Second Thoughts

Gregory Mitchell*

ABSTRACT

Biases in judgment and decision-making often arise at the level of first-order thoughts. If these initial thoughts are not overridden by second-order thoughts, they may lead to biased outputs. Current psychological models of legal actors assume that individuals are largely incapable of overcoming these first-order biasing thoughts and that these thoughts ultimately lead to irrational and discriminatory behavior. These models, however, ignore considerable evidence that individuals often naturally engage in self-correction and that situational pressures often encourage self-correction. I discuss the conditions under which self-correction may occur and the possibilities and limits for the law in promoting self-correction to overcome biased judgments, decisions, and behavior.

Seals . . . exemplify one of the strongest impulses of legal ordering, to signal formally to people at critical moments that they are about to leave the world of social interaction for the world of compulsion, bureaucracy and impersonality. It is a way of saying "this counts," "this is for keeps," "we're not kidding around anymore" and similar things. . . . With seals, of course, the modern problem has been that as the form of the seal has been attenuated (down to a preprinted "seal"), its cautionary effect has been vitiated apace.

—Arthur Leff[1]

Our instincts and intuitions sometimes lead us in directions that, on second thought, we'd rather not go. The impulse to hit the brakes when one's car encounters a patch of ice can be consciously overridden by preparing for that possibility. The attractive actor's endorsement of a product may resonate at some affective level, but this appeal often falls flat when its informational content is analyzed. Our initial impressions of students and co-workers may be based on stereotypes of the groups to which they are perceived to belong, but these initial impressions often give way to more nuanced beliefs as we learn more about these

* Daniel Caplin Professor of Law & E. James Kelly, Jr.-Class of 1965 Research Professor, University of Virginia School of Law, 580 Massie Road, Charlottesville, VA 22903-1738, greg_mitchell@virginia.edu. This paper benefitted from discussions with Hal Arkes, Richard Petty, and Philip Tetlock, as well as comments by participants at the McGeorge School of Law talk based on an earlier version of this paper and by students in the University of Illinois Law and Economics seminar.

1. Arthur Allen Leff, *A Letter from Professor Leff to a Prospective Publisher*, 94 YALE L.J. 1852, 1852-53 (1985) (entry on "seals" from Professor Leff's legal dictionary proposal).

people as individuals. These second thoughts, which may be the product of conscious effort or may come to us just as rapidly and unbidden as the initial thoughts, can serve as important checks on judgments, decisions, and behavior. We learn through experiences that initial reactions to stimuli should sometimes be distrusted or avoided, and we develop a variety of techniques for employing second thoughts to overcome suspect, undesirable, or maladaptive first thoughts.[2]

The propensity to engage in self-doubt and self-correction varies across persons and situations. Some people, due to their education, upbringing, values, or genetic endowment, naturally engage in reflection and revision more often than others. Nevertheless, all persons with cognitively normal functioning possess the ability to engage in some amount of deliberation, "metacognition," or thought about one's own thoughts—indeed, metacognition is sometimes described as a uniquely human characteristic.[3] Metacognition often follows a feeling of difficulty, surprise, or unease when trying to process information. For example, a fleeting feeling of knowing something that cannot quite be expressed may prompt a memory search, while a feeling of unease upon encountering a member of another social group may lead to vigilance about what one thinks and says during the encounter.

But metacognitive processes that lead to adjustments in judgments and beliefs may also occur without our awareness through the operation of automatic processes and associative networks within the mind. Both forms of second

---

2. By "first thoughts" I mean initial responses to stimuli, while "second thoughts" refer to any subsequent processing of the stimuli or thoughts initiated by the first thoughts. *Cf.* Deanna Kuhn, *Metacognitive Development*, 9 CURRENT DIRECTIONS PSYCHOL. SCI. 178, 178 (2000) (defining metacognition as "cognition that reflects on, monitors, or regulates first-order cognition"); Richard E. Petty & Pablo Briñol, *Persuasion: From Single to Multiple to Metacognitive Processes*, 3 PERSP. ON PSYCHOL. SCI. 137, 142 (2008) ("Primary thoughts are those that occur at a direct level of cognition and involve our initial associations of some object with some attribute or feeling. Following a primary thought, people can also generate other thoughts that occur at a second level, involving reflections on the first-level thoughts."). A stimulus may at times trigger multiple, conflicting first thoughts, with second thoughts then mediating which of these first thoughts will gain dominance or attention (e.g., meeting another person may activate categories related to this person's sex, race, age, and occupation, with motivation and inhibition mechanisms determining which category, and its related content, captures attention and influences subsequent thought). *See* C. Neil Macrae & Galen V. Bodenhausen, *Social Cognition: Thinking Categorically About Others*, 51 ANN. REV. PSYCHOL. 93, 102 (2000).

3. *See, e.g.*, Guy Lories et al., *From Social Cognition to Metacognition*, *in* METACOGNITION: COGNITIVE AND SOCIAL DIMENSIONS 1, 1 (Vincent Y. Yzerbyt et al. eds., 1998) ("The possibility of metacognition seems typical of the human species and may be related to our being linguistic animals. It stands as one of the important differences between animal and human cognition and the very existence of psychology is proof of our interest in our own mental processes."). Whether metacognition is in fact uniquely human is now the subject of debate. *Compare* J. David Smith & David A. Washburn, *Uncertainty Monitoring and Metacognition by Animals*, 14 CURRENT DIRECTIONS PSYCHOL. SCI. 19 (2005) (summarizing research confirming animals' capacity for metacognition), *with* Peter Carruthers, *Meta-Cognition in Animals: A Skeptical Look*, 23 MIND & LANGUAGE 58 (2008) (arguing that metacognition processes within non-humans should not be accepted from the current public data).

The capacity for metacognition appears to vary developmentally as well. Kuhn and Pease argue, for instance, that metacognitive executive control of first-order thoughts develops considerably during adolescence. *See* Deanna Kuhn & Maria Pease, *Do Children and Adults Learn Differently?*, 7 J. COGNITION & DEV. 279 (2006).

thoughts—effortful deliberation and automatic correction through metacognitive processes working at the conscious and unconscious levels[4]—have important implications for the likelihood that our judgments, decisions, and ultimately behaviors, will exhibit biases, either of the sort associated with deviations from norms of rationality or of the sort associated with judgments about other people based on their appearances and perceived group memberships.

In this Article, I discuss research demonstrating the importance of second thoughts in the correction of bias in the domains of rationality and interpersonal relations and consider the implications of this research for legal theory and lawmaking that aims to protect against the effects of irrationality and interpersonal biases.[5] Little attention has been given to the role of second thoughts in either domain, but especially the domain of interpersonal bias, where second thoughts can overcome initial categorization and evaluation processes that might otherwise result in stereotyping, prejudice, and discrimination.[6] I argue that we may conceive of the law as a cognitive force operating in the form of both conscious and unconscious second thoughts and that viewing the law in this way explains why legal regulations typically thought of as requiring conscious, intentional thought can be effective even with respect to judgments, decisions, and behaviors that have their origins at the unconscious level. In this view, the law creates metacognitive thoughts that serve as important checks on judgments, decisions, and behaviors. Lastly, I discuss how neglect of the role of second thoughts can lead to perverse legal policy.

---

4. Wilson and his colleagues refer to "deliberative and theory-driven correction" as "debiasing" and to "rapid and nonconscious correction" as "implicit adjustment." Timothy D. Wilson et al., *Mental Contamination and the Debiasing Problem*, *in* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT 185, 188-89 (Thomas Gilovich et al. eds., 2002). I refer to both types of processes as self-correction, though I distinguish between corrective process occurring above and below the threshold of conscious awareness.

5. The two domains—of rationality and interpersonal relations—are not exclusive in terms of the phenomena studied, for the same judgment and decision-making processes may be at work in each domain. The distinction arises from differences in the events or outcomes sought to be explained. The focus of rationality research is on understanding when judgment and decision-making processes lead to outcomes that are consistent or inconsistent with normative standards for achieving internal coherence and logical consistency in one's beliefs and preferences and why deviations may occur. *See, e.g.*, B.A. Mellers, A. Schwartz & A.D.J. Cooke, *Judgment and Decision-Making*, 49 ANN. REV. PSYCHOL. 447, 449-50 (1998). The focus of much interpersonal relations research is on understanding how people perceive others, form impressions or judgments about other people, and make decisions relevant to others and how these psychological processes may lead to systematic biases in judgments and decisions about members of different social groups. *See, e.g.*, Linda Tropp, *Intergroup Relations*, *in* 1 ENCYCLOPEDIA OF SOCIAL PSYCHOLOGY 494, 494-95 (Roy F. Baumeister & Kathleen D. Vohs eds., 2007).

6. Within psychology "stereotyping" refers to the association of group traits or characteristics with a particular individual perceived to belong to a particular group, "prejudice" refers to group-based, as opposed to trait-based, evaluative associations with an individual, and "discrimination" refers to behavior directed at an individual because of his or her group memberships. *See* Gregory Mitchell & Philip E. Tetlock, *Antidiscrimination Law and the Perils of Mindreading*, 67 OHIO ST. L.J. 1023, 1034-40 (2006). For instance, on encountering a law student, one may assume the person is argumentative and ambitious, may expect to dislike this person, and may avoid this person in the future because of his or her status as a law student.

I. SECOND THOUGHTS ABOUT EMERGING PSYCHOLOGICAL PORTRAITS OF LEGAL ACTORS

Over the last few decades, legal scholars have become increasingly aware of psychological research on judgment and decision-making that casts into doubt traditional legal assumptions about human nature. These scholars have sought to use this research to develop more realistic models of human psychology to serve as the foundations for legal theorizing and lawmaking.[7] The primary targets of these reform efforts have been models of humans, first, as intentional actors who can control their thoughts and behaviors if adequately motivated to do so, and second, as rational actors who respond to incentives and information efficiently and effectively.

A. The Unintentional Actor

In 1987, leading the attack on the intentional actor model, Professor Charles Lawrence published his influential article on unconscious racism that used psychoanalytic and cognitive theory to argue that unconscious motives and processes cause discriminatory acts toward minorities.[8] In 1995, Professor Linda Hamilton Krieger published her equally influential article on the role of automatic categorization processes in human judgment and decision-making, in which she argued that most discrimination occurs through subtle and often unconscious forms of bias driven by normal cognitive processes.[9] Many other important works during this time period endorsed the view expressed in

---

7. *E.g.*, Gary Blasi & John T. Jost, *System Justification Theory and Research: Implications for Law, Legal Advocacy, and Social Justice*, 94 CAL. L. REV. 1119, 1120 (2006) ("A behavioral realist approach to law posits that law can better realize its normative aims if it is based on an accurate view of how individuals behave and how social institutions function."); Christine Jolls et al., *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471, 1487 (1998) ("The project of behavioral law and economics, as we see it, is to take the core insights and successes of economics and build upon them by making more realistic assumptions about human behavior.").

8. *See* Charles R. Lawrence, III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 322-23 (1987).

> [A] large part of the behavior that produces racial discrimination is influenced by unconscious racial motivation. There are two explanations for the unconscious nature of our racially discriminatory beliefs and ideas. First, Freudian theory states that the human mind defends itself against the discomfort of guilt by denying or refusing to recognize those ideas, wishes, and beliefs that conflict with what the individual has learned is good or right. . . . Second, the theory of cognitive psychology states that the culture—including, for example, the media and an individual's parents, peers, and authority figures—transmits certain beliefs and preferences. Because these beliefs are so much a part of the culture, they are not experienced as explicit lessons.

Id.

9. *See* Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1165 (1995) ("[A] broad class of biased employment decisions now analyzed under Title VII's disparate treatment theory results not from discriminatory motivation, but from a variety of categorization-related judgment errors characterizing normal human cognitive functioning.").

Lawrence and Krieger's articles that discrimination often occurs without intentional thought or action and that a strict reading of the intentionality element within antidiscrimination statutes causes the law to miss the most pervasive forms of discrimination.[10]

The advent of new psychological tools, most notably the Implicit Association Test (IAT),[11] which can supposedly detect prejudice and stereotypes operating at the unconscious, or implicit, level, provided new support for this anti-intentionality perspective.[12] This research supposedly reveals that the great majority of Americans implicitly associate many historically disadvantaged groups with negative attributes and historically advantaged groups with positive attributes.[13] Legal scholars quickly embraced this new evidence to bolster their continued assault on the intentional actor model within antidiscrimination law,[14]

---

10. *See id*. at 1242.

> Most fundamentally, under the approach I propose, courts would reformulate disparate treatment doctrine to reflect the reality that disparate treatment discrimination can result from things other than discriminatory intent. To establish liability for disparate treatment discrimination, a Title VII plaintiff would simply be required to prove that his group status *played a role* in causing the employer's action or decision. Causation would no longer be equated with intentionality.

Id.; *see also* Lawrence, III, *supra* note 8, at 387 ("The intent requirement is a centerpiece in an ideology of equal opportunity that legitimizes the continued existence of racially and economically discriminatory conditions and rationalizes the superordinate status of privileged whites."); Martha Chamallas, *Listening to Dr. Fiske: The Easy Case of* Price Waterhouse v. Hopkins, 15 VT. L. REV. 89 (1990); Barbara J. Flagg, *"Was Blind, but Now I See": White Race Consciousness and the Requirement of Discriminatory Intent*, 91 MICH. L. REV. 953, 973 (1993); Mary F. Radford, *Sex Stereotyping and the Promotion of Women to Positions of Power*, 41 HASTINGS L.J. 471 (1990).

11. *See* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464, 1473-74 (1998) (describing the original race IAT).

The IAT measures millisecond differences in reaction times to pairings of concepts that vary in their putative stereotypic or prejudicial connotations. For instance, if a test-taker responds more quickly to the pairing of photographs of African-American faces with negative character trait words than to the pairing of photographs of European-American faces with the same negative character trait words, then the test-taker is said to exhibit an implicit negative stereotype toward African-Americans. Or if a test-taker responds more quickly to the pairing of "White-sounding" names with the term "pleasant" than to the pairing of "Black-sounding" names with the term "pleasant," then the subject is said to exhibit an implicit negative attitude (or prejudice) toward African-Americans. Numerous IATs have been developed to test for possible implicit biases against a wide range of groups. Demonstration versions of many of these IATs can be taken online at https://implicit.harvard.edu/implicit/.

12. For detailed discussions of the many problems with using data from the IAT as a basis for legal theory and policy, see Mitchell & Tetlock, *supra* note 6; Philip E. Tetlock & Gregory Mitchell, *Calibrating Prejudice in Milliseconds*, 71 SOC. PSYCHOL. Q. 12 (2008); Amy L. Wax, *The Discriminating Mind: Define It, Prove It*, 40 CONN. L. REV. 979 (2008).

13. *See* Brian A. Nosek et al., *Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site*, 6 GROUP DYNAMICS: THEORY, RES. & PRAC. 101, 112 (2002) ("From young to old, male to female, Black to White, and conservative to liberal, implicit biases are not held by a select few but are readily observed among all social groups."); Michael A. Olson & Russell H. Fazio, *Relations Between Implicit Measures of Prejudice: What Are We Measuring?*, 14 PSYCHOL. SCI. 636, 636 (2003) ("[P]rejudiced IAT scores are found in 70 to 90% of Whites." (citation omitted)).

14. *See, e.g.*, Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1490 (2005); Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate*

and to call for a host of structural reforms to supplement, if not replace, intentionality-focused legal incentives as the primary remedy for discrimination.[15]

B. The Quasi-Rational Actor

While antidiscrimination scholars were questioning the intentional actor model, a diverse group of scholars was questioning law and economics' model of humans as rational processors of information and maximizers of utility.[16] These scholars, working within the "behavioral law and economics" movement, argued that judgments made under factual uncertainty (e.g., probability judgments or causal inferences) are determined primarily by non-deliberative thought processes based on cognitive heuristics, or mental rules of thumb, rather than by careful application of the laws of probability, the rules of logic, or scientific rules for causal inference; in other words, intuitive thought ordinarily predominates, but analytical thought may sometimes override intuitions.[17] For instance,

---

Treatment, 94 CAL. L. REV. 997 (2006).

15. *E.g.*, Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action,"* 94 CAL. L. REV. 1063, 1080 (2006) ("[W]e need a new model of discrimination for implicit bias—one based on a more accurate model of human cognition and emotion, especially its constraints. This new model must promote proactive structural interventions that minimize harm without relying solely on potential individual litigation.").

16. For overviews of these attacks on the rationality assumption, see Gregory Mitchell, *Taking Behavioralism Too Seriously? The Unwarranted Pessimism of the New Behavioral Analysis of Law*, 43 WM. & MARY L. REV. 1907 (2002) [hereinafter Mitchell, *Taking Behavioralism Too Seriously?*]; Gregory Mitchell, *Why Law and Economics' Perfect Rationality Should Not Be Traded for Behavioral Law and Economics' Equal Incompetence*, 91 GEO. L.J. 67 (2002) [hereinafter Mitchell, *Equal Incompetence*].

17. The intuitive/analytical distinction roughly parallels the distinction I am drawing between first-order and second-order thoughts. While it is overly simplistic to think of the mind as consisting of two separate systems of thought, many current theories of cognition contrast automatic, intuitive processes with effortful, analytical processes. *See* Jonathan St. B.T. Evans, *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition*, 59 ANN. REV. PSYCHOL. 255, 270 (2008). Kahneman and Frederick, building on the work of Stanovich and West, provide a useful description of the two modes of thought captured by the two-systems view:

> The ancient idea that cognitive processes can be partitioned into two main families—traditionally called *intuition* and *reason*—is now widely embraced under the general label of *dual-process theories*. Dual-process models come in many flavors, but all distinguish cognitive operations that are quick and associative from others that are slow and rule-governed. We adopt the generic labels *System 1* and *System 2* from Stanovich and West. These terms may suggest the image of autonomous homunculi, but such a meaning is not intended. We use *systems* as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate.
>
> . . . .
>
> In the particular dual-process model we assume, System 1 quickly proposes intuitive answers to judgment problems as they arise, and System 2 monitors the quality of these proposals, which it may endorse, correct, or override. The judgments that are eventually expressed are called *intuitive* if they retain the hypothesized initial proposal without much modification. The roles of the two systems in determining stated judgments depend on features of the task and of the individual, including the time available for deliberation, the respondent's mood, intelligence, and exposure to

judgments about the likelihood of some future event, such as a terrorist attack, may be based on how easily similar past events come to mind rather than a consideration of the unique probabilities associated with this potential event.

Similarly, in this view, choices occur not through consultation of a stable, well-defined, coherently-ordered menu of preferences and careful deliberation about the costs and benefits of different courses of action weighted by the probability of those actions, but through the often unconscious construction and ranking of preferences based on the information available in memory and salient to the decision-maker, along with fleeting affective influences. As a result, preference formation and expression are susceptible to transitory influences that can cause choices to appear contradictory across time and situations and in violation of axioms of rational choice.[18] For instance, framing choices as gains or losses relative to a baseline (e.g., emphasizing the number of lives saved versus the number lost by a program) may lead to significant differences in choice tendencies.[19]

Just as implicit biases toward disadvantaged groups can be seen as inevitable byproducts of ordinary cognitive processes that help us organize our world into manageable categories and associate characteristics and evaluations with those categories,[20] these cognitive heuristics and preference construction processes provide an efficient means to manage the numerous and complex demands that the environment places on our limited mental resources. The cost of this efficient

---

statistical thinking.

Daniel Kahneman & Shane Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, *in* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT, *supra* note 4, at 49, 51 (citations omitted); *see also* Steven A. Sloman, *Rational Versus Arational Models of Thought*, *in* THE NATURE OF COGNITION 573 (Robert J. Sternberg ed., 1999) (reviewing evidence in support of two systems of thought); Keith D. Stanovich & Richard F. West, *Individual Differences in Reasoning: Implications for the Rationality Debate?*, *in* HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT, *supra* note 4, at 421, 436-38 (discussing "System 1" and "System 2" reasoning processes).

I discuss cognitive overrides—"second thoughts"—that may occur within each system, which sets this argument apart from the now generally accepted view that deliberative, or "System 2," processes can, under certain circumstances, debias intuitive, or "System 1," processes. In other words, I argue that, while many second thoughts may appear to be of the System 2 variety, self-correction should not be assumed to require conscious effort or deliberation.

18. John W. Payne et al., *Behavioral Decision Research: A Constructive Processing Perspective*, 43 ANN. REV. PSYCHOL. 87, 89 (1992) ("[The] notion of constructive preferences means . . . that preferences are not necessarily generated by some consistent and invariant algorithm such as expected value calculation." (citation omitted)).

19. Paul Slovic, *The Construction of Preference*, 50 AM. PSYCHOLOGIST 364, 365 (1995) ("Preferences appear to be remarkably labile, sensitive to the way a choice problem is described or 'framed' and to the mode of response used to express the preference." (citations omitted)).

20. *See, e.g.*, Max H. Bazerman & Mahzarin R. Banaji, *The Social Psychology of Ordinary Ethical Failures*, 17 SOC. JUST. RES. 111, 111 (2004) ("These ordinary unethical behaviors are conceived to be ordinary because they are assumed to be rooted in the basic mechanics of the mind's abilities and constraints. They are also ordinary in that such unethical behaviors are not characteristic of a special group of unethical people . . . but rather of all of us." (citation omitted)); Krieger, *supra* note 9, at 1188 ("[S]tereotypes, like other categorical structures, are cognitive mechanisms that *all* people, not just 'prejudiced' ones, use to simplify the task of perceiving, processing, and retaining information about people in memory.").

processing, however, is predictable biases in our judgment and decisions.[21] Thus, in the example above, when a person relies on the "availability heuristic" to judge the probability of a future terrorist event, he or she will often successfully avoid harm or disappointment; but relying on what is available in memory rather than a cold calculation based on reliable data may sometimes lead to erroneous judgments and even fatal choices.[22]

The predominant response of behavioral law and economics scholars to evidence of systematic irrational tendencies has been to call for greater governmental regulation of consumer behavior and less reliance on market competition to produce efficient outcomes. But the prescriptions of behavioral law and economics extend far beyond the buyer-seller context. No legal actor has been immune from calls for greater oversight and paternalistic protection from the government: voters, judges, jurors, financial brokers, white- and blue-collar workers, the young and the old, and the educated and the uneducated, have all been the subject of regulatory proposals due to their supposed irrational tendencies.[23]

C. The Imperfect but Self-Correcting Actor

Both of these lines of scholarship—moving toward models of humans as unintentional, quasi-rational actors—de-emphasize the role of controlled, deliberative thought processes and emphasize the role of automatic, intuitive thought processes.[24] These new models leave little room for self-correction of biased thought; indeed, the possibility of self-correction is frequently discounted on grounds that individuals lack the self-awareness, insight, and computational abilities needed to overcome automatic and intuitive responses to stimuli that lead to systematic biases.[25] Thus, incentives to be unbiased and education about

---

21. *See, e.g.*, Jeffrey J. Rachlinski, *Heuristics and Biases in the Courts: Ignorance or Adaptation?*, 79 OR. L. REV. 61, 61 (2000).

> The human brain is extremely efficient, but it is not a computer. The brain has a limited ability to process information but must manage a complex array of stimuli. In response to its natural constraints the brain uses shortcuts that allow it to perform well under most circumstances. Reliance on these shortcuts, however, leaves people susceptible to all manner of illusions: visual, mnemonic, and judgmental.

Id.

22. Following the terrorist attacks of September 11, 2001, many people overestimated the likelihood of another airline hijacking and underestimated the risks associated with the alternative of driving. By one estimate, an extra 1595 persons lost their lives due to increased car travel following 9/11. *See* Gerd Gigerenzer, *Out of the Frying Pan into the Fire: Behavioral Reactions to Terrorist Attacks*, 26 RISK ANALYSIS 347 (2006).

23. *See generally* Jonathan Klick & Gregory Mitchell, *Government Regulation of Irrationality: Moral and Cognitive Hazards*, 90 MINN. L. REV. 1620 (2006) (discussing the way government regulates lay citizens).

24. The subtitle of Malcolm Gladwell's phenomenally successful popular science treatment of psychological research on judgment and decision making illustrates the importance that has been assigned to rapid, intuitive thought in much of recent psychological research. *See* MALCOLM GLADWELL, BLINK: THE POWER OF THINKING WITHOUT THINKING (2005).

25. *E.g.*, Jeremy A. Blumenthal, *Emotional Paternalism*, 35 FLA. ST. U. L. REV. 1, 52 (2007).

potential biases are seen as ineffective because we lack the capacity to detect our biases and to govern our unruly unconscious through intentional thought and action.

This discounting of self-correction processes within legal scholarship reflects a similar discounting within much psychological scholarship.[26] Empirical studies of rationality have often emphasized how difficult it is for subjects to recognize and avoid errors, even when warned that the errors might occur. Empirical studies of automatic categorization and evaluation processes have likewise emphasized the limited control that people can exert over these automatic processes.[27] Nevertheless, while pessimistic views about the potential for self-correction persist,[28] recent research has softened the strong view against the potential for self-correction in both domains, primarily by demonstrating that conscious awareness of bias is not a necessary precondition to self-correction.

Conscious vigilance and deliberate introspection certainly can lead to efforts to avoid bias, but we now know that bias avoidance can also occur as a result of vague or inchoate thoughts, feelings operating at the fringe of consciousness, and even through processes operating fully below the level of consciousness. The figure below arranges the sources of self-correction along a rough continuum of consciousness, each of which will be discussed with respect to its implications for rational behavior and intergroup relations.

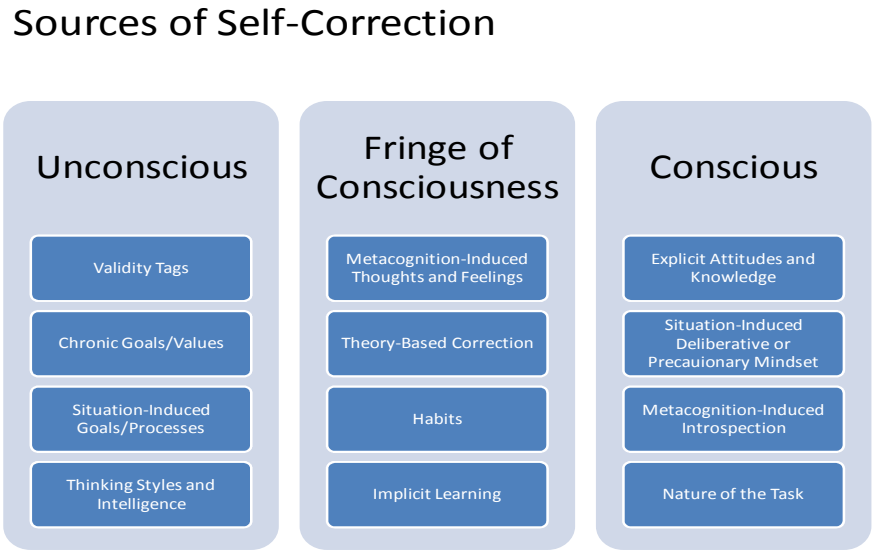
> Further, in order for self-correction to have a chance to succeed, the individual must be aware of the bias and must be motivated to correct it. There is some evidence that when awareness and motivation are both present, cognitive biases can be attenuated. But people typically assume that they are unbiased.

Id.

26. Behavioral law and economics scholarship recently acknowledged that individuals often do deliberate and reach different judgments and decisions than those prompted by intuitive responses to stimuli. For instance, Guthrie and colleagues recently proposed the "Intuitive-Override Model of Judging," in which some judges naturally engage in deliberation to police their intuitions and others can be encouraged to do so through legal structures. *See* Chris Guthrie et al., *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1 (2007).

27. *See* Irene V. Blair, *The Malleability of Automatic Stereotypes and Prejudice*, 6 PERSONALITY & SOC. PSYCHOL. REV. 242, 242 (2002) ("People may often not be aware of what they are doing, they might even intend to be doing something else; perhaps worst of all, the operation of stereotypes and prejudice may be outside of their control." (citation omitted)).

28. *E.g.*, John A. Bargh, *The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects*, *in* DUAL-PROCESS THEORIES IN SOCIAL PSYCHOLOGY 361, 378 (Shelly Chaiken & Yaacov Trope eds., 1999) ("Once a stereotype is so entrenched that it becomes activated automatically, there is really little that can be done to control its influence."); Wilson et al., *supra* note 4, at 190 ("Although the question of how often people detect bias and try to correct for it in everyday life is difficult to answer, we are not optimistic.").



The middle, fringe-consciousness category, in particular, should be treated as a fuzzy category. This middle category encompasses feelings and thoughts that have discernible content but indiscernible origins, such as the "tip of the tongue" phenomenon associated with the near-comprehension of memory items and feelings of low confidence associated with an impression, factual judgment, or risky choice.[29] Also in this middle category are learned habits of thought and behavior whose self-corrective effects usually go unnoticed until those habits become the focus of attention.[30] Mental associations and thought patterns that are the result of implicit learning fall into this middle fringe category, not because they may sometimes be consciously detected, but because some argue that implicit learning is an unconscious process. However, debate continues over the extent to which implicit learning constitutes unconscious versus simply hard-to-detect learning.[31] Rather than seeing the phenomena as occurring fully above or

29. *See, e.g.*, Mark C. Price & Elisabeth Norman, *Intuitive Decisions on the Fringes of Consciousness: Are They Conscious and Does It Matter?*, 3 JUDGMENT & DECISION MAKING 28, 30-31 (2008).

> [E]xperience-based metacognitive judg[]ments are based on rapid automatic inferences that are in one sense conscious and in another sense non-conscious. For example, we might have a *Feeling of Knowing* that we would be able to recogni[z]e the correct answer to a question that we cannot currently recall. The feeling is non-conscious in the sense that we do not have detailed conscious access to its information processing antecedents . . . . It is nevertheless conscious in that there is a distinct phenomenology—something it feels like to have the feeling.

Id. (citations omitted). I borrow the term "fringe of consciousness" from Price and Norman, who adapted this concept to describe phenomena that inhabit this middle ground territory between conscious or unconscious thought. *See id*. at 32-34.

30. For instance, the adaptive practice of habitually saving word-processing files at regular intervals often goes unnoticed until an untimely power outage brings that habit into focus.

31. "Implicit learning" refers to "where complex regularities in our environment are learned without full

below the threshold of conscious thought, the category boundaries in the figure should be seen as porous, with phenomena assigned to the different categories capable of occurring at different levels of consciousness depending on the circumstances of a particular judgment or decision.

1. Overcoming Thoughts that Might Lead to Irrational Behavior

Within the domain of rational behavior, studies of individual differences in rationality have led to the realization that sometimes large numbers of people avoid what were once thought to be pervasive and hard-to-avoid biases in judgment and decision-making. This avoidance of bias occurs in part because some individuals naturally possess greater ability (as measured typically by intelligence) and motivation (as measured by various scales designed to tap into different thinking styles) than others to engage in debiasing deliberative thought absent any extrinsic incentives or conscious decision to do so.[32] Moreover, some individuals naturally feel greater uncertainty about their judgments than others and, as a result, are more motivated to monitor and correct their judgments.[33]

Bias avoidance also occurs because some situations prompt debiasing deliberative thought across a wide range of people, such as the self-critical reflection prompted by knowledge that one's judgments and decisions must be explained to an audience with unknown views,[34] or by an organization's standard operating procedures.[35] These situational influences on rationality tend to operate

---

awareness of what has been learned, or sometimes even without any awareness that learning has occurred at all." Price & Norman, *supra* note 29, at 32. People appear to learn from their experiences, for instance, that the accuracy of numerical estimates increases with sample size. *See* Peter Sedlmeier, *From Associations to Intuitive Judgment and Decision Making: Implicitly Learning from Experience*, *in* THE ROUTINES OF DECISION MAKING 83, 91 (Tilmann Betsch & Susanne Haberstroh eds., 2005). "There has been a vigorous debate over whether so-called implicit learning is really based on non-conscious learning, or could instead be mediated by consciously learned fragments of the target knowledge." Price & Norman, *supra* note 29, at 32.

32. *See, e.g.*, Donna Torrens et al., *Individual Differences and the Belief Bias Effect: Mental Models, Logical Necessity, and Abstract Reasoning*, 5 THINKING & REASONING 1 (1999); Andrew F. Simon et al., *Decision Framing: Moderating Effects of Individual Differences and Cognitive Processing*, 17 J. BEHAV. DECISION MAKING 77 (2004); Shoshana Shiloh et al., *Individual Differences in Rational and Intuitive Thinking Styles as Predictors of Heuristic Responses and Framing Effects*, 32 PERSONALITY & INDIVIDUAL DIFFERENCES 415 (2002). *See generally* KATHLEEN M. GALOTTI, MAKING DECISIONS THAT MATTER: HOW PEOPLE FACE IMPORTANT LIFE CHOICES 117-25 (2002) (providing an overview of research on individual differences in decision-making styles); Mitchell, *Equal Incompetence*, *supra* note 16, at 83-105 (discussing research on individual differences in rationality).

33. *See* Leigh Ann Vaughn & Gifford Weary, *Causal Uncertainty and Correction of Judgments*, 39 J. EXPERIMENTAL SOC. PSYCHOL. 516, 522-23 (2003).

34. *See* Jennifer S. Lerner & Philip E. Tetlock, *Accounting for the Effects of Accountability*, 125 PSYCHOL. BULL. 255, 256-57 (1999).

35. *See generally* Chip Heath et al., *Cognitive Repairs: How Organizational Practices Can Compensate for Individual Shortcomings*, 20 RES. ORGANIZATIONAL BEHAV. 1 (1998). Cutting across the individual/ situational differences divide is the role of enculturation, which plays an important role in one's tendencies toward rationality. *See* RICHARD E. NISBETT, THE GEOGRAPHY OF THOUGHT: HOW ASIANS AND WESTERNERS THINK DIFFERENTLY . . . AND WHY (2003). *See* Mitchell, *Equal Incompetence*, *supra* note 16, at 105-19 (discussing research on situational differences in rationality).

with little conscious awareness that a different mode of thought has been activated, though some situations will make the importance of a decision or judgment palpable, leading to conscious effortful deliberation, or to an intentionally precautionary mindset. In other situations, one may become aware of the effort being exerted through observation of one's own behavior. Thus, whereas individual predispositions operate largely beyond consciousness, situational influences to engage in debiasing efforts operate above and below the threshold of consciousness.

The shift to a more deliberative, rational mode of thought may also be triggered by thoughts and feelings operating at the fringe of consciousness that arise from metacognitive monitoring of one's own thought processes.[36] In particular, the ease with which information can be accessed or processed provides an important cue to the possible need for self-correction.[37] With respect to the accessibility of information:

> Thoughts about focal aspects of a given issue produce bias when they are easy to bring to mind, but reduce bias when they are difficult to bring to mind; conversely, thoughts about alternatives reduce bias when they are easy to bring to mind, but produce bias when they are difficult to bring to mind. Hence, encouraging people to "consider the opposite" can be a successful debiasing strategy when consideration of the opposite is experienced as easy; the strategy backfires when consideration of the opposite is difficult—and even furrowing one's brow can be enough to produce a backfire effect. Finally, encouraging people to generate many focal thoughts can, paradoxically, be a successful debiasing, strategy, provided that thought generation is difficult.[38]

Thus, asking subjects to think of three ways in which an event might have turned out differently—an easy task in terms of information accessibility—leads to debiasing of the hindsight bias (knowledge of an outcome biases judgments about the likelihood of that outcome). Whereas asking subjects to think of twelve alternative outcomes—a hard task in terms of information accessibility—does

---

36. In addition to thinking of metacognition as thoughts about thoughts, it may be defined as "subjective experiences that accompany the thinking process." Lawrence J. Sanna & Norbert Schwarz, *Metacognitive Experiences and Human Judgment: The Case of Hindsight Bias and Its Debiasing*, 15 CURRENT DIRECTIONS PSYCHOL. SCI. 172, 172-73 (2006).

37. *See* Norbert Schwarz et al., *Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns*, 39 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 127, 132 (2007). Another important metacognitive feeling is that of a feeling of rightness or wrongness in one's judgments or choices. "One source of rightness or wrongness feelings is experiencing a good or a poor fit between one's regulatory focus and one's goal pursuit strategies." Leigh Ann Vaughn et al., *When Two Wrongs Can Make a Right: Regulatory Nonfit, Bias, and Correction of Judgments*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 654, 655 (2006).

38. Schwarz et al., *supra* note 37, at 143.

not lead to debiasing, and may even exacerbate the bias.[39] The ease with which alternative outcomes can be generated mediates the expression of the hindsight bias: when an information-generation task is easy, people tend to see the actual outcome as less striking and certain when viewed in hindsight; when the task is hard, they tend to see the actual outcome as more deterministic.[40] With respect to the fluency with which information can be processed:

> [P]eople correctly assume that familiar information is easier to process than novel information. Applying this naïve theory of mental processes, they infer from experienced processing fluency that the information is familiar—even when the fluency derives from presentation variables, like good figure-ground contrast or long exposure times, or from contextual influences, like preceding semantic primes. In memory experiments, this fluency-familiarity link gives rise to the erroneous "recognition" of fluently processed but previously unseen stimuli.
>
> More important for the present purposes, the perceived familiarity of information influences the likelihood that the information is accepted as true or flagged for closer scrutiny. . . . [J]udgments are likely to be consistent with the implications of fluently processed declarative information, but inconsistent with the implications of disfluently processed declarative information.[41]

Furthermore, "high fluency is experienced as hedonically positive," and "[t]his affective response can itself serve as a basis for judgment and fluently processed stimuli are evaluated more favorably than less fluently processed ones."[42] Accordingly, making information difficult to read (a low processing fluency manipulation), for instance, leads to a feeling of unfamiliarity and prompts closer scrutiny of information that might otherwise produce flawed intuitive judgments.[43]

An important corollary to the proposition that metacognitive experiences mediate irrational behavior is the proposition that individuals develop naïve theories about the meaning of their metacognitive experiences and how they should react to these experiences.[44] If a naïve theory causes the individual to

---

39. *See id*. at 138.

40. The ease with which information can be generated has also been shown to affect the likelihood of overconfidence and the planning fallacy (the tendency to underestimate time to complete a project). *See id*. at 135-36, 139-40.

41. *Id*. at 143 (citations omitted).

42. *Id*. at 132 (citation omitted).

43. Providing information about an outcome in a difficult to read format reduces the hindsight bias as well as susceptibility to distorted questions (for example: "How many animals of each kind did Moses take on the Ark?"). *See id*. at 144-45.

44. For example:

> [People] assume that information that is well represented in memory is easier to recall than

second-guess his or her intuitive response, the metacognitive experience is likely to lead to corrective processes; if the theory leads to an interpretation of the metacognitive experience that reinforces the intuitive response, the bias is likely to persist and perhaps become stronger.[45]

Much experience-based learning, which can lead to habitual responses to tasks and new associations[46] takes place on the fringe of consciousness as well.[47] A dynamic approach to judgment and decision-making recognizes that, through education, experimentation, and feedback, individuals learn which choices and judgments are most likely to produce desirable outcomes and develop competence in their ability to compile, rank, and make those choices.[48] Because much of this learning occurs without full awareness and affects automatic associative processes, the products of experience-based learning are often intuitive responses to judgment and decision problems.[49] These intuitions,

---

> information that is poorly represented, making ease of recall a cue for memory judgments; that recent events are easier to recall than distant events, making ease of recall a cue for temporal distance; that important events are easier to recall than unimportant ones; and that thought generation is easier when one has high rather than low expertise, making ease a cue for importance and expertise.

Id. at 154 (citations omitted).

45. If the metacognitive experience is discredited as a source of information—by giving a reason to believe that information should not be accessible or should be hard to process—then it will not affect the likelihood of bias. *See id*. at 133.

46. Such learning can be particularly helpful in the context of frequency judgments. *See* Sedlmeier, *supra* note 31, at 87.

> The basic input to associative learning is repeatedly occurring events. Therefore, it seems to be the natural basis for all kinds of frequency-related judgments. An overview of the literature reveals that intuitive frequency judgments, especially judgments of relative frequency, are quite accurate if the encoding process is not biased in any systematic way.

Id. (citations omitted).

47. I follow here Verplanken and colleagues' definition of habits as "learned sequences of acts that have become automatic responses to specific cues, and are functional in obtaining certain goals or end states." Bas Verplanken et al., *The Measurement of Habit*, *in* THE ROUTINES OF DECISION MAKING, *supra* note 31, at 231, 231 (internal quotations omitted). "In terms of social cognition models, habits may be considered as behavior that is guided by implicit structures like schemas or implicit attitudes, rather than by explicit evaluations of behavior or conscious decision making." *Id.* at 232 (citation omitted). The key here, from a rationality perspective, is the assumption that a habit is functional in obtaining a goal. The social desirability or wisdom of the goal may be debated, but the habit arises because it is effective at goal attainment. This distinction parallels the distinction between real world success and coherence norms for the evaluation of judgments and decisions, with the latter being emphasized in studies of rationality. *See* Mitchell, *Taking Behavioralism Too Seriously?*, *supra* note 16, at 1996-2002.

48. *See* Klick & Mitchell, *supra* note 23, at 1627-38 (discussing learning in judgment and decision making, with an emphasis on Byrnes' self-regulation model of decision making); *see also* Adam S. Goodie & Diana L. Young, *The Skill Element in Decision Making Under Uncertainty: Control or Competence?*, 2 JUDGMENT & DECISION MAKING 189, 202 (2007) ("The unique contribution of control to decision making has important theoretical and applied implications, suggesting that decisions may be influenced by the opportunity to improve at tasks that can be learned, even at short-term expected loss, in order to create more advantageous prospects in the future.").

49. For one theory of the relation between implicit learning and intuition, see Matthew D. Lieberman, *Intuition: A Social Cognitive Neuroscience Approach*, 126 PSYCHOL. BULL. 109, 126 (2000) ("The claim of this review is that intuition is a phenomenological and behavioral correlate of implicit learning . . . ."). For

however, will often differ from intuitions prompted by cognitive heuristics, which involve the substitution of simple processes for computationally more complex processes. Thus, while the latter, heuristic-based intuitions, may lead to systematic biases where the simpler computation produces an answer at odds with the proper product of the more complex computation, the former, experiential intuitions, may lead to quite accurate judgments and satisfactory choices.

Although individuals often fail to identify accurately all of the influences on their judgments and decisions, and thus fail to be conscious of biasing influences, they may nonetheless engage in deliberation due to the nature of the task without any extra situational prompting or special internal predisposition to do so. Choice behavior in particular, where competing alternatives are apparent, will often involve deliberation.[50] While mere deliberation will not overcome all reasoning fallacies—for some problems are computationally difficult even with effort and full attention, and some decision structures are confusing or deceptive to even the most careful observer[51]—pre-decision deliberation does tend to shift one toward

---

discussions of how conscious or explicit processes may become automatized, see John R. Anderson, *Automaticity and the ACT* Theory*, 105 AM. J. PSYCHOL. 165 (1992) and Gordon D. Logan, *Toward an Instance Theory of Automatization*, 95 PSYCHOL. REV. 492 (1988).

50. *See* Itamar Simonson, *In Defense of Consciousness: The Role of Conscious and Unconscious Inputs in Consumer Choice*, 15 J. CONSUMER PSYCHOL. 211, 212 (2005) ("[T]he notion that automatic (System 1) influences are the default, with relatively infrequent override by conscious (System 2) processes may fit many psychological phenomena but does not adequately describe choice, where System 2 is usually the primary influence."); *id*. ("[C]onsciously considered inputs tend to play a major role in choice (including consumer choice), and . . . many potential unconscious influences in typical consumer-choice environments (e.g., in stores) create high 'noise' level and potential interactions that tend to diminish the measurable significance of unconscious relative to conscious choice inputs.").

51. The Monty Hall problem provides perhaps the best example of a decision structure that evades understanding. Derived from the name of the host on the game show "Let's Make a Deal," the problem involves choosing among three doors, one of which has a prize behind it. After the initial door choice, one of the two remaining doors is opened to reveal no prize behind that door; the player is then given the option of sticking with the initial choice or switching to the remaining, unopened door. The rational choice is to switch because the odds that the initial door conceals the prize remain at 1/3 while the odds of the prize being behind the other, unopened door are 2/3, because we know that the third, now opened door did not conceal the prize. However, without experience playing the game, few people switch because few people comprehend that important information has been revealed by opening one of the two unchosen doors. For instance, Granberg and Brown found that, in a single-trial study, only 13% of all subjects chose to switch doors. *See* Donald Granberg & Thad A. Brown, *The Monty Hall Dilemma*, 21 PERSONALITY & SOC. PSYCHOL. BULL. 711, 713 (1995). In a multi-trial study, they found that only 10% switched doors on the initial trial when there was no incentive to switch (i.e., success after sticking paid the same as success after switching), but 43% switched on the first trial where success after switching yielded double points toward a prize and 36% switched on the first trial where success after switching yielded quadruple points compared to success after sticking. *Id*. at 714. Considering only the last ten of the fifty trials, Granberg and Brown found that 55% of these decisions were switches in the condition where success after sticking and switching paid the same, 73% were switches in the condition where success after switching paid double, and 88% were switches in the condition where success after switching increased points fourfold. *Id*. at 714-15. Thus, with experience, players can learn that switching is the optimal choice. However, as demonstrated by another study that found a similar improvement in choice behavior over time, this improvement appears to come from monitoring outcomes rather than gaining insight into the probability structure of the outcomes in the game. *See* Talia Ben-Zeev et al., Increasing Working Memory Demands Improves Probabilistic Choice But Not Judgment on the Monty Hall Dilemma 13 (unpublished manuscript on

rule-based reasoning and away from heuristic-based reasoning.[52] Pre-decision deliberation may also lead to metacognitive recognition of one's shortcomings and uncertainties, which in turn may lead to precaution. One response to such a precautionary mindset may be greater reliance on mechanical rules and decision aids, which makes it more likely that simple computational errors and inappropriate weighting of data points will be avoided.[53] And, of course, once one's thoughts are reduced to writing or expressed verbally, one will often check the answer against conscious beliefs and attitudes, attempting corrections if the situation allows.

2. Overcoming Thoughts that Might Lead to Discriminatory Behavior

Second thoughts play an equally important debiasing role within the domain of interpersonal relations. Studies into individual and situational differences in automatic responses to members of various demographic and social groups have revealed that

> automatic stereotypes and prejudice can be moderated by a wide variety of events, including, (a) perceivers' motivation to maintain a positive self-image or have positive relationships with others, (b) perceivers' strategic efforts to reduce stereotypes or promote counterstereotypes, (c) perceivers' focus of attention, and (d) contextual cues. In addition, the research shows that group members' individual characteristics can influence the extent to which (global) stereotypes and prejudice are automatically activated.[54]

This research indicates that we should not assume that automatic appraisal processes will inevitably reflect divisions along demographic lines: the

---

file with the *McGeorge Law Review*) ("[P]articipants who decided to switch doors most often held the belief that the probability of winning the prize was still .50. This finding suggests a dissociation between choice and probability judgment; making a correct probabilistic choice does not entail an understanding of the underlying probabilities."). The abstract version of Wason's four-card selection task is also notorious for its difficulty, though contextualized versions of the task are much easier. *See* Mitchell, *Taking Behavioralism Too Seriously?*, *supra* note 16, at 1986-88.

52. Post-judgment deliberation may lead some to reconsider their initial responses and provide better responses following second thoughts, but it may lead others simply to create more elaborate justifications of their initial responses. *See* Jody M. Shynkaruk & Valerie A. Thompson, *Confidence and Accuracy in Deductive Reasoning*, 34 MEMORY & COGNITION 619, 629-30 (2006).

53. *See* Hal R. Arkes & Virginia A. Shaffer, *Should We Use Decision Aids or Gut Feelings?*, *in* HEURISTICS AND THE LAW 411 (Gerd Gigerenzer & Christoph Engel eds., 2006) (discussing the conditions under which decision aids typically outperform intuitive judgment). Use of decision aids absent some form of motivation, such as a lack of confidence in one's independent ability to solve a problem or organizational directives, is, unfortunately, often quite low. *See* Winston R. Sieck & Hal R. Arkes, *The Recalcitrance of Overconfidence and Its Contribution to Decision Aid Neglect*, 18 J. BEHAV. DECISION MAKING 29, 30 (2005).

54. Blair, *supra* note 27, at 255.

unconscious is less prejudiced and less stereotype-driven than many psychologists and legal scholars have assumed.

Just as people possess naïve theories about the circumstances under which their behavior may be irrational, they possess naïve theories about when their judgments about others are likely to be biased.[55] Within the domain of social judgments, where people are engaged in interpreting social interactions and evaluating others, people develop theories about what constitutes good and bad outcomes and attitudes toward different types of people and they develop metacognitive theories about when these interpretive theories and attitudes should be applied.[56] We learn to avoid certain types of thoughts because of their personally or socially aversive nature, and we develop theories about the biasing influence of different stimuli and conditions so that we can adjust our beliefs and correct our judgments. These second-order thoughts greatly affect how first-order thoughts are interpreted and encoded and whether first-order thoughts control the outputs of mental processing.[57]

One important naïve theory about interpersonal bias and the need for self-correction is that judgments about women and minorities may be tainted by improper considerations.[58] Accordingly, people may engage in concerted efforts

---

55. *See, e.g.*, Vincent Y. Yzerbyt et al., *Social Judgeability and the Dilution of Stereotypes: The Impact of the Nature and Sequence of Information*, 23 PERSONALITY & SOC. PSYCHOL. BULL. 1312, 1319 (1997) [hereinafter Yzerbyt et al., *Dilution of Stereotypes*] ("In accordance with the naïve theory that one is not supposed to judge a specific individual on the basis of one's stereotypes, a clear pattern of dilution emerged."); Vincent Y. Yzerbyt et al., *Social Judgeability: The Impact of Meta-Informational Cues on the Use of Stereotypes*, 66 J. PERSONALITY & SOC. PSYCHOL. 48 (1994) [hereinafter Yzerbyt et al., *Social Judgeability*].

> The core idea of the present series of studies is that, when making a social judgment, people are not only influenced by cognitive and motivational factors but also by social rules that declare the target judgeable.
>
> . . . .
>
> Another rule, tested in the present studies, is that one should not judge a given individual on the basis of stereotypical information only. This in no way means that the stereotype is not activated, but that it is not considered valid if no other information is available.

Yzerbyt et al., *Social Judgeability*, *supra*, at 53-54.

56. *See, e.g.*, Christine Hepburn & Anne Locksley, *Subjective Awareness of Stereotyping: Do We Know When Our Judgments Are Prejudiced?*, 46 SOC. PSYCHOL. Q. 311, 312, 317 (1983) ("The data indicated that subjects relied on a variety of external cues to infer how much their own stereotypic beliefs were influencing their target judgments. . . . The general effect of these inferential strategies is to overestimate actual effects of stereotyping on judgments of individuals."). In addition to metacognitive doubt, these theories may be activated by external cues, which may provide poor estimates of true bias. *Id.* at 317.

57. *See* Richard E. Petty et al., *The Role of Metacognition in Social Judgment*, *in* SOCIAL PSYCHOLOGY: HANDBOOK OF BASIC PRINCIPLES 254, 270 (Arie W. Kruglanski & E. Tory Higgins eds., 2007) ("Individuals' evaluations of their thoughts and perceptions can have sweeping effects on judgment and behavior.").

58. For example, Sczesny and Kühnen found evidence of an expectation that gender stereotypic beliefs might bias personnel judgments and found evidence of subsequent self-correction for such potential bias in judgments of hypothetical applicants. However, they did not find evidence for an expectation of, or correction for, the biasing influence of applicants' physical appearance even though appearance did bias judgments. *See* Sabine Sczesny & Ulrich Kühnen, *Meta-Cognition About Biological Sex and Gender-Stereotypic Physical Appearance: Consequences for the Assessment of Leadership Competence*, 30 PERSONALITY & SOC. PSYCHOL. BULL. 13, 15, 18 (2004).

to be even-handed when they become aware of demographic or social group information, to the point that some overcorrect for potential bias.[59] Aberson and Dora found, for instance, that heterosexual college students who had little contact with gay persons engaged in self-overcorrection when reviewing hypothetical candidates for a job, such that they rated "flawed (alcoholic) gay candidates the same as unflawed gay candidates" but they rated flawed heterosexual candidates worse than unflawed heterosexual candidates.[60] Yet college students with gay friends rated the flawed gay candidate worse than the unflawed gay candidate, suggesting less concern on the part of these subjects that they would be biased against gay persons as a group. As Aberson and Dora say,

> [c]ontact mitigated the overcorrection effect. Individuals with gay friends do not overcorrect. This group apparently ignored category information such as gay/heterosexual and only varies evaluations based on candidate flaws. As such, individuals with gay male friends exhibited truly egalitarian ratings.[61]

The impetus to engage in self-correction may be a conscious concern about one's bias toward a particular group or, as in the domain of rationality, vague concerns operating at the fringe of consciousness. "[T]he mere feeling of a contextual influence may trigger a correction process,"[62] and discrepancies between how one thinks one *will* respond to a member of another group versus how one thinks one *should* respond can give rise to feelings of general discomfort and even guilt that motivate pre-emptive self-correction efforts.[63]

---

59. "[T]o the extent that potential for bias is salient, people become more oriented toward taking steps to identify and avoid any biases at work." *Id*. For example, Sommers and Ellsworth have found that making race salient in mock jury settings eliminates racial biases in White jurors' decision making, whereas racial bias appeared when race was not salient. In other words, under conditions that triggered self-concerns about expressions of bias, White jurors were vigilant and appeared to self-correct. *See* Samuel R. Sommers & Phoebe C. Ellsworth, *White Juror Bias: An Investigation of Prejudice Against Black Defendants in the American Courtroom*, 7 PSYCHOL. PUB. POL'Y & L. 201 (2001).

60. Christopher L. Aberson & Jessica Dora, *Reactions to Gay Men: Contact and Overcorrection in an Employee Selection Simulation*, 22 CURRENT PSYCHOL.: DEVELOPMENTAL LEARNING PERSONALITY SOC. 164, 172 (2003).

61. *Id.*

62. Lorella Lepore & Rupert Brown, *The Role of Awareness: Divergent Automatic Stereotype Activation and Implicit Judgment Correction*, 20 SOC. COGNITION 321, 346 (2002). Brown and Lepore found that subjects in an experiment who suspected that there was a relation between the first task, in which a supraliminal prime relating to race was presented, and a second task, in which subjects were asked to form impressions about persons of different races, corrected for possible race bias in their impressions, but those who saw no connection between the tasks did not (though correction by suspicious but low-prejudiced subjects actually resulted in more negative judgments about targets, whereas correction by suspicious but high-prejudiced subjects resulted in less negative judgments about targets, with both suspicious groups having similar judgments following correction). *See id*. at 344-47.

63. *See* Margo J. Monteith & Aimee Y. Mark, *Changing One's Prejudiced Ways: Awareness, Affect, and Self-Regulation*, 16 EUR. REV. SOC. PSYCHOL. 113, 149 (2005).

> Our research suggests that many people are aware of their proneness to prejudice-related discrepancies, that they feel bad about themselves for having prejudice responses, and that they can

Furthermore, recent research establishes that self-correction for intergroup bias can occur without conscious or even semi-conscious awareness of the bias. As Glaser and Kihlstrom explain, this research is radically altering our understanding of the unconscious:

> Evidence from studies of automatic affect and cognition suggests that, in addition to the ability to process the meaning of, categorize, and evaluate perceived stimuli automatically, the human mind is capable of maintaining unconscious vigilance over its own automatic processes. This suggests a volitional nature of the unconscious, an idea that to many may seem self-contradictory. . . . That goals can operate at the unconscious level, and subsequently influence explicit judgments and behaviors, is now well demonstrated. . . . This thesis, and the findings supporting it, represents a departure from traditional conceptions of the unconscious as passive and reactive, suggesting an unconscious that is, paradoxically, "aware."[64]

This research establishes that chronic goal orientations, such as a commitment to egalitarianism, can serve as important checks on intergroup bias at both the conscious and unconscious levels of responding to other groups.[65]

> build associations between this affect, environmental stimuli present when prejudiced responses occur, and prejudiced responses themselves, so as to establish cues for control. The presence of these cues appears to signal the need for behavioural inhibition and prospective reflection, which enables people to interrupt ongoing responding when prejudiced responses might otherwise occur and generate alternative, more personally acceptable responses.

Id.

64. Jack Glaser & John F. Kihlstrom, *Compensatory Automaticity: Unconscious Volition Is Not an Oxymoron*, *in* THE NEW UNCONSCIOUS 171, 189-90 (Ran R. Hassin et al. eds., 2005) (citations omitted).

65. *See, e.g.*, Patricia G. Devine et al., *The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice*, 82 J. PERSONALITY & SOC. PSYCHOL. 835, 845 (2002).

> [W]e expected that participants who reported high levels of internal motivation and low levels of external motivation and, thus, were theoretically highly autonomous would be the most effective at regulating expressions of race bias, even on difficult-to-control responses. Consistent with our expectations, these individuals responded with lower levels of implicit race bias than did all other participants.

Id.

> These results demonstrate that control over stereotype activation is being exerted by chronics; the failure to use stereotypes cannot be due to an effortful process of correcting or debiasing one's judgments. This would suggest a change in how terms such as *intended* and *deliberate* are used in the literature, so that they are not equated with consciousness; one can exert the will in an intentional fashion but without effortful processing.

Gordon B. Moskowitz et al., *Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals*, 77 J. PERSONALITY & SOC. PSYCHOL. 167, 180 (1999) (citations omitted).

> We expected internal motivation to predict lower implicit prejudice, inasmuch as internal motivation represents a chronic goal to be nonprejudiced that derives from one's personal values. Moreover, it was expected that external motivation would predict greater implicit prejudice, and that this effect would be mediated by efforts to control prejudiced responses. . . . Our results were consistent with these expectations. Internal motivation was negatively related, and external motivation was positively related, to implicit prejudice.

Importantly, such correction does not appear to require controlled processing, as persons internally motivated to control prejudice have been shown to self-correct even when cognitively busy.[66] Rather, goals to which one has a long-standing commitment and which are regularly endorsed and implemented at the conscious level appear to create routines and mental structures that operate automatically at the unconscious level to achieve those goals.[67]

But it is not only consciously-endorsed, chronic values that can counter automatic stereotyping and prejudice processes. The priming of temporary goals and situational or social norms that are inconsistent with stereotyping or prejudice may also lead to self-correction at subconscious levels.[68] For instance,

---

Leslie R. M. Hausmann & Carey S. Ryan, *Effects of External and Internal Motivation to Control Prejudice on Implicit Prejudice: The Mediating Role of Efforts to Control Prejudiced Responses*, 26 BASIC & APPLIED SOC. PSYCHOL. 215, 222 (2004).

66. *See* Devine et al., *supra* note 65, at 846 ("In Study 3, we introduced a cognitive busyness manipulation that theoretically should have disrupted any control efforts that required cognitive resources. Even when cognitively taxed, the high internal, low external participants reported much lower levels of bias on the implicit measure than did any of the other participants.").

67. The precise mechanisms whereby the correction occurs remains unclear and may involve the avoidance of activation of stereotypes or automatic inhibition of them through practice and/or the creation of new mental structures, such as counter-associations or links between group categories and egalitarian goals, the alteration of existing structures by weakening existing associations, or the prevention of biased structures in the first place. *See id*.; William W. Maddux et al., *Saying No to Negativity: The Effects of Context and Motivation to Control Prejudice on Automatic Evaluative Responses*, 41 J. EXPERIMENTAL SOC. PSYCHOL. 19, 32-33 (2005); Moskowitz et al., *supra* note 65, at 180-81; Gordon B. Moskowitz et al., *Preconsciously Controlling Stereotyping: Implicitly Activated Egalitarian Goals Prevent the Activation of Stereotypes*, 18 SOC. COGNITION 151, 171-72 (2000).

68. *See, e.g.*, Brian S. Lowery et al., *Social Influence Effects on Automatic Racial Prejudice*, 81 J. PERSONALITY & SOC. PSYCHOL. 842, 851 (2001).

> Taken together, the four experiments presented here provide evidence that automatic racial attitudes are subject to both tacit and explicit social influence. Across two measures of automatic prejudice, European Americans but not Asian Americans exhibited less automatic prejudice in the presence of a Black experimenter than a White experimenter. Experiment 3 demonstrated that explicit experimental instruction to avoid racism reduced automatic prejudice for European Americans and Asian Americans alike.

Id.; Gordon B. Moskowitz et al., *The Implicit Volition Model: On the Preconscious Regulation of Temporarily Adopted Goals*, 36 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 317, 396 (2004) ("People who had fairness goals triggered by contemplating such a prior experience (thinking about failures at being egalitarian) did not show activation of the stereotype. They compensated for the triggered goal by engaging in goal-compatible responses such as inhibiting the stereotype."); Kai Sassenberg & Gordon B. Moskowitz, *Don't Stereotype, Think Different! Overcoming Automatic Stereotype Activation by Mindset Priming*, 41 J. EXPERIMENTAL SOC. PSYCHOL. 506, 511-12 (2005) (stating that situational cues to think creatively activated a "think different" mindset in which stereotypic associations were inhibited, not activated, or not retrieved); Stacey Sinclair et al., *Social Tuning of Automatic Racial Attitudes: The Role of Affiliative Motivation*, 89 J. PERSONALITY & SOC. PSYCHOL. 583 (2005).

> The reported experiments provide clear support for the notion that automatic racial attitudes are subject to affiliative social tuning. Across two experiments, two measures of automatic prejudice, and three operationalizations of affiliative motivation, automatic prejudice shifted toward the ostensible attitudes of a social actor to the degree that individuals were motivated to get along with him or her. . . . This research adds to the growing literature demonstrating that automatic attitudes flexibly respond to social motives.

Sinclair et al., *supra*, at 590.

the mere presence of others, even those within one's in-group, has been found to lead to automatic reductions in measures of implicit prejudice toward an out-group; the presence of others apparently primes thoughts about equality and fairness that offset prejudicial associations.[69]

Petty and colleagues discuss another means whereby second thoughts may come to override first thoughts. In their metacognitive model of attitudes, metacognitive information about one's association of objects to evaluative information can be stored in the form of "validity tags,"[70] which create validation (or invalidation) links between first- and second-order thoughts. These tags make judgments and decisions more efficient by streamlining the metacognitive oversight of first-order thoughts, with validity tags confirming first thoughts, negating them, or triggering inhibitory processes.[71] Thus, while validation checking may occur online through conscious and fringe-consciousness processes, it may also occur offline, through automatic activation of validity tags.[72] "To the extent that the retrieval of validity tags becomes automatic, it even becomes possible for people to quickly correct for undesired evaluations that might come to mind."[73]

The metacognitive validity tag model provides a particularly useful way of understanding how we deal with living in a world where we encounter considerable information that may turn out to be inaccurate, irrelevant, or lead to behavior contrary to our existing values and beliefs. As Petty explains,

> [a]ssociations can come about for a large number of reasons. One is that the person believes in or endorses the association and may express it

---

69. Luigi Castelli & Silvia Tomelleri, *Contextual Effects on Prejudiced Attitudes: When the Presence of Others Leads to More Egalitarian Responses*, 44 J. EXPERIMENTAL SOC. PSYCHOL. 679, 683-84 (2008).

> The actual presence of other individuals significantly modified spontaneous responses demonstrating that changes in the surrounding social environment were effective in producing a positive shift in automatic intergroup attitudes. . . . In the presence of other persons, egalitarian-related concepts were more easily accessed after the presentation of Black faces demonstrating that the group context triggered the goal of not being prejudiced and behaving fairly toward Blacks.

Id. (citations omitted).

70. Richard E. Petty et al., *The Meta-Cognitive Model (MCM) of Attitudes: Implications for Attitude Measurement, Change, and Strength*, 25 SOC. COGNITION 657, 667 (2007) ("[A] validity tag is a stored form of meta-cognition (i.e., secondary cognition) . . . .").

> [T]he MCM goes beyond the idea that validation is solely an online process and holds that perceived validities, like evaluations themselves, can be stored for later retrieval. In other words, the MCM assumes that just as it is adaptive to store evaluations to guide decision making and action, so too is it adaptive to know if any activated evaluation is a reliable guide.

Id. at 664 (citation omitted).

71. *See* Richard E. Petty & Pablo Briñol, *A Metacognitive Approach to "Implicit" and "Explicit" Evaluations: Comment on Gawronski and Bodenhausen (2006)*, 132 PSYCHOL. BULL. 740, 741 n.4 (2006) ("Endorsement (validation) can be represented as true/false, confidence/doubt, yes/no, or even good/bad. If a validity tag is retrieved along with the evaluative association, then there is no need for an online validation process.").

72. Richard E. Petty, *A Metacognitive Model of Attitudes*, 33 J. CONSUMER RES. 22, 22 (2006).

73. Petty et al., *supra* note 70, at 664.

> often, in which case the association does represent the attitude. Another is that some idea is expressed so often in the culture (e.g., "apples are good") that people have ready access to it even if they do not endorse it personally. Indeed, even if a person believes the opposite, the counterassociation may still be a quick one. Imagine a long-standing egalitarian who engages in diversity training for a living. As part of the training, the person must constantly explain stereotypes to people and then express why these stereotypes are wrong. The ready access to stereotypic content should not lead us to label the person as prejudiced. Rather, the attitude structure may be [one] where a quick association is accompanied by an invalidity tag (which may or may not be retrieved on all occasions).[74]

Thus, the metacognitive model of attitudes distinguishes attitudes from first-order object-evaluation associations, with attitudes including the metacognitive information stored on top of these object-evaluation associations. Just because we may associate women and minorities with negative outcomes as a result of living in a world where women and minorities may, on average, fare worse than White males on a number of outcome measures, we are not cognitively compelled to dislike women or minorities or react negatively to them as a result of these first-order associations.

Finally, it is important to note that explicit, or conscious, attitudes and beliefs perform an important checking role on implicit first thoughts in the domain of intergroup bias, even more so than in the domain of rationality. Whereas some irrational tendencies are difficult to overcome because the judgment or decision problem is computationally complex or because one lacks the basic logical and statistical training needed to decompose the problem and then solve it, figuring out how to avoid responding to another person in a prejudiced or stereotypic way is not computationally complex: one need only consult one's explicit unbiased beliefs and attitudes or be attuned to social norms. Accordingly, relations with other persons do not implicate the kinds of computational problems that the axioms of rationality sometimes do.[75] This ease in trumping biased implicit first thoughts through explicit second thoughts explains why implicit intergroup bias is most likely to have an independent influence on behavior, if at all, in situations where there is little motive or opportunity to monitor or control one's behavior (such as split-second decisions about whether a suspect is holding a weapon).[76]

---

74. Petty, *supra* note 72, at 23 (citations omitted).

75. Of course, intergroup relations may be complex and complicated in many ways, but applying one's explicit beliefs or attitudes to an interaction does not present the kinds of hard-to-grasp computations and dissection of problems that we see in some of the tests of rationality. *See supra* note 51.

76. *See, e.g.*, Russell H. Fazio & Michael A. Olson, *Implicit Measures in Social Cognition Research: Their Meaning and Use*, 54 ANN. REV. PSYCHOL. 297, 304 (2003).

> The MODE model suggests that the magnitude of the relation between an implicit and an explicit measure will depend on the motivation and opportunity to deliberate. If either motivation or

C. Summary

As this growing body of research shows, we are not captives to our automatic first thoughts, whether those thoughts take the form of heuristic responses to judgment and decision-making tasks or unconscious reactions to other persons based on their apparent membership in various demographic groups. Even without specific, conscious awareness that first thoughts may be biasing our judgments and decisions, we employ a variety of corrective second thoughts to combat our irrational and discriminatory tendencies. Thus, not only may we be unaware of biases in our first thoughts, but we may be unaware of our corrections taking place through second thoughts. These second thoughts are certainly not perfect; sometimes they are inadequate, at other times they bolster our first thoughts, and sometimes they result in over-correction. But these second-order corrective processes constitute an important influence on our thoughts and behavior that should not be ignored when assessing our capacity for rational behavior and impartial intergroup relations.

II. THE LAW AND SECOND THOUGHTS

What might this empirical research mean for legal theory? The specific contention here is that this research indicates that current, popular models of judgment and decision-making within antidiscrimination theory and behavioral-law-and-economics theory portray humans in too simplistic a light; these models fail to give sufficient weight to the impact of second thoughts on first thoughts. More broadly, by understanding how second-order thoughts interact with first-order thoughts, we can better predict when biases will result in unwanted behaviors and what kinds of debiasing efforts are likely to succeed.

> opportunity is relatively low at the time that the explicit response is being considered, then explicit measures should correlate with implicit ones. However, when both motivation and opportunity are relatively high, they are less likely to correlate.

Id.; Wilhelm Hofman et al., *Implicit and Explicit Attitudes and Interracial Interaction: The Moderating Role of Situationally Available Control Resources*, 11 GROUP PROCESSES & INTERGROUP REL. 69, 83 (2008) (finding that the race IAT better predicted visual contact with an outgroup member when subjects were operating under memory load than when not); B. Keith Payne, *Weapon Bias: Split-Second Decisions and Unintended Stereotyping*, 15 CURRENT DIRECTIONS PSYCHOL. SCI. 287, 290 (2006) ("Race can bias snap judgments of whether a gun is present, and that bias can coexist with fair-minded intentions."). Note that even under these conditions, the likelihood of discriminatory behavior occurring depends on a host of other factors, such as individual value orientations, levels of intergroup contact, and practice. *See* Jack Glaser & Eric D. Knowles, *Implicit Motivation to Control Prejudice*, 44 J. EXPERIMENTAL SOC. PSYCHOL. 164, 170 (2008) ("[T]hose who implicitly view prejudice as especially bad show no relationship between implicit stereotypes and spontaneous behavior," as measured in a Shooter bias study.); Payne, *supra*, at 290. Note also that high implicit bias has been found to be related to positive interracial encounters, making clear that bias at the level of first-order thoughts is not a direct indicator of biased behavior. *See* J. Nicole Shelton et al., *Ironic Effects of Racial Bias During Interracial Interactions*, 16 PSYCHOL. SCI. 397, 401 (2005) ("Black participants evaluated Whites with higher automatic-bias scores more positively than Whites with lower automatic-bias scores.").

In the succeeding sections I discuss two particular ways that a greater appreciation of second thoughts can inform legal theory and policy. First, I discuss how research on second thoughts reveals that antidiscrimination scholars have been too quick to read measures of unconscious bias as measures of discriminatory propensity and to reject Title VII's intentional discrimination prohibition as an effective check on unconscious bias. Second, I discuss how laws that regulate informational content without addressing the metacognitive experience of processing information may lead to perverse effects and how taking this metacognitive experience into account can lead to more effective policy.

A. The Law as Second Thoughts

Failing to distinguish between first and second thoughts can lead to misconceptions about the nature of intergroup bias, discrimination, and their regulation. Two particularly serious misunderstandings arise from the disregard of second thoughts: (1) because second thoughts often override bias in first thoughts, it is error to assume that an expression of first-order bias at one time or in one setting will generalize to other times and settings or that an expression of first-order bias holds any direct implications for behavior; (2) because the law can create metacognitive goals and validity tags that serve to check biased first thoughts, even without conscious awareness, it is error to assume that legal prohibitions on intentional discrimination will be ineffective against unconscious biases.

First, we cannot accurately determine any individual's tendency to discriminate from a measure of first-order, implicit bias. Given the intervening effects of second-order thoughts, psychological instruments designed to measure intergroup bias at the level of first-order thoughts, such as the IAT,[77] tell us nothing about the likelihood of bias occurring at the level of judgments, decisions, or behaviors. "[I]ndirect measures of attitudes cannot assess discrimination, because discrimination is overt behavior, not an attitude. Thus, even if a reliable and valid measure of implicit prejudice were developed, additional research would have to establish the link between these measured attitudes and behavior."[78] To date, no empirical research has established that any

77. *See supra* note 11 and accompanying text. It should be noted that it is not even clear if the IAT measures first-order associations between membership in demographic or social groups and evaluations or traits also associated with those groups. *See* Mitchell & Tetlock, *supra* note 6, at 1059-94.

78. C. Miguel Brendl et al., *How Do Indirect Measures of Evaluation Work? Evaluating the Inference of Prejudice in the Implicit Association Test*, 81 J. PERSONALITY & SOC. PSYCHOL. 760, 761 (2001). Gehring and colleagues echo this warning:

> One must be cautious . . . about claims that any measure provides a direct window into racially biased behavior. This caution is particularly warranted for the IAT: researchers in social psychology disagree about the meaning of IAT scores, yet much research is reported in academic journals and the popular press as if its validity as a measure of racial bias and prejudice were well-established.

particular score on the IAT reliably predicts any particular behavior in any particular setting.[79] In fact, one study found that *higher* implicit bias readings on the race IAT correlated with *better* interracial interactions.[80]

Further, because of the complex relationship between first- and second-order thoughts, measures of implicit bias should not be viewed as measures of durable intergroup preferences, much less discriminatory tendencies.[81] To put it another

---

William J. Gehring et al., *Thinking About Interracial Interactions*, 6 NATURE NEUROSCIENCE 1241, 1241 (2003).

79. In a re-analysis of data from two studies examining the relationship between scores on the race IAT to interracial judgments and behavior, we found no range of IAT scores for which we could reliably predict whether a White subject would express positive, negative, or neutral behavior toward a Black person. *See* Hart Blanton et al., Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the Race IAT (2008) (unpublished manuscript on file with the *McGeorge Law Review*); *see also* Hart Blanton & James Jaccard, *Arbitrary Metrics in Psychology*, 61 AM. PSYCHOLOGIST 27 (2006).

> [T]he meaning of different IAT scores must be established through research that links specific scores to the observable events that are relevant to the underlying psychological dimension of interest. In the case of the race IAT, this means that its metric becomes meaningful to the extent that one knows just how much "relative implicit preference for Whites versus Blacks" is implied by any given IAT score.

Id. at 33.

80. *See* Shelton et al., *supra* note 76, at 401 ("Black participants evaluated Whites with higher automatic-bias scores more positively than Whites with lower automatic-bias scores."). This finding is in line with other research showing, paradoxically, that awkwardness in interracial relations may be the result not of Whites' racism but Whites' concerns about being labeled racist. *See, e.g.*, Phillip Atiba Goff et al., *The Space Between Us: Stereotype Threat and Distance in Interracial Contexts*, 94 J. PERSONALITY & SOC. PSYCHOL. 91, 104 (2008) ("The four studies presented here provide support for the hypothesis that stereotype threat may cause Whites to distance themselves from Blacks. This distancing was unrelated to racial prejudices, either implicit or explicit."); Jacquie D. Vorauer & Cory A. Turpie, *Disruptive Effects of Vigilance on Dominant Group Members' Treatment of Outgroup Members: Choking Versus Shining Under Pressure*, 87 J. PERSONALITY & SOC. PSYCHOL. 384, 395 (2004) ("[Study Participants'] concerns regarding the impressions conveyed by their behavior prompted choking effects in individuals who were lower in prejudice or racial ingroup identification."). Vorauer and Tupie's analysis showed that discomfort evident in Whites' nonverbal behavior toward Blacks was the result of discordance between general racial attitudes at the implicit level and evaluations of a specific Black person. *See id.* Thus, metacognitive concerns about racial bias can lead to over-correction that may result in the very (mis)perceptions sought to be avoided. Social commentators who suggest that most White Americans suffer from hard-to-avoid unconscious biases may be unwittingly contributing to this over-correction. *See, e.g.*, GLADWELL, *supra* note 24, at 77-88 (describing IAT research).

> In all likelihood, you won't be aware that you're behaving any differently than you would around a white person. But chances are you'll lean forward a little less, turn away slightly from him or her, close your body a bit, be a bit less expressive, maintain less eye contact, stand a little farther away, smile a lot less, hesitate and stumble over your words a bit more, laugh at jokes a bit less. Does that matter? Of course it does. Suppose the conversation is a job interview. . . . What this unconscious first impression will do . . . is throw the interview hopelessly off course.

Id. at 85-86.

81. The IAT creators' use of the term "automatic preference" (e.g., "automatic preference for Whites") to refer to scores on the IAT has likely contributed to misunderstandings about the meaning of an IAT score, for this term implies more than mere association between concepts in the mind—it implies that the score reflects how one will relate to groups and invokes notions of choice-based preferences. As an answer to Frequently Asked Question No. 23 on the Project Implicit website indicates, however, what is really meant by "automatic preference" is just an "association between a concept and an evaluation such as good-bad, positive-negative, or pleasant-unpleasant." S*ee* Project Implicit, Frequently Asked Questions, https://implicit.harvard.edu/implicit/demo/background/faqs.html#faq6 (last visited June 4, 2008) (on file with the *McGeorge Law Review*).

way, because measures of first-order bias are extremely sensitive to individual and situational conditions, a measure of implicit biases accesses only a transient psychological state rather than a stable, cross-situational preference for one group over another:

> [T]hough [implicit measures] originally were assumed to be highly stable and resistant to change, considerable research now indicates that responses on these measures are highly context dependent. For example, priming and IAT results are sensitive to subtle contextual features of the stimuli, temporary changes in the accessibility of different features of the attitude object, variations in the context in which the measure is administered, fluctuations in respondents' physical and motivational states, and many other factors.[82]

Accordingly, suggestions that IAT scores be used as evidence of discriminatory propensities for purposes of litigation or personnel selection should be rejected.[83] These suggestions assume a stability in implicit bias, a reliable relation between implicit bias and behavior, and a reliability in the measurement of implicit bias that do not exist. The IAT has a median test-retest reliability coefficient of approximately 56,[84] which is poor from a psychometric perspective and which means that a person's score on the IAT at Time 1 is likely to be very different from the score at Time 2.

While it is understandable that legal scholars view estimates of widespread intergroup bias operating at the level of first thoughts as cause for alarm,[85]

---

82. Jeffrey W. Sherman et al., *The Self-Regulation of Automatic Associations and Behavioral Impulses*, 115 PSYCHOL. REV. 314, 321 (2008) (citations omitted).

83. *Contra* IAN AYRES, PERVASIVE PREJUDICE? UNCONVENTIONAL EVIDENCE OF RACE AND GENDER DISCRIMINATION 424-25 (2001) (suggesting that scores on the IAT might be used as evidence in litigation, employee screening, and racial sensitivity programs); Reshma M. Saujani, *"The Implicit Association Test": A Measure of Unconscious Racism in Legislative Decision-Making*, 8 MICH. J. RACE & L. 395, 413 (2003) ("The IAT (or psychological testing more generally) should be added to the non-exclusive list enumerated in Arlington Heights for determining racial intent." (emphasis omitted)). Such uses in litigation would likely run afoul of the character evidence rule. *See* FED. R. EVID. 404(a). Plaintiffs' experts in employment class action litigation have recently begun to use IAT research on implicit bias to support their opinions that company management likely engaged in systematic discrimination against women or minorities. *See* William T. Bielby, *Can I Get a Witness? Challenges of Using Expert Testimony on Cognitive Bias in Employment Discrimination Litigation*, 7 EMP. RTS. & EMP. POL'Y J. 377, 379-81 (2003).

84. *See* Brian A. Nosek et al., *The Implicit Association Test at Age 7: A Methodological and Conceptual Review*, *in* AUTOMATIC PROCESSES IN SOCIAL THINKING AND BEHAVIOR 265, 274 (John A. Bargh ed., 2007). This lack of measurement reliability should not be surprising given that multiple first thoughts may be primed by observing another person, even in the context of the IAT, and given the ability of second thoughts to override first thoughts even at the implicit level.

85. *E.g.*, Stephen P. Garvey, *Self-Defense and the Mistaken Racist*, 11 NEW CRIM. L. REV. 119, 166 (2008) ("[T]he prevailing wisdom among cognitive scientists is that more or less all of us are burdened with racist beliefs."); John A. Powell, *Structural Racism: Building upon the Insights of John Calmore*, 86 N.C. L. REV. 791, 799 (2008).

> In recent years, social science research has shown that we all have subconscious or implicit biases—beliefs, attitudes, and expectations that are based on stereotypes about the race, gender, age, or

particularly given implicit-bias researchers' failure to discuss clearly the gap between first thoughts and outputs in the forms of judgments, decisions, and behavior.[86] From an associative learning perspective, it would be surprising if minority groups and females were not more often associated with negative outcomes and related evaluative terms than White males in societies where there are racial, ethnic, and sexual inequalities.[87] But, as I have been arguing, to

---

> groups to which an individual belongs. Though most of us are completely unaware of their influence on our subconscious, these biases affect how we perceive, interpret, and understand others' actions.

Id.

86. *See* Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CAL. L. REV. 945, 946 (2006); Kang & Banaji, *supra* note 15. Drs. Greenwald and Banaji created the Implicit Association Test and co-authored these articles. The articles, which were aimed at popularizing IAT research within the legal academy, omit discussion of many of the published criticisms of, and acknowledged limitations on, the IAT research. For instance, neither article notes the low reliability of the IATs or that a significant change was made in 2003 to the rules for scoring IATs due to acknowledged measurement artifacts in the tests that cause older adults' scores to be artificially inflated in the direction of greater bias. *See* Anthony Greenwald et al., *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197, 212 (2003) ("The present findings call strongly for replacing the IAT's conventional scoring procedure.").

> [T]he improved algorithm offers a gain in construct purity. That is, the improved algorithm, compared with the conventional scoring procedure, is less contaminated by extraneous variables. One such contaminant is the conventional IAT measure's production of spuriously extreme IAT scores for slow responders . . . . The new algorithm almost completely eliminates this artifact . . . . Resistance to the response-speed artifact should be useful in studies that compare IAT scores for groups, such as children versus adults, that differ in speed of responding. The new algorithm likewise should provide more valid correlations of IAT measures with individual difference measures, such as age or working memory capacity, that correlate with response speed.

Id. at 214-15. Whether the new method for scoring the IAT appropriately controls for processing speed artifacts remains a point of debate. *See, e.g.*, Hart Blanton et al., *Decoding the Implicit Association Test: Implications for Criterion Prediction*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 192, 209-10 (2006). Thus, there is no warning to legal scholars that the results of IAT research based on pre-2003 scoring rules may be contaminated by known artifacts and that IAT scores are very unstable.

87. *See generally* DAVID R. SHANKS, THE PSYCHOLOGY OF ASSOCIATIVE LEARNING 20-37 (1995); Eva Walther et al., *Evaluative Conditioning in Social Psychology: Facts and Speculations*, 19 COGNITION & EMOTION 175 (2005); *see also* GLADWELL, *supra* note 24, at 85 ("'You don't choose to make positive associations with the dominant group . . . . But you are required to. All around you, that group is being paired with good things. You open the newspaper and you turn on the television, and you can't escape it.'" (quoting Mahzarin Banaji, one of the IAT's creators)). For evidence that the results of IATs are affected by available associative knowledge that may be irrelevant to one's personal evaluations of minorities or women, see H. Anna Han et al., *The Influence of Experimentally Created Extrapersonal Associations on the Implicit Association Test*, 42 J. EXPERIMENTAL SOC. PSYCHOL. 259, 269 (2006) ("The two experiments reported here suggest that the traditional IAT is influenced both by one's personal associations and by extrapersonal associations—ones that are attitude-irrelevant but that are valenced and available in memory." (citation omitted)); Michael A. Olson & Russell H. Fazio, *Reducing the Influence of Extrapersonal Associations on the Implicit Association Test: Personalizing the IAT*, 86 J. PERSONALITY & SOC. PSYCHOL. 653, 663 (2004) ("Data from the four experiments reported here suggest that the IAT has the potential to be contaminated by associations that although available in memory are irrelevant to one's evaluation of the attitude object.").

As discussed shortly, it is important to note that the strength and nature of these contingency judgments can be altered by our second thoughts and countervailing associations. *See, e.g.*, Jan De Houwer & Tom Beckers, *A Review of Recent Developments in Research and Theories on Human Contingency Learning*, 55B Q.J. EXPERIMENTAL PSYCHOL. 289, 306 (2002) ("[B]eliefs and retrospective recoding of events can have . . . a profound impact on contingency judgments."). Furthermore, to the extent that the IAT does measure

recognize a contingent relationship between group membership and material outcomes at the level of first thoughts does not compel the conclusion that these first thoughts bias our judgments, decisions, and behaviors. One may rationally associate the New York Yankees with winning and success, but that association does not compel one to root for the Yankees or root against the Red Sox.

Second, the fact that intergroup biases may be present in first thoughts without conscious awareness does not mean that regulations aimed at preventing intentional discrimination will be ineffective against these biases.[88] It is error to equate regulations that reward or punish certain goals (i.e., regulations directed at intentions) with regulations that require conscious mediation of thoughts and behavior;[89] accordingly, it is error to assume that prohibitions against intentional discrimination cannot combat unconscious or implicit bias.[90]

Because goals and norms can operate at both conscious and unconscious levels and override initial stereotypic and prejudiced reactions to women and minorities, one should not assume that implicit bias is immune to legal regulations that prohibit using protected-group memberships as a basis for personnel decisions. When discrimination occurs in a workplace governed by Title VII, the cause may well be the lack of effective priming of the nondiscrimination norm (or outright rejection of the norm) rather than an inability to check subtle or unconscious bias. By holding managers accountable for the neutrality of their personnel decisions and for the performance of the work groups they assemble, or by using diverse evaluation teams, employers can

---

associations within the mind as opposed to other phenomena, *see* Mitchell & Tetlock, *supra* note 6, at 1059-94, it is not a pure measure of associations and association strength. *See* Frederica R. Conrey et al., *Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance*, 89 J. PERSONALITY & SOC. PSYCHOL. 469, 483 (2005) ("[O]ur findings suggest that researchers should exercise caution in assuming that the implicit prejudice scores they calculate with priming measures or the IAT reflect exclusively the strength of automatic associations. Clearly, attempts to overcome these associations also contribute to performance on these tasks.").

88. The most prominent regulation against intentional disparate treatment is, of course, found within Title VII's direction not to make employment decisions on the basis of race, color, religion, sex, or national origins. *See* 42 U.S.C. § 2000e-2 (2006).

89. *See supra* notes 62-69 and accompanying text.

90. *See, e.g.*, LU-IN WANG, DISCRIMINATION BY DEFAULT: HOW RACISM BECOMES ROUTINE 135 (2006) ("[L]egal prohibitions against discrimination are inadequate to redress the largest share of modern discrimination, particularly under the dominant model of intentional discrimination."); Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CAL. L. REV. 1, 3 (2006) ("Unconscious bias, interacting with today's 'boundaryless workplace,' generates inequalities that our current antidiscrimination law is not well-equipped to solve."); Martha Chamallas, *Deepening the Legal Understanding of Bias: On Devaluation and Biased Prototypes*, 74 S. CAL. L. REV. 747, 753 (2001) ("[A]ntidiscrimination law is inadequate because it targets mainly intentional discrimination, missing the more prevalent contemporary forms of bias that are often nondeliberate or unconscious."); Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CAL. L. REV. 969, 980 (2006) ("The central focus of existing antidiscrimination law is on prohibiting consciously biased decisionmaking—a focus that has produced intense criticism from those interested in implicit bias.").

activate nondiscrimination and accuracy goals that operate at both conscious and unconscious levels.[91]

Laws aimed at good and bad intentions may debias our first thoughts in another important way. Conscious attention to the law's prohibitions may lead to the online monitoring of our behavior for bias, but conscious thoughts about the appropriateness or inappropriateness of certain considerations may lead to offline debiasing as well, through the creation of metacognitive validity tags. As the mind repeatedly pairs stereotypic or prejudicial thoughts with metacognitive thoughts about the illegality or immorality of such first thoughts, the mind tags the first thoughts as invalid, which may result in automation of the metacognitive process. Through this metacognitive process, the law becomes associated with (or tags) first-order thoughts and thereby weakens or prevents the negative effects that biased first-order thoughts might otherwise have.

Presently, antidiscrimination law's ability to create metacognitive validity tags is conjectural, given the early state of research into the metacognitive model of attitudes. But it would be surprising if our conscious thoughts and feelings about how we should act did not have some lasting or residual effects on associative networks and the metacognitive process overseeing these networks similar to that of the validity-tagging concept.[92] Viewing the law primarily as a system of second thoughts that functions both consciously and unconsciously opens up new perspectives on the cognitive representation and operation of the law, with implications well beyond the regulation of intergroup bias. For instance, an understanding of the ease or difficulty with which certain legal associations may arise, how their effects may generalize within spreading

---

91. *See, e.g.*, Susan T. Fiske, *Stereotypes Work . . . But Only Sometimes: Comment on How to Motivate the "Unfinished Mind,"* 3 PSYCHOL. INQUIRY 161, 162 (1992); Thomas E. Ford et al., *The Role of Accountability in Suppressing Managers' Preinterview Bias Against African-American Sales Job Applicants*, 24 J. PERS. SELLING & SALES MGMT. 113 (2004); Steven L. Neuberg & Susan T. Fiske, *Motivational Influences on Impression Formation: Outcome Dependency, Accuracy-Driven Attention, and Individuating Processes*, 53 J. PERSONALITY & SOC. PSYCHOL. 431 (1987); Janet B. Ruscher & Laura Lawson Duval, *Multiple Communicators with Unique Target Information Transmit Less Stereotypical Impressions*, 74 J. PERSONALITY & SOC. PSYCHOL. 329 (1998). *See generally* Lerner & Tetlock, *supra* note 34, at 341-42 (discussing the multiple psychological processes relevant to judgmental bias that are affected by accountability). Two interesting questions for workplace debiasing are (a) just how subtle the priming of the nondiscrimination norm can be and still be effective and (b) whether diversity training can instill chronic egalitarian goals or create metacognitive uncertainty about the accuracy and efficacy of group stereotypes that has debiasing effects. On this last point, Weary and colleagues found that the activation of temporary causal uncertainty beliefs (or doubts about one's understanding of causal relations in social interactions) led to reduced reliance on stereotypes. *See* Gifford Weary et al., *Chronic and Temporarily Activated Causal Uncertainty Beliefs and Stereotype Usage*, 81 J. PERSONALITY & SOC. PSYCHOL. 206, 216 (2001).

92. Most Americans do consciously endorse nondiscrimination norms, but according to IAT research, most Americans also show some degree of bias in their implicit associations. *See, e.g.*, Greenwald & Krieger, *supra* note 86, at 955 ("[T]he IAT measures consistently revealed greater bias in favor of the relatively advantaged group (averaging almost three-quarters of respondents across all the topics) than did the explicit measures (for which an average of slightly over one-third of respondents showed bias favoring advantaged groups)."); *see also* Lincoln Quillian, *New Approaches to Understanding Racial Prejudice and Discrimination*, 32 ANN. REV. SOC. 299, 309-12 (2006).

activation networks, and how difficult it may be to extinguish legal associations, may help to explain how governments can achieve such high levels of legal compliance despite the low risk of punishment for most offenses. And we may find that "intuitive" feelings of unease about certain behaviors represent not just moral reactions but the automatic operation of legal associations or validity tags as well. In short, the metacognitive approach to legal regulation moves away from seeing the law simply as changing the price of different behaviors—for the purpose of a rational analysis of the costs and benefits of different courses of action—and directs attention to how the law alters our processing of information.[93] In the next section, I illustrate how this broader view of the cognitive effects of law may inform our views about the importance of structure and format in legal transactions and legal instructions.

B. Legal Structure and Second Thoughts

Professor Leff's quotation that opens this paper, lamenting the degradation of the meaning and impact of seals in modern commercial life, reveals an important insight about the psychological utility of legal formalities.[94] Professor Leff recognized that the importance of a legal structure extends well beyond what the law requires or encourages in terms of the disclosure of informational content; legal structure also affects how information is disclosed and, consequently, how information is processed. It is this last metacognitive effect that Leff appreciated, recognizing that a feeling of legal formality may lead to a different attitude toward a transaction, causing some to treat the transaction with greater solemnity and care and perhaps causing some not to go forward with the transaction at all (given the sense of foreboding elicited by the trappings of the deal).

When we consider the metacognitive effects of removing formality from the law, particularly through plain language movements that seek to translate complex legal concepts into familiar terms,[95] we realize that efforts to simplify legal transactions may have perverse effects: removing legalese from an

---

93. An associative perspective on the law's effects on thought leads to very different assumptions about information processing than the computational perspective underlying most economic approaches to the law:

> "Associations, unlike symbols, are not added, subtracted, multiplied, and divided in order to generate new associations." In contrast, cognitive accounts assume that the brain calculates such things as the intervals and contingencies between events and "records the results in memory for later use in the computations that mediate decisions" on whether or not to behave in a particular way.

David R. Shanks, *Associationism and Cognition: Human Contingency Learning at 25*, 60 Q. J. EXPERIMENTAL PSYCHOL. 291, 297 (2007) (quoting C. R. Gallistel & John Gibbon, *Computational Versus Associative Models of Simple Conditioning*, 10 CURRENT DIRECTIONS PSYCHOL. SCI. 146, 146-47 (2001)).

94. *See supra* note 1 and accompanying text.

95. *See* David M. LaPrairie, Note, *Taking the "Plain Language" Movement Too Far: The Michigan Legislature's Unnecessary Application of the Plain Language Doctrine to Consumer Contracts*, 45 WAYNE L. REV. 1927, 1927 (2000) ("'Plain language' is generally defined as something 'written in a clear and coherent manner using words and phrases with common and everyday meanings.'" (quoting H.B. 4028 § 2(c), Reg. Sess. (Mich. 1997))).

insurance form or other contract document may increase feelings of comprehension due to greater ease of processing and familiarity with the terms without increasing actual comprehension of the legal significance and meaning of the plainer language. And there is some evidence to this effect. In an empirical study examining the effects of different simplifications to legal contracts, Masson and Waldron found that using more familiar terms and shorter sentences did make the contracts more readable and comprehensible, but overall comprehension of the legal meaning of the contracts remained low.[96] Therefore, after converting contracts to plainer language, people may continue to enter into contracts they do not truly understand—indeed, they may increase their contracting rates—because they literally feel good about the "clearer" contracts as they read them, when a feeling of dread or apprehension may be the more adaptive metacognitive experience in such situations.[97] This is not to say that moves to plainer language should be rejected, but that reformers need to measure whether the reforms are having the desired effects: do laypeople better understand the intended meaning after the redrafting, or do they just find the contract less intimidating? Less intimidation with less comprehension represents the worst of all worlds.[98]

---

96. *See* Michael E. J. Masson & Mary Anne Waldron, *Comprehension of Legal Contracts by Non-Experts: Effectiveness of Plain Language Redrafting*, 8 APPLIED COGNITIVE PSYCHOL. 67, 78-79 (1994). As Masson and Waldron say, "[h]owever much of law's inaccessible nature may be explained by obscurantism, not all of it melts away in the face of plain language." *Id*. at 79.

97. Likewise, making a technical term, such as a warranty disclaimer or remedy limitations, more conspicuous through larger, bolded font may make processing of the term easier (and thus less threatening), without any gains in comprehension. Of course, if the goal is primarily to give notice rather than increase comprehension, then this concern is misplaced. If the goal is also to instill cautionary thoughts, however, as used to be the case with seals, then making the font stand out but leaving it slightly fuzzy and difficult to read would likely have that metacognitive effect, without diminishing notice. *See* Schwarz et al., *supra* note 37, at 145 (noting that a difficult-to-read font can lead to low processing fluency, which "flags material for closer scrutiny").

Similar paradoxical effects may occur in public information campaigns if only the substance and not the form of the campaign is considered:

> From a metacognitive perspective, frequent exposure also facilitates increasingly fluent processing of the message and increased perceptions of familiarity, which, in turn, increase the likelihood of message acceptance. Rhyming slogans and presentation formats that facilitate fluent processing will further enhance this effect.
>
> . . . [T]his logic implies that false information is better left alone. Any attempt to explicitly discredit false information necessarily involves a repetition of the false information, which may contribute to its later familiarity and acceptance.

Id. at 146-47. An example of a successful public information campaign is the use of graphic warnings on cigarette packages, which were more noticeable, elicited more negative thoughts, and lead to greater awareness of health risks compared to text warnings. *See* David Hammond et al., *Text and Graphic Warnings on Cigarette Packages: Findings from the International Tobacco Control Four Country Study*, 32 AM. J. PREVENTIVE MED. 210, 215 (2007). However, once these ads become more familiar and easier to process, they risk losing their persuasive and informational effects. *See id*. Pictures of graphic tobacco warnings used in various countries can be viewed at: Picture Based Health Care Warnings, http://www.smoke-free.ca/warnings/countries%20and%20laws.htm (last visited June 9, 2008) (on file with the *McGeorge Law Review*).

98. This conclusion depends on how one calculates social welfare. Depending on one's conception of welfare, it may be appropriate and desirable to trick people into agreeing to certain contract terms. For instance,

One important effect of metacognitive doubt can be inaction, and legal transactions may be structured to take advantage of this metacognitive effect: when individuals become aware of a possible bias but are unsure how to correct it, they may stop going forward with a judgment or a behavior. In empirical studies of rationality and social judgment, non-responses are often discouraged or excluded from data analyses, yet in real life, withdrawing from a situation or withholding judgment may be a common, smart response. Yzerbyt and colleagues' "social judgeability" research finds that people may withhold judgment when they do not feel justified in rendering such judgment absent better or more information;[99] my research has found that individuals who experience doubt about the normative response to a problem will often refuse to provide an answer rather than venture a guess.[100]

Manipulating metacognitive experiences to create doubt about the proper approach may be an effective tool for sorting individuals who are well- and ill-suited for certain types of transactions, without affirmatively barring the transactions or erecting categorical exclusions that may be over-inclusive (i.e., not all members of a group that might in general be ill-suited for a transaction will, in fact, be ill-suited for it). Likewise, erecting accountability mechanisms that encourage metacognitive reflection (and possibly doubt) about the neutrality or legality of one's decisions may debias judgments in legally sensitive or risky situations, such as personnel selection settings, though the risk of overcorrection may arise as well. Setting the proper balance between under- and over-correction risks arising from metacognitive manipulations requires a value judgment about the relative social undesirability of the harms attached to under- and over-correction. This judgment is informed by empirical research of the likelihood of correction deficits or excesses occurring under different conditions.

One promising area for the use of second thoughts to advance legal goals lies in the domain of evidence law and its desire to have jurors disregard evidence that they have observed but that turns out to be inadmissible or admissible for only limited purposes. Experimental studies of the ability of jurors to honor an instruction to disregard evidence have shown mixed results, but overall, these

---

Sunstein and Thaler's "libertarian paternalism" emphasizes using default rules that, due to psychological biases, irrational persons are likely to accept to enhance their well-being. *See* Cass R. Sunstein & Richard H. Thaler, *Libertarian Paternalism Is Not an Oxymoron*, 70 U. CHI. L. REV. 1159, 1161-62 (2003).

99. *See, e.g.*, Yzerbyt et al., *Dilution of Stereotypes*, *supra* note 55, at 1315 ("According to social judgeability theory, social perceivers have a fair intuition of the kind of information that allows them to make sound judgments when a real individual is at stake."); *see also* Jean-Claude Croizet & Susan T. Fiske, *Moderation of Priming by Goals: Feeling Entitled to Judge Increases Judged Usability of Evaluative Primes*, 36 J. EXPERIMENTAL SOC. PSYCHOL. 155, 177 (2000) (discussing importance of the notion of the "usability" of information to the expression of social judgments).

100. I regularly administer surveys involving a wide range of judgment and decision-making problems in conjunction with my seminars on law and psychology, and I then use aggregate data from the surveys to illustrate the phenomena being discussed. Even in these anonymous surveys, large numbers of my students skip the more difficult problems without even venturing a guess, and many report doing so due to feelings of difficulty and uncertainty.

studies have found that admonitions to ignore the inadmissible evidence do not lead to results equal to those of jurors who never observed the inadmissible evidence.[101] However, one of the ways to make such admonitions more effective is to give jurors a reason why the evidence should not be relied on,[102] an approach that recognizes the importance of jurors' metacognitive processes. "Clearly, jurors respond to specific information they can understand and appreciate."[103] Also, one study recently found that a "Neutralization" judicial instruction (which explicitly warns jurors that the inadmissible evidence might bias their judgments improperly and thus encourages metacognitive self-correction) and an "Elaborate Forget" instruction (which very directly and repeatedly informs jurors to put the evidence out of their minds)[104] both lead to effective debiasing for very incriminating inadmissible evidence.[105] Mock jurors who observed the inadmissible evidence and received these instructions understood the potential probative value but were able to ignore it when rendering verdicts. Thus, judicial instructions that expressly take into account how jurors will process information, and that appeal to jurors' metacognitive self-correction processes, offer a promising solution to a persistent evidentiary problem.[106]

Another category of judicial instruction that might benefit from a metacognitive overhaul is that of burden-of-persuasion instructions. These instructions, particularly directions about the proper subjective weight that should accompany a finding of criminal guilt, have received harsh academic criticism.[107] Along with criticisms of the definitional inconsistencies and confusion surrounding these instructions, criticism flows from empirical studies demonstrating that verbal descriptions of the decision standard fail to elicit appropriately high decision standards as applied by jurors.[108] Indeed, studies have

---

101. Nancy Steblay et al., *The Impact on Juror Verdicts of Judicial Instruction to Disregard Inadmissible Evidence: A Meta-Analysis*, 30 LAW & HUM. BEHAV. 469, 477-78 (2006).

102. *Id*. at 486.

103. *Id*. Interestingly, telling mock jurors that evidence is inadmissible because it was illegally obtained can lead to even greater reliance on the evidence, also indicating that jurors are exercising their own judgments about the propriety of considering the evidence but reaching conclusions opposite to those that the law desires in the face of this explanation. *See id*.

104. *See* Linda J. Demaine, *In Search of an Anti-Elephant: Confronting the Human Inability to Forget Inadmissible Evidence*, GEO. MASON L. REV. (forthcoming) (manuscript on file with the *McGeorge Law Review*) (providing verbatim descriptions of such jury instructions).

105. *Id*. The neutralization instruction did perform better than the elaborate forget instruction on some measures within the study, leading the author to recommend the former over the latter for use by judges. *See id*.

106. *See* Steblay et al., *supra* note 101, at 470.

107. *See, e.g.*, Larry Laudan, *Is Reasonable Doubt Reasonable?*, 9 LEGAL THEORY 295 (2003); Lawrence M. Solan, *Refocusing the Burden of Proof in Criminal Cases: Some Doubt About Reasonable Doubt*, 78 TEX. L. REV. 105 (1999).

108. *See* Erik Lillquist, *Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability*, 36 U.C. DAVIS L. REV. 85, 88 (2002).

> Ever since the publication of Harry Kalven and Hans Zeisel's *The American Jury* in 1966, researchers have been subjecting reasonable doubt, as well as other standards of proof, to empirical

found that mock jurors often do not distinguish between the clear-and-convincing-evidence standard and the reasonable-doubt standards, and that they even occasionally impose a higher standard under the former than the latter.[109] Much has been written on how to revise jury instructions to elicit appropriate decision thresholds, and it is certainly possible to revise verbal descriptions to elicit more or less care on the part of jurors.[110] The most significant reform proposal from a metacognitive standpoint, however, is to switch descriptions of burdens from verbal to numerical formats.[111]

Advocates of quantitative descriptions of persuasion burdens favor them over verbal descriptions because the quantitative descriptions elicit greater consistency and decision thresholds, which is more in line with the theoretical understandings of how the preponderance, clear-and-convincing, and reasonable-doubt standards should operate.[112] I posit the use of quantitative burdens, however, not because they might produce verdicts closer to some normative ideal, but because framing the jurors' task in quantitative terms may activate a more deliberate, rational evaluation of the evidence. There is some evidence that asking jurors to express judgments as verbal statements of probability activates an intuitive assessment mode, while asking them to express judgments as numerical statements of probability activates a deliberative assessment mode.[113] Numerical formats make accuracy and precision considerations more salient, evoking a greater feeling of a need for proof of a result instead of a mere statement of opinion. Verbal formats,

---

> tests. This research has consistently shown that the jurors in criminal cases will often be satisfied with much less certainty than is conventionally assumed.

Id.

109. *See* Dorothy K. Kagehiro, *Defining the Standard of Proof in Jury Instructions*, 1 PSYCHOL. SCI. 194, 197 (1990); Dorothy K. Kagehiro & W. Clark Stanton, *Legal vs. Quantified Definitions of Standards of Proof*, 9 L. & HUM. BEHAV. 159, 173 (1985).

110. *See generally* Nancy S. Marder, *Bringing Jury Instructions into the Twenty-First Century*, 81 NOTRE DAME L. REV. 449 (2006); Elisabeth Stoffelmayr & Shari Seidman Diamond, *The Conflict Between Precision and Flexibility in Explaining "Beyond a Reasonable Doubt,"* 6 PSYCHOL. PUB. POL'Y & L. 769 (2000).

111. *See, e.g.*, Kagehiro, *supra* note 109, at 198; Peter Tillers & Jonathan Gottfried, *Case Comment—* United States v. Copeland*, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim that Proof Beyond a Reasonable Doubt Is Unquantifiable?*, 5 LAW, PROBABILITY & RISK 135, 155-56 (2006).

112. *See, e.g.*, Kagehiro, *supra* note 109, at 197.

113. Paul D. Windschitl & Gary L. Wells, *Measuring Psychological Uncertainty: Verbal Versus Numeric Methods*, 2 J. EXPERIMENTAL PSYCHOL.: APPLIED 343, 358 (1996); *see also* Amnon Rapoport et al., *Revision of Opinion with Verbally and Numerically Expressed Uncertainties*, 74 ACTA PSYCHOLOGICA 61 (1990). *But see* Ido Erev & Brent L. Cohen, *Verbal Versus Numerical Probabilities: Efficiency, Biases, and the Preference Paradox*, 45 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 1, 15-16 (1990) (finding no difference in performance under verbal and numerical response formats). Furthermore, for some choice problems, the processing activated by numerical formats may actually exacerbate errors relative to verbal formats. *See* Claudia González-Vallejo & Thomas S. Wallsten, *Effects of Probability Mode on Preference Reversal*, 18 J. EXPERIMENTAL PSYCHOL.: LEARNING MEMORY & COGNITION 855 (1992) (finding greater preference reversal with numeric than verbal formats due to the latter format eliciting less risk aversion). Thus, one must be cognizant that more elaborate processing may not be a sufficient condition to avoidance of a bias and may even exacerbate some biases under some conditions. The key is to fit the appropriate presentation format to the preferred processing mode for the particular task at hand.

on the other hand, provide comfort in their imprecision, and thus are likely to evoke less effortful processing: if one is asked to state how probable three "Heads" in a row are, a safe and easy response is "unlikely" or "improbable" because one's answer provides no specific point estimate for comparison to the normative answer, while a numerical response is either precisely right or wrong.[114] Also, because communicating a vote quantitatively is less natural and more difficult than expressing a vote in a verbal format, a juror who must communicate his or her vote numerically is likely to experience greater metacognitive discomfort during deliberations. In sum, the juror is more likely to engage in greater monitoring of his or her information processing.[115]

The format in which information is presented will have effects far beyond the specific examples discussed here.[116] As with all legal uses of empirical research, whether a particular different effect associated with one or another presentation format is desirable depends on the policy goal and the trade-offs presented by moving from one format to another.[117] Nevertheless, we should recognize that

---

114. And many opinions cannot be stated with great precision, leading to a preference for verbal statements of probability. *See* Thomas S. Wallsten et al., *Measuring the Vague Meanings of Probability Terms*, 115 J. EXPERIMENTAL PSYCHOL.: GEN. 348, 348 (1986).

115. *See* Thomas S. Wallsten et al., *Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numerical Terms*, 31 BULL. PSYCHONOMIC SOC'Y 135, 137 (1993); Tzur M. Karelitz & David V. Budescu, *You Say "Probable" and I Say "Likely": Improving Interpersonal Communication with Verbal Probability Phrases*, 10 J. EXPERIMENTAL PSYCHOL. APPLIED 25, 26 (2004). While most people prefer to communicate information using verbal probability phrases, they prefer to receive information in numerical form. *See* Karelitz & Budescu, *supra*, at 26. Erev and Cohen, who observed this same pattern, labeled it the "communication mode preference (CMP) paradox." Erev & Cohen, *supra* note 113, at 2.

116. *See, e.g.*, James R. Bettman & Pradeep Kakkar, *Effects of Information Presentation Format on Consumer Information Acquisition Strategies*, 3 J. CONSUMER RES. 233, 239 (1977) (discussing and illustrating how consumers' information acquisition strategies are affected by the format in which product information is presented); Eric R. Stone et al., *Foreground: Background Salience: Explaining the Effects of Graphical Displays on Risk Avoidance*, 90 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 19, 20-23 (2003) (discussing and illustrating how the graphical displays of risk information can be more effective in encouraging risk avoidance than numerical displays); Yung-Cheng Shen & Chih-Wei Hue, *The Role of Information Presentation Formats in Belief Updating*, 42 INT'L J. PSYCHOL. 189, 191-92 (2007) (discussing and illustrating the different modes of processing, and their effects on belief updating, associated with numerical and verbal displays of consumer product information). For a good survey of research on the effects of presentation formats, with an emphasis on the communication of risk, see Isaac M. Lipkus, *Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations*, 27 MED. DECISION MAKING 696 (2007).

Within each format, there are many ways in which information can be displayed, with these sub-formats producing important differences. For instance, numeric descriptions might be displayed via single-event probabilities or frequency formats with very different effects. *See, e.g.*, Paul Slovic et al., *Violence Risk Assessment and Risk Communication: The Effects of Using Actual Cases, Providing Instruction, and Employing Probability Versus Frequency Formats*, 24 LAW & HUM. BEHAV. 271, 289-90 (2000) (finding that use of a frequency format response scale by forensic psychologists and psychiatrists led to lower estimates of likelihood of harm but higher perceived risk relative to a probability scale); *see also* Mitchell, *Taking Behavioralism Too Seriously?*, *supra* note 16, at 1989-92.

117. For discussions of policy arguments against a move to quantitative burden-of-persuasion instructions, see James Franklin, *Case Comment*—United States v. Copeland*, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): Quantification of the 'Proof Beyond a Reasonable Doubt' Standard*, 5 LAW, PROBABILITY & RISK 159, 159-65 (2006); Kagehiro, *supra* note 109, at 198; Tillers & Gottfried, *supra* note 111, at 141-56.

legal structure is unlikely to be neutral in its effects, for the manner in which information is presented affects the use and processing of that information.

III. CONCLUSION

A large and growing body of empirical research establishes that, while our intuitive and automatic first thoughts may at times lead us toward irrational and discriminatory behaviors, our second thoughts impose considerable control over these first thoughts. This research indicates that, while people are certainly not perfectly rational or law-abiding, they possess considerable potential for self-correction, including the abilities to suppress stereotypic thoughts, respond unconsciously to egalitarian goals, and infer from feelings of uncertainty that judgments or decisions should be issued with care. Accordingly, the reader should be wary of legal theories that assume that unintentional, quasi-rational actors are at the mercy of their first thoughts.

This Article has emphasized the importance of second thoughts for legal transactions that might be affected by irrational proclivities and for the law's regulation of intergroup relations that might be affected by unconscious intergroup biases. But by suggesting the range of ways to activate second thoughts and exercise control over judgments, decisions, and behavior, I hope to encourage others to consider how the law shapes our second thoughts, or fails to do so when perhaps it should. My most provocative claim is that the law functions primarily through second thoughts, via metacognitive amendments to our existing attitudes and beliefs. This testable hypothesis, if correct, may open up a new perspective on how the law functions. Rather than conceiving the law as primarily establishing prices that can be bargained around or added to rational calculations at will, or as a system primarily of online thoughts about proscribed and prescribed behavior that can easily be ignored or forgotten absent heavy-handed monitoring, it may also be appropriate to see the law as functioning automatically and pervasively, creating unconscious associations that cannot easily be ignored, but that can be altered or extinguished over time in line with principles of conditioning. A role for deliberative thought and rule-based calculation remains important in this perspective, for conscious deliberation is an important determinant of behavior, and the law will often be a conscious input to calculations. But we should also recognize the effects of the law on unconscious and fringe-conscious thoughts, with the law effectively serving as a brooding omnipresence in our heads if not in the sky.