linked to legal doctrines of negligence. Given the civil libertarian norms of early 21st century American political culture, few are likely to embrace overt punitiveness (opening the Orwellian specter of “thought crime”), but many may be open to quasi-punitive measures, such as the idea that organizations should risk greater penalties when things go wrong if they failed to test their employees for unconscious biases.

Put plainly, the FBC model expects intuitive prosecutors to play ideological favorites but not to be heavy-handed in how they do it. To explore this possibility, this study focuses on observers’ willingness to embrace hypothetical technological breakthroughs in assessing unconscious attitudes that many people might prefer to conceal but that society arguably has an interest in discovering because of the technology’s potential to predict norm violations. The study tests the following specific hypotheses:

(1) Consistent with the fairness component of the FBC model, few observers will deem it justifiable to take *overtly punitive* measures against people based solely on unconscious attitudes that have yet to translate into harmful acts and that people are not even aware of possessing. But many observers will see good justifications for interpreting the harm principle broadly and supporting *covertly punitive* measures that impose special compliance burdens on those with potentially harm-producing unconscious attitudes. Covertly punitive measures involve supporting judges and regulators who create incentives for organizations to be proactive—and to take steps to avoid harmful conduct that might flow from unconscious attitudes.

(2) Individual-difference research on value hierarchies (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. Consistent with the bias or motivated-reasoning component of the FBC model, conservatives will be: (a) 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; (b) less willing to downplay fairness and libertarian concerns on behalf of lower-ranked values and correspondingly more prone to mobilize counter-arguments for resisting adoption of the technology, such as concerns about false-positive labeling of high scorers as racists, concerns about an activist scientific community, and concerns about creating an excessively intrusive and oppressive accountability regime.

(3) Research on value hierarchies (Rokeach, 1973; Schwartz, 1992; Tetlock, 1986) also indicates that liberals put higher priority on the values of equality and remedying past collective wrongs and lower values on crime control and national security. Consistent with the bias component of the FBC model, liberals will be: (a) more willing 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; (b) less willing to downplay fairness and libertarian concerns on behalf of lower-ranked values and correspondingly more prone to mobilize counter-arguments for resisting adoption of the technology, such as concerns about false-positive labeling of high scorers as terrorist threats, concerns about an activist scientific community, and concerns about creating an excessively intrusive and oppressive accountability regime.

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

(5) The FBC implies that people feel a need for socially acceptable rationales for unfamiliar and potentially controversial decisions and that, depending on the subculture, these rationales are likely to include ontological justifications (claims about the pervasiveness of destabilizing unconscious attitudes), epistemic justifications (claims about the objectivity of the scientific community) and ethical justifications (claims about the relative dangers of false-positive vs. false-negative attributions of attitudes). It follows

that the more one’s ideological outlook predisposes one to see false-positive attributions as more serious than false negatives, the more it predisposes one to see unconscious attitudes as pervasive and corrosive of core values, and the more it disposes one to be suspicious of the scientific community, the more that outlook should predict opposition to societal applications.

(6) When people are asked questions that highlight the reputational risk of harboring double standards, the correction component of the FBC model predicts the activation of a reflective mindset in which people balance the need to appear consistent (“I am not a hypocrite”) against their affinity for one technical application over the other. The expectation is that people who embraced the first-presented application (strong liberals and conservatives who respectively welcomed advances in detecting unconscious racism and unconscious anti-Americanism) will feel consistency pressure to adopt the same technology when it is now in the service of a less congenial cause. They will then have three value-conflict-reduction options: (a) accept what would otherwise be an unacceptable application; (b) defend a double standard by explaining why one application is more acceptable than the other; (c) acknowledge a possible error and 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 created by this experiment, we predict an exception to the empirical generalization that those at the political extremes will be most unwilling to reconsider their positions. We predict that, 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 now find the second

application a bitter ideological pill should find reconsideration of the first application the most attractive option.

3. METHOD

3.1 Participants

Ninety-five managers from executive MBA programs at the University of California, Berkeley (M_{age} = 34; 64 men, 31 women) 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-conservatism self-identification scale (1 = *strongly liberal in conventional sense of the term*, 5 = *moderate/not consistently one direction or the other,* and 9 = *strongly conservative in 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 = *strongly 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: (1)

Participants assigned to the *control scenario* reacted to a description of the new technology that mentioned no specific intended application; (2) participants assigned to the *unconscious-prejudice scenario* judged the same technology but learned that its primary application was for detecting unconscious bias against African-Americans by employers; (3) participants assigned to the *unconscious-terrorism scenario* judged the same technology but learned that its primary application was for detecting unconscious anti-Americanism 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," and both technologies were described as detecting various unconscious attitudes that people may hold. 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 unconscious-prejudice scenario was identical to the control scenario except that the technology was described as detecting unconscious prejudicial attitudes among European Americans and the last paragraph of this scenario added: "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 unconscious-terrorism scenario was likewise identical to the control scenario except that the technology was described as detecting unconscious anti-American attitudes among American Muslims and the last paragraph of this scenario added: "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 the assigned scenario, participants indicated their level of agreement with the following statements on a 9-point scale (unless otherwise noted, 1 = *strong disagreement,* 5 = *somewhat agree*, and 9 = *strong agreement*):

(1) Misuse potential: "All technologies can, of course, be abused. Do you agree that this technology has unusually serious potential to be abused?";

(2) Scientific value: "Do you agree that this technology has potentially great scientific value?";

(3) Perceptions of pervasiveness (wording varied by condition): "The researchers are probably right about the pervasiveness of unconscious of undesirable unconscious attitudes/unconscious prejudice against African-Americans among European Americans/unconscious anti-American attitudes among American Muslims";

(4) Harm principle: "Taking legal action against individuals based solely on claims about their unconscious attitudes (not their behavior) would be unacceptable";

(5) Researcher bias: “The scientists doing this research may have a political agenda that is biasing their work”;

(6) Appropriate use (wording differed slightly by condition as noted): “Society should use this technology to ensure that managers with undesirable unconscious attitudes/unconscious prejudice against African-Americans/unconscious anti-American attitudes are prevented from making harmful decisions”;

(7) False positive vs. false negatives (wording differed slightly by condition as noted): “Which error do you see as more serious: an employer who concludes that someone has an unconscious undesirable attitude/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 undesirable attitude/prejudice against African-Americans/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*);

(8) 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 undesirable attitudes/prejudice against African-Americans/anti-American attitudes. As a result, a manager who would otherwise have been screened out made a flawed decision that led indirectly to an accidental death/was in a position to make flawed decisions that damaged the careers of African-American employees and that led indirectly to an accidental death/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*);

(9) Reflection on opinions: (a) Participants in the control condition were given the chance to alter their level of support for the technology if it were used to screen managers from unconscious prejudice against African-Americans or to screen managers making sensitive national security decisions for unconscious anti-American attitudes (1 = *much less support*, 5 = *exactly the same support*, and 9 = *much more support*). (b) Participants in the unconscious-prejudice condition were asked if they would change their level of support for the technology if it were used to detect unconscious anti-Americanism among managers making sensitive national security decisions, and participants in the unconscious-terrorism condition were asked if they would change their level of support for the technology if it were used to detect unconscious prejudice against African-Americans among managers (1 = *much less support*, 5 *exactly the same support*, and 9 = *much more support*). (c) Participants in all conditions were asked, "looking back at your answers, do you think you were initially too eager to embrace or too quick to reject use of the technology?" (1 = *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 dependent measures, participants were debriefed, and the experimental session concluded.

4. RESULTS

As shown in Table 1, scores on the ideology scale and the five value scales indicated that self-identified conservatives were traditional in orientation (attaching lower value to social equality and higher value to national security), whereas liberals were social democratic in orientation (displaying the mirror-image priorities). We conducted a maximum likelihood factor analysis with oblimin rotation, and the first factor accounted for 74% of the variance, with the following variable loadings on that factor: ideology (.91), egalitarianism (-.84), libertarian constraint on equality (-0.74), national security (0.64), and libertarian constraint on national security (0.47). Participant scores on this ideology factor served as the measure of individual ideology in the following analyses unless indicated otherwise.

Table 2 presents the means and standard deviations for all of the dependent measures. We ran a set of three OLS and Ordered Probit regressions for each dependent variable that tested the main-effect and ideology-by-context hypotheses while controlling for gender and age (Green, 2009). We focus on the OLS results, but the similarity of these results to those of the probit regressions demonstrated the robustness of our analyses across metric assumptions about the dependent variables (Cameron & Travedi, 2005). Table 3 reports the key OLS and Ordered Probit findings.

Insert Tables 1, 2 and 3 about here

4.1. Test of Hypothesis 1

Consistent with the fairness component of the FBC model, there was near unanimity across conditions that it was unacceptable to take legal action against individuals based solely on their unconscious attitudes (M_{control} = 7.93, M_{race} = 8.15, $M_{\text{anti-Americanism}}$ = 8.22, $F(2,92) = 1.18$, $p = 0.31$). Thus, there was general support for the harm principle when the punitive action toward those with undesirable unconscious attitudes would be direct or overt. 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_{control} = 2.3, M_{race} = 2.09, $M_{\text{anti-terrorism}}$ = 2.25, F(2, 92)= 0.32, p = 0.73). Participants were more accepting of the proposition that society should use the technology to seek to prevent managers with undesirable attitudes from making harmful decisions **(**M_{control} = 4.8, M_{race} = 4.88, $M_{\text{anti-terrorism}}$ = 4.97, F(2, 92)= 0.17, p = 0.85). Thus, participant 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.

4.2. Tests of Hypotheses 2, 3 and 4

Consistent with the motivated-reasoning component of the FBC model, the correlations between ideology and support for the unconscious-mindreading technology shifted as a function of which political values the technology was purportedly protecting. Using the control group as the baseline, when the purported goal was to identify unconscious negative attitudes toward African-Americans, conservatives were more likely to see serious misuse ($\beta_{\text{ideology x race}}$ = 1.30), t(87)= 3.45 $p < 0.01$) and were skeptical of the researchers claims about the pervasiveness of these negative unconscious attitudes ($\beta_{\text{ideology x race}}$ = -1.28, t(87) = -3.43 $p < 0.01$), to view false-positive classifications of people as prejudiced as the more serious error ($\beta_{\text{ideology x race}}$ = -1.20, t(87) = -2.95 $p < 0.01$), to oppose using the technology in routine business operations ($\beta_{\text{ideology x}}$

$_{\text{race}}$ = -0.83, $t(87) = -2.34$ $p < 0.05$), and to oppose increasing the civil liability of companies that reject using the technology, even though using a technology could have prevented harm ($\beta_{\text{ideology x race}}$ = -1.11, $t(87) = -3.06$ $p < 0.01$).

By contrast, when the purported goal was to identify unconscious anti-Americanism among American Muslims, the Ideology x Treatment coefficients reversed signs in many instances. Although liberals were not more likely to see serious misuse potential ($\beta_{\text{ideology x anti-Americanism}}$= -0.55, $t(87) = -1.45$ $p > 0.05$), they were skeptical that the technology had much scientific value ($\beta_{\text{ideology x anti-Americanism}}$ = 1.31, $t(87) = 3.30$ $p < 0.01$), were skeptical of researchers' claims about the pervasiveness of these negative unconscious attitudes ($\beta_{\text{ideology x anti-Americanism}}$= 0.76, $t(87) = 2.01$, $p < 0.05$), strongly suspected that the scientists have a political agenda ($\beta_{\text{ideology x anti-Americanism}}$ = -1.52, $t(87) = -4.05$ $p < 0.001$), strongly opposed using the technology in routine business operations ($\beta_{\text{ideology x anti-Americanism}}$ = 1.50, $t(87) = 4.22$ $p < 0.001$), and opposed increasing the civil liability of companies that reject using the technology, even though using a technology could have prevented harm ($\beta_{\text{ideology x anti-Americanism}}$ = 1.08, $t(87) = 2.97$ $p < 0.01$).

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 "supportiveness" 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 mindreading technology, moderates showed no shift in support for the technology as a function of its intended use: liberals supported the technology when aimed at unconscious prejudice but conservatives did not (*M* = 5.75 vs. *M* = 4.23; t(16) = 4.45, p<0.001); conservatives supported the technology when aimed at anti-American attitudes but liberals did not (*M* = 5.9 vs. *M* = 4.14; t(19) = -4.29, p<0.001); (*M* = 5.75 vs. *M* = 4.23); moderates showed moderate support for use of the technology across conditions ($M_{\text{unconscious terrorism}}$ = 4.91 vs. $M_{\text{unconscious prejudice}}$ = 4.90 vs. M_{control} = 5.15; *F*(2,28) = 0.58, *p* = 0.57) .

4.3. Test of Hypothesis 5

The FBC model predicted that interactions between ideology and error aversion, ideology and pervasiveness of bias, and ideology and researcher bias would track the ideology by opposition-to-technology interactions. To explore possible reasons underlying ideological selectivity in support of the technology, we ran a series of OLS mediational analyses. To simplify the analysis, we used the average of the responses to the questions on use of the technology to screen managers and increased civil liability for company failure to use the technology to prevent harm as the measure of participant support for use of the technology (*r* = .44). Results of the mediation analysis demonstrate that, although attitudes toward false negatives versus false positives completely mediate the relationship between ideology and attitudes toward applications of the technology when the technology is used to detect unconscious bias against African-Americans, researcher bias plays virtually no mediating role. By contrast, when the technology is used to detect anti-Americanism among Muslims, researcher bias partially mediates the relationship between ideology and attitudes toward policy applications, whereas attitudes towards false negatives versus false positives play no mediating role. Table 4 presents these meditational analyses.

Insert Table 4 about here

These findings were confirmed by a series of Sobel tests and 95% confidence intervals from bootstrapped re-samplings of the indirect effect (a1 x b1). The Sobel test has become the de facto standard for mediation in social psychology, but it has come under attack by psychometricians who argue that, although the standard errors for each coefficient in the mediation analysis are accurate as long as regression assumptions are met, the standard errors for interaction coefficients in the Sobel test are not, especially for smaller sample sizes (Shrout & Bolger 2002; Preacher & Hayes, 2004; Zhao et al., 2010). We therefore used both sets of tests to reduce doubt that our findings hinged on method-specific assumptions.[1] Results of the Sobel tests and bootstrap of the a1 x b1 interaction with 5000 replications were consistent. Strong evidence of mediation is found in the unconscious-prejudice condition for false negative/false positive balancing (DV 6), but not researcher bias (DV 5) using the Sobel test (DV 6: Sobel $z = -2.35$, $p < 0.05$, DV 5: Sobel $z = -1.33$, $p > 0.10$) and using the bootstrapped a1 x b1 interaction (DV 6: 95% CI [-0.33,-0.03], DV 5: 95% CI [-0.27, 0.05]).[2] In the unconscious-terrorism condition, however, strong evidence of mediation is found for researcher bias but not false positive/false negative balancing using the Sobel test (DV 6: Sobel $z = 0.11$, $p > 0.10$, DV 5: Sobel $z = 2.80$, $p < 0.01$) and using the bootstrapped a1 x b1 interaction (DV 6: 95% CI [-0.04, 0.07], DV 5: 95% CI [0.07, 0.37]).

[1] Bootstrapped a1 x b1 coefficients and empirical confidence intervals were calculated using the bootstrap algorithm discussed in Shrout and Bloger (2002) and were implemented in the R statistical package.

[2] When the confidence interval for the empirical distribution of a1xb1 does not pass through zero, we can reject the null hypothesis that the true indirect effect, a_1 x b_1 , equals zero (Zhao et al., 2010).

4.4. Test of Hypothesis 6

The correction component of the FBC model predicts that initial positions would constrain later ones but that people would abandon initial positions if consistency pressures called on them to embrace an application that fell in their latitude of rejection. To test this prediction, we examined reactions to four switches: (1) from no-specified-use (control) to use for unconscious-prejudice detection, (2) from control to use for unconscious-terrorism detection; (3) from unconscious-prejudice detection to use for unconscious-terrorism detection; (4) from unconscious-terrorism detection to use for unconscious-prejudice detection.

In the control-to-prejudice switch, we find a significant liberal-conservative cross-over in which liberals offered more support for the technology on knowing its intended use (M_{liberals} = 5.7 vs. $M_{\text{conservatives}}$ =4.08; $t(19) = 4.26$, $p < .001$). But when the use switched from unconscious-terrorism detection to unconscious-prejudice detection, support among liberals and conservatives did not differ (M_{liberals} = 4.44 vs. $M_{\text{conservatives}}$ = 4.80; $t(14.38) = -1$, $p = 0.33$). Similarly, when the application switched from unconscious-prejudice detection to unconscious-terrorism detection, support among liberals and conservatives was indistinguishable ($M_{\text{ liberals}}$ = 4.83 vs. $M_{\text{conservatives}}$ = 4.83; $t(19.6) < 1$, $p = 0.67$). The disappearance of a robust between-subjects effect in a repeated-measures context 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 these alternative applications of the technology on willingness to reconsider initial support for the technology, we ran regressions exploring the relationship between ideology and interest in reconsidering initial support and the possibility of an inverted-U relationship using deviations of participants' scores from the midpoint of the ideology scale as the measure of tendency toward ideological extremes in thinking. Our analysis revealed that considering potentially dissonant applications in the unconscious-prejudice and unconscious-terrorism conditions caused liberals and conservatives, respectively, to reassess their feelings toward the technology. Using the control group as our baseline, we found evidence that, in the unconscious-prejudice condition, liberals were more likely to believe that they were too quick to embrace the technology and conservatives were more likely to say that they were too quick to reject it ($\beta_{\text{ideology x race}} = 0.76$, $t(87) = 1.94$, $p < 0.10$). In the unconscious-terrorism condition, a stronger but opposite reaction came into play: conservatives believed they were too quick to embrace the technology and liberals believed they were too quick to reject it ($\beta_{\text{ideology x anti-Americanism}} = -0.86$, $t(87) = -2.18$, $p < 0.05$).

To test the ideologue hypothesis that extremists would be less willing to change their minds, we created a "relative-extremism" dummy variable based on the distribution of ideology scores. Political "extremists" were defined as those who scored 2 or 3 (left extreme) or 7 or 8 (right extreme) on the ideological self-identification scale. We included this "extremism" dummy variable in a regression equation with the control as the baseline group, in addition to the covariates included in the regression above. In both the unconscious-prejudice and unconscious-terrorism conditions, extremists were more likely than non-extremists to perceive that they were too eager to embrace the technology ($\beta_{\text{extremists x race}} = -0.44$, $t(84) = -2.28$, $p < 0.05$; $\beta_{\text{extremists x anti-Americanism}} = -0.55$, $t(84) = -3.79$, $p < 0.001$). Table 5 presents descriptive statistics for extremists and non-extremists on this measure of willingness to second-guess initial responses.

Insert Table 5 about here

----

5. DISCUSSION

Our results underscore how easily a new technology can become politicized. When we examine how participants in the unconscious-prejudice-detection condition and unconscious-terrorism-detection condition respond to the technology in comparison to those in the control condition (where no use was specified), we find that strong relationships emerge 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 when aimed at unconscious prejudice toward African-Americans, and conservatives were consistently more open to the technology when aimed at unconscious anti-Americanism among American Muslims. And in each case liberals were more supportive of using the technology in ways that imposed quasi-punitive burdens on those showing greater unconscious prejudice toward African-Americans, whereas conservatives were more supportive of using the technology in ways that imposed quasi-punitive burdens on those showing greater unconscious anti-Americanism.

Our data were consistent with the hypothesis that ideologically-selective willingness to apply the technology in punitive ways is partially mediated by selective skepticism toward the scientific community that produced the technology and with the hypothesis that this willingness to use technology is fully mediated by prior beliefs about the relative seriousness of false positives versus false negatives in the domains of discrimination versus terrorism. These patterns suggest that intuitive prosecutors play favorites and draw on well-defined ideological scripts to justify this favoritism (Kunda, 1990). Prosecutorial priorities require 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), and 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 technological opposition, a reluctance that may be explained by MacCoun and Paletz's (2009) finding that citizens tend to believe scientists hold liberal rather than conservative political views. If scientists are expected to be liberals, then liberal participants should discount the likelihood of researcher bias as an explanation for findings in the unconscious-terrorism line of research, which our participants did, but conservatives should see researcher bias as a cause for concern about the unconscious-prejudice line of research, which our participants did. Left liberal opposition to the use of the IAT as anti-terrorism technology 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 the IAT when used as anti-discrimination technology. In short, conservatives worry that liberal scientists have smuggled their value judgments into the line of research that happens to advance a liberal agenda, while liberals worry that valid science may be used to advance a conservative agenda (i.e., that companies or policymakers will reach a trade-off of Type I and II errors different from their own).[3]

[3] We do not claim to have exhausted all possible mediators of motivated reasoning about science and technology. For example, Kahan, Jenkins-Smith and Braman (2011) found that persons holding different cultural risk profiles

Overall, the data do not paint a picture of unconstrained prosecutorial discretion. There are limits to how far people are prepared to go in holding others accountable. 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 covert 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 between-subjects design 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 where 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

systematically overestimated the degree of scientific consensus in support of positions consistent with those risk profiles (e.g., persons seeing climate change as a serious risk believed there was greater consensus among climate scientists than those less concerned with climate change). Our results and those of MacCoun and Paletz (2009) suggest that liberals would be more likely than conservatives to cite scientific consensus as a basis for technology support, while conservatives would be likely to dismiss the consensus as value-driven as opposed to science-driven.

they made a mistake. At first glance, this runs counter to political-psychological research suggesting extremists are more likely to display intolerance of ambiguity and rigidity (McCloskey & Chong, 1985; Tetlock, 1984, 2005). The contradiction is, however, more apparent than real. 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.

In closing, it is worth noting that those who argue for the most ambitious applications of unconscious bias research to the law and public policy typically eschew the rhetoric of punitiveness aimed at delinquent agents (e.g., Bagenstos, 2007; Kang & Banaji, 2006). Instead, they prefer the public-health rhetoric of disease control: the right mindset for approaching these issues is not the intuitive prosecutor but rather the intuitive epidemiologist. Given the difficulty of mobilizing even the ideologically predisposed to adopt a prosecutorial stance toward unconscious bias (especially when there are suspicions of double standards), it is easy to see the political temptations of this public-health rhetoric—and it is worth testing the efficacy of such reframing. The next phase of this unfolding debate over legal applications of research based on the IAT will be worthy of sustained attention (Mitchell & Tetlock, 2006). We should expect civil-libertarian skeptics to insist on unpacking the implications of this metaphorical shift in antidiscrimination law from blameworthy intentional actors to blameless neuro-transmitters of contagious ideas, and we should expect conservative skeptics to dismiss the underlying research as contaminated by researcher bias and see this “scientific” reframing of the debate as merely the continuation of politics by other means.

It is also worth noting that our results shed potential light on the actual, not just hypothetical, political and legal debates surrounding the policy-relevance of unconscious-bias assessment techniques. When social scientists became part of an explicit effort to expand antidiscrimination law and invoked IAT research in support of that effort (Potier, 2004), these public political statements were likely to raise suspicion about researcher bias, especially among conservatives. Later advocates of IAT research for legal purposes seem to have understood the credibility-corrosive risks of this tactic and have sought to defend the scientific status of the research by dismissing doubts about the statistical stability and external validity of IAT research as politically-motivated backlash (Bagenstos, 2007; Kang, 2010; Lane, Kang & Banaji, 2007). Our results suggest that these counter-attacks are themselves likely to be assimilated to fit pre-existing ideological viewpoints for extremists, but their effects on moderates await further study. Most fundamentally, however, our results raise serious questions about the role of scientists in public policy debates and the dangers of crossing the traditional, neo-positivist fact-value divide. Our participants understood that the use of even sound scientific technology requires value judgments. Deference to science will only take scientist-policy advocates so far, for once scientists have been categorized as “issue advocates” (Pielke, 2007) on a particular policy trade-off dimension, those scientists risk losing any deference that their linkages to the scientific community once bestowed.

REFERENCES

Alicke, Mark. D. 2000. Culpable Control and the Psychology of Blame. *Psychological Bulletin* 126: 556-574.

Ames, Susan L., Jerry L. Grennard, CarolienThush, Steve Sussman, Reinout W. Weirs, 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: 204-218.

Axelrod, Robert and William D. Hamilton. 1981. The Evolution of Cooperation. *Science* 211: 1390–1396.

Ayres, I. 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 & Policy Review* 1: 477-493.

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 & Policy Review* 4: 149-171.

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

de Quervain, Dominique J.-F.., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. The Neural Basis of Altruistic Punishment. *Science* 305:1254-1258.

Edgerton, Robert B. 1985. *Rules, Exceptions, and Social Order*. Berkeley, CA: University of California Press.

Fehr, Ernst, and Urs Fischbacher. 2004. Third-party Punishment and Social Norms. *Evolution and Human Behavior* 25: 63-87.

Fiedler, Klaus, Claude Messner, and Matthias Bluemke. 2006. Unresolved Problems With the “I”, the “A” and the “T”: Logical and Psychometric Critique of the Implicit Association Test (IAT). *European Review of Social Psychology* 17: 74-147.

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley.

Gazzaniga, Michael S. 2007. *The Ethical Brain*. New York: Dana Press.

Goldberg, Julie H., Jennifer S. Lerner, and Philip E. Tetlock. 1999. Rage and Reason: The Psychology of the Intuitive Prosecutor. *European Journal of Social Psychology* 29: 781-795.

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, July.

Greenwald, Anthony G. 2006. Expert Report, Satchell v. FedEx Express, Case Nos. C 03-2659 and C 03-2878, U.S. District Court, Northern District of California.

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: 1464-1480.

Harcourt, Bernard. 1999. The Collapse of the Harm Principle. *Journal of Criminal Law & Criminology* 90: 109-194.

Kahan, Dan M., Hank Jenkins-Smith, and Donald Braman. 2011. Cultural cognition of scientific consensus. *Journal of Risk Research* 14: 147-174.

Kang, Jerry. 2010. Implicit Bias and Pushback from the Left. *St. Louis University Law Review* 54: 1139-1149.

Kang, Jerry and Mahzarin R. Banaji. 2006. Fair Measures: A Behavioral Realist Revision of "Affirmative Action." *California Law Review* 94: 1063-1118

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

Kunda, Ziva. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108: 480–498.

Lane, Kristen A., Jerry Kang, and Mahzarin R. Banaji. 2007. Implicit Social Cognition and Law. *Annual Review of Law & Social Science* 3: 427-451.

Lerner, Jennifer S., Julie H. Goldberg, and Philip E. Tetlock. 1998. Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin* 24: 563-574.

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: 329-363.

Malle, Bertram F., & Sarah E. Nelson. 2003. Judging *MensRea*: The Tension Between Folk Concepts and Legal Concepts of Intentionality. *Behavioral Sciences and the Law* 21: 563-580.

Mill, John Stuart. 1859/1978. *On Liberty*. Indianapolis, IN: Hackett Publishing Co.

Mitchell, Gregory, and Philip E. Tetlock. 2006. Anti-discrimination Law and the Perils of Mind Reading. *The Ohio State University Law Review* 67: 1023-1121.

Molesworth, Brett R. C., and Betty Chang. 2009. Predicting Pilots' Risk-taking Behavior Through an Implicit Association Test. *Human Factors* 51: 845–857.

Nock, Matthew K., and Mahzarin R. Banaji. 2007a. Assessment of Self-injurious Thoughts Using a Behavioral Test. *American Journal of Psychiatry* 164: 820–823.

Nock, Matthew K., and Mahzarin R. Banaji. 2007b. Prediction of Suicide Ideation and Attempts Among Adolescents Using a Brief Performance-based Test. *Journal of Consulting and Clinical Psychology* 75: 707–715.

Ostafin, Brian D., G. Alan Marlatt, and Anthony G. Greenwald. 2009. Drinking Without Thinking: An Implicit Measure of Alcohol Motivation Predicts Failure to Control Alcohol Use. *Behaviour Research and Therapy* 46: 1210-1219.

Perkins, Andrew, Mark Forehand, Anthony Greenwald, and Dominika Maison. 2008. Measuring the Nonnconscious: Implicit Social Cognition in Consumer Behavior. Pp. 461-475 in *Handbook of Consumer Psychology*, edited by Curtis P. Haugtvedt, Paul M. Herr and Frank R. Kardes. New York: Lawrence Erlbaum.

Petty, Richard E., Pablo Briñol, Zakary L. Tormala, Duane T. Wegener. 2007. The Role of Meta-cognition in Social Judgment. Pp. 254-284 in S*ocial psychology: Handbook of Basic Principles*, 2d ed., edited by Arie W. Kruglanski and E. Tory Higgins. New York: Guilford Press.

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.

Project Implicit. 2011. “Services” page at http://www.projectimplicit.net/services.php.

Reskin, Barbara F. 2006. Declaration, Ellis v. Costco Wholesale Corp., Case No. C-04-3341, U.S. District Court, Northern District of California.

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

Giuseppe Sartori, Sara Agosta, Cristina Zogmaister, Santo Davide Ferrara, and Umberto Castiello. 2008. How to Accurately Assess Autobiographical Events. *Psychological Science* 19: 781–788.

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.

Scott, Marvin B., and Stanford M. Lyman. 1968. Accounts. *American Sociological Review* 33: 46-62.

Shrout, Patrick E., and Niall Bolger. 2002. Mediation in Experimental and Non-experimental Studies: New Procedures and Recommendations. *Psychological Methods* 7: 422-445.

Sniderman, Paul M., and Philip E. Tetlock. 1986. Symbolic Racism: Problems of Motive Attribution in Political Analysis. *Journal of Social Issues* 42: 129-150.

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 & Psychology* 15: 620-641.

Steffens, Melanie C., 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. 107-128 in *Foreign Policy Decision-making: Perception, Cognition, and Artificial Intelligence,* edited by S. Chan & D. Sylvan. Boulder, CO: Westview Press.

Tetlock, Philip E. 1986. A Value Pluralism Model of Ideological Reasoning. *Journal of Personality and Social Psychology* 50: 819-827.

Tetlock, Philip E. 2000. Cognitive Biases and Organizational Correctives: Do Both Disease and Cure Depend on the Ideological Beholder? *Administrative Science Quarterly* 45: 293-326.

Tetlock, Philip E. 2002. Social-functionalist Frameworks for Judgment and Choice: The Intuitive Politician, Theologian, and Prosecutor. *Psychological Review* 109: 451-472.

Tetlock, Philip E. 2005. Gauging the Heuristic Value of Heuristics. *Behavioral and Brain Sciences* 28: 562-563.

Tetlock, Philip E., and Gregory Mitchell. 2009a. 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., and Gregory Mitchell. 2009b. Adversarial Collaboration Aborted, But Our Offer Still Stands. 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: 388-395.

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: 1367–1383.

Tooby, John, and Leda Cosmides. 1989. Evolutionary Psychology and the Generation of Culture, Part I. Theoretical Considerations. *Ethology & Sociobiology* 10: 29-49.

Tooby, John, and Leda Cosmides. 1996. Friendship and the Banker's Paradox: Other Pathways to the Evolution of Adaptations for Altruism. *Proceedings of the British Academy* 88: 119-143.

Woolfolk, Robert L., John M. Doris, and John M. Darley. 2006. Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition* 100: 282-301.

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.

Table 1

Correlation Matrix for Ideology Questions

	1.	2.	3.	4.	5.
1. Ideology Self-Report	1.00				
2. Egalitarianism	-0.83	1.00			
3. Libertarianism-Egalitarianism Balancing	-0.67	0.58	1.00		
4. National Security	0.65	-0.61	-0.44	1.00	
5. Libertarianism-National Security Balancing	0.58	-0.58	-0.19	0.6	1.00

Table 2

Summary Statistics by Experimental Condition

Dependent Variable	All (N=95)		Control (n=30)		Racism (n=33)		anti – Americanism (n=32)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Misuse potential	5.12	1.10	4.97	1.19	5	1.03	5.38	1.07
2. Scientific potential	5.11	1.12	5.3	0.99	5.09	0.98	4.94	1.37
3. Believe pervasiveness claims	5.07	1.43	4.9	1.52	5	1.52	5.31	1.26
4. No liability for unconscious attitudes	8.11	0.76	7.93	0.91	8.15	0.67	8.21	0.71
5. Researchers are biased	4.84	1.21	4.83	0.99	4.76	1.32	4.94	1.32
6. False positive vs. false negative	4.87	1.14	4.97	0.76	4.67	1.31	5	1.24
7. Use Technology to Screen	4.88	1.15	4.8	0.81	4.88	1.39	4.97	1.18
8. Enhanced civil liability for non-use	2.21	1.08	2.3	0.99	2.09	1.07	2.25	1.19
9a. More support	4.84	0.88	4.8	1	*	*	4.81	0.69
9b. More support	5.00	0.74	5.07	0.64	4.94	0.83	*	*
9c. Too quick to embrace/reject	4.69	0.88	4.93	0.74	4.69	0.95	4.47	0.95

* Not Applicable.

Table 3

Ideology By Experimental Condition Contrasts

	OLS Beta-Weights		OLS Coefficients		Probit Coefficients	
	Ideology x Racism	Ideology x Anti-Americanism	Ideology x Racism	Ideology x Anti-Americanism	Ideology x Racism	Ideology x Anti-Americanism
	$\beta_{\text{ideology x race}}$	$\beta_{\text{ideology x anti-Americanism}}$	$b_{1\text{ ideology x race}}$	$b_{2\text{ ideology x anti-Americanism}}$	$b_{1\text{ ideology x race}}$	$b_{1\text{ ideology x anti-Americanism}}$
DV1	1.30**	-0.55	0.54**	-0.23	0.61**	-0.29
DV2	-0.35	1.31**	-0.15	0.57**	-0.15	0.60**
DV3	-1.28**	0.75*	-0.70**	0.42*	-0.60**	0.37*
DV4	0.01	-0.36	0	-0.11	0.01	-0.17
DV5	0.60	-1.52***	0.28	-0.72***	0.26	-0.73***
DV6	-1.19**	-0.43	-0.52**	-0.19	-0.50**	-0.19
DV7	-0.83*	1.49***	-0.36*	0.66***	-0.38*	0.78***
DV8	-1.11**	1.08**	-0.46**	0.45**	-0.67**	0.51**

Note—One model was run on each dependent variable with the control as the reference group: DV# = a + b_1 Ideology x Race + b_2 Ideology x anti-Americanism + b_3 Race + b_4 anti-Americanism + b_5 Gender + b_6 Age + b_7 Ideology; * $p < .05$, ** $p < .01$, *** $p < .001$; $N = 95$ and df = 87 for each model.

Table 4

Mediation Analyses Using Error Trade-off and Researcher Bias Responses

	Racism Condition		Anti-Americanism Condition	
	Without Mediator	With Mediator	Without Mediator	With Mediator
	Type I/II Error Balancing as Mediator			
DF	57	56	56	55
R-sq	0.37	0.48	0.52	0.52
	$\beta_{unmediated}$	$\beta_{mediated}$	$\beta_{unmediated}$	$\beta_{mediated}$
Ideology x Treatment	-1.20**	-0.67	1.78***	1.78***
Ideology	-0.11	-0.15	-0.08	-0.08
DV6: False Positive/Negative Balancing	n/a	0.38**	n/a	-0.01
	Research Bias as Mediator			
DF	57	56	56	55
R-sq	0.37	0.55	0.52	0.61
	$\beta_{unmediated}$	$\beta_{mediated}$	$\beta_{unmediated}$	$\beta_{mediated}$
Ideology x Treatment	-1.20**	-0.92**	1.78***	1.13**
Ideology	-0.11	-0.01	-0.08	-0.01
DV5: Scientists Conducting Research are Biased	n/a	-0.47***	n/a	-0.38**

Note—Models were run on two subsets of data: (1) Racism and Control Conditions; (2) Anti-Americanism and Control Conditions; *p < .05, **p < .01, ***p < .001.

Table 5

Too Eager to Embrace or Reject Technology by Extremity of Ideological Commitments

	Extremists			Non-Extremists		
	N	Mean	SD	N	Mean	SD
Control	12	5.08	1	18	4.83	0.51
Racism	13	4.46	1.2	20	4.85	0.59
anti-Americanism	12	3.75	1.14	20	4.9	0.45

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