Banaji, Mahzarin R. & Bhaskar, R. (2000). Implicit Stereotypes and Memory: The Bounded Rationality of Social Beliefs. In Schacter & Scarry (Eds.), Memory, Brain, and Belief (pp. 139-175). Cambridge, Massachusetts: Harvard University Press.

Implicit Stereotypes and Memory: The Bounded Rationality of Social Beliefs

5

Mahzarin R. Banaji
R. Bhaskar

The constructs of *belief* and *memory* have become closely intertwined in contemporary research on stereotyping. Yet even a few decades ago psychology's treatment of belief (stereotype) and memory moved along trajectories so disconnected that a link between the two might have seemed implausible. The theme of this book prompts a glance at the consequences of the relationship between belief and memory in experimental social psychology, the field that laid claim to a rigorous understanding of how beliefs about social groups shape judgments of their members.

Stereotyping and Prejudice

In the connection that developed between the social psychological study of belief and the cognitive study of memory, the concept of stereotype came to be demystified in two fundamental ways. First, the link between belief and memory allowed stereotyping and its uglier cousin, prejudice, to be viewed as ordinary in origin and pervasive in the sweep of its contagion.[1] Instead of considerations of the inherently evil nature of humans that produced prejudice and resulting group conflict, social psychology came to understand these phenomena as rooted in the very nature of how knowledge is acquired and used—that is, in the ordinary, constrained operations of an information processing system. Second and more recently, discoveries of the implicit aspects of memory have exposed the substantial unconscious component of social beliefs as well. In contrast to the first hundred years of research, which conveyed a view of memory and beliefs as operating exclusively in a conscious mode, the past two decades have shown increasingly that both memory and belief also operate implicitly in powerful yet unconscious ways, outside the actor's awareness or control (Greenwald and Banaji, 1995; Schacter, 1987).

This chapter draws on both developments, exploring the premise that the *ordinary* and *implicit* character of stereotypes and prejudice provides a more accurate representation of their nature, and reveals their full influence in human affairs. Together the properties of ordinariness and implicitness raise questions about how the limits on human thought and preferences diminish the rationality of stereotyped beliefs and prejudicial judgments. Conventionally, boundedly rational behaviors have been viewed as cognitive curiosities that influence human thought. Our purpose however, is to emphasize that cognitive acts are social acts and inherently have moral dimensions. This simple extension from ordinary and implicit social judgments to their moral consequences changes the scope of social cognition's view of stereotyping and prejudice in deep and permanent ways.

The Ordinary Nature of Stereotyping and Prejudice

The classic tract *The Authoritarian Personality* (Adorno, Frenkel-Brunswik, Levinson, and Sanford, 1950) provides the clearest statement of the intuitive position that stereotypes and prejudice are rooted in the structure of the prejudiced personality of special individuals. As averred in the preface to the book, "The central theme of the work is a relatively new concept—the rise of an 'anthropological' species we call the authoritarian type of man. In contrast to the bigot of the older style he seems to combine the ideas and skills which are typical of a highly industrialized society with irrational or anti-rational beliefs" (p. xi). The mission of *The Authoritarian Personality* was to identify and understand "the *potentially fascistic* individual, one whose structure is such as to render him particularly susceptible to antidemocratic propaganda" (p. 1; emphasis in original). This was a remarkable book providing a complex theory that merged psychoanalytic thinking with social science findings, and it encouraged a great deal of research. Although its central message was never formally challenged and is still endorsed in some quarters, it failed to have permanent theoretical impact. It was another volume, containing ideas quite ahead of their time, that stood waiting for academic psychology to catch up. In a beloved and now widely cited book, *The Nature of Prejudice,* Gordon Allport offered a radically different view of prejudice as reflecting "[man's] normal and natural tendency to form generalizations, concepts, categories, whose content represents an oversimplification of his world of experience" (1954, p. 27).

In the decades since Allport's book, and especially since the emphasis on social cognition in the mid-1970s, a steady transformation has occurred in thinking about stereotypes and prejudice, removing from them associations of unnaturalness and uniqueness. At least in their academic discourse, social psychologists have moved from the view that stereotypes and prejudice reflect the warped beliefs and preferences of distasteful individuals who threaten harmonious social existence, to the view that such processes are best considered the unhappy and even tragic outcomes of the ordinary workings of human cognition. As a consequence, a fundamental interconnectedness between the cognitive processes of memory, perception, attention, categorization, and reasoning on the one hand and the social processes of stereotyping and prejudice on the other became permanently established (see Banaji and Greenwald, 1994; Hamilton, 1981).[2]

Research has borne out Allport's (1954) claim that "categories have a close and immediate tie with what we see, how we judge, and what we do. In fact, their whole purpose seems to be to facilitate perception and conduct—in other words, to make our adjustment to life speedy, smooth, and consistent. This principle holds even though we often make mistakes in fitting events to categories and thus get ourselves in trouble" (p. 21). In this chapter we agree with Allport's prescient view that stereotypes and prejudice are rooted in the ordinary mechanisms of perception and categorization, and our data will provide evidence in support. We treat these mistakes as not only "getting ourselves in trouble," but as quite systematically getting others in trouble as well. Fortunately, we have access to an additional forty-five years of empirical research and national debate about intergroup relations, and these are the backdrop against which we discuss the implications of ordinary prejudices for imagining the ideal of a just society.

The Unconscious Nature of Stereotyping and Prejudice

A second shift in thinking has allowed yet another link between stereotypes and memory to emerge; this one stems from the inclusion of unconscious processes. Abelson's (1986) comment that "beliefs are like possessions" captured the idea that psychological beliefs are capable of evoking the feelings of ownership, endowment, and attachment typically elicited by material possessions. As if referring to tangible entities, we speak of acquiring, inheriting, and adopting beliefs, or of losing, disowning, and abandoning them. Such metaphors reveal in our language a view of beliefs as entities that are available to conscious awareness and responsive to conscious control (for instance, "I used to be an atheist, but I gave it up for Lent").[3] Such an understanding of beliefs as residing largely within conscious awareness and control is not only intuitive, but its widespread acceptance within the scientific community has determined methods of research used through the first century of experimental social psychology (see Banaji and Greenwald, 1994; Greenwald and Banaji, 1995).

Without debating the assumption that beliefs are consciously available and deployed, we worked with the assumption that the opposite of well-worn truths may also be true (cf. McGuire, 1973). For the past several years, one of us (MRB) has been interested in the discovery of the unconscious or implicit operation of beliefs and preferences.[4] This goal has involved creating in research participants the effortless articulation of stereotypes and prejudice, in some instances without their conscious awareness of such use or without conscious control over their expression. The research is part of a wide range of experiments designed to show the many ways in which stereotypes can be automatically activated and utilized even by actors who may intend a quite different response (Banaji, 1997; Chen and Bargh, 1997; Devine, 1989; Fazio, Jackson, Dunton, and Williams, 1995; Greenwald and Banaji, 1995; see Fiske, 1998, for a review).

The main message of this new body of research is the inevitability *of unconscious stereotyping and prejudice. The best of intentions do* not and cannot override the unfolding of unconscious processes, for the triggers of automatic thought, feeling, and behavior live and breathe outside conscious awareness and control. Again, Allport (1954) predicted the empirical findings decades in advance and in surprisingly modern language: "A person with dark brown skin will activate whatever concept of Negro is dominant in our mind. If the dominant category is one composed of negative attitudes and beliefs we will automatically avoid him, or adopt whichever habit of rejection is most available to us" (p. 21).

In the remainder of this chapter we champion two views, one directly reflecting the state of the science, the other a speculation about what the discoveries imply for social justice. Our position is that all humans are implicated to varying degrees in the operation of implicit stereotypes and prejudice. The pervasiveness of such expressions has been underestimated because large portions occur outside the awareness and control of both perceivers and targets. Based on evidence of the ways in which perception, attention, categorization, and memory operate to produce biases in judgment, stereotyping and prejudice too must be viewed as the outcome of ordinary and automatic thinking and feeling (Allport, 1954; Banaji and Greenwald, 1994; see Hamilton, 1981, for many chapters that include this assumption).

Such a position is allied more generally with psychological research that has offered a humbling view of human thought and rationality (Nisbett and Ross, 1980; Simon, 1983; Tversky and Kahneman, 1974). In contrast to a traditional nineteenth-century assumption that behavior is rational, the established modern view is that humans are boundedly rational, having "neither the facts nor the consistent structure of values nor the reasoning power at their disposal" to make decisions in line with subjective expected utility, "a beautiful object deserving a prominent place in Plato's heaven of ideas, . . . [but] impossible to employ . . . in any literal way in making actual human decisions" (Simon, 1983). Over the past few decades, such a depiction of decisionmaking has forced an acceptance of bounded rationality as the proper characterization of stereotyping and prejudice.[5]

Second, when stereotypes are unconsciously activated and used, two direct challenges to the implementation of fairness are posed: (a) perceivers and targets are often unaware of the steady and continuous rendering of judgments, and (b) judgments are based on beliefs about targets' social groups rather than on targets' actions. Such issues concerning fairness are not inventions of twentieth-century concepts of justice. Resolutions of the concerns they raise were implemented in relatively ancient systems. When judgments about humans are made, it is a fundamental principle of justice, now almost a thousand years old in Anglo-American jurisprudence (Assize of Clarendon, 1166), that targets of judgment should be made aware of the judgment. The main modern purpose of this rule, is, of course, to ensure that judgments are not based on factual error—although a deeper principle is also involved, that justice is better served when an opportunity to be heard exists (Ptahotep scrolls, 2400 B.C.). An unaware judge subverts this principle, because targets of judgment are denied the opportunity to contest, contradict, or modify the judgment.

Further, it is an equally hoary and fundamental principle of justice that judgments about individuals must be based on the individuals' own behavior, involving specific acts of commission and omission. Societies in which punishment was based on association (as when families of traitors were beheaded in seventh-century T'ang China) are regarded as barbaric by the standards of contemporary democracies. In this century, social science research in which beliefs about groups have been shown to influence judgments of individuals has been increasingly interpreted as representing bias. This interpretation arises not from a concern with the correctness of perceivers' beliefs about the group, but because the application of group-level knowledge (some X are Y) to individuals (X is Y) is deemed incorrect.[6]

The purpose of presenting the experimental evidence that follows is largely to help deconstruct an opposing view of stereotyping as correct and rational. This perspective is perhaps expressed most infelicitously by McCauley, Jussim, and Lee (1995), who endorse with evident approval the decision of a taxi driver not to stop for "a lone Black male at midnight in a bad neighborhood" (pp. 300–301). Such a view is so commonly held in contemporary American culture that the legitimacy of group-based judgments brings us up against a troubling tension between what is so socially prevalent as to be self-evident and what is correct upon sustained reflection. We will argue in the concluding section of this chapter that stereotypic behaviors of the sort endorsed by McCauley and colleagues may well be termed reasonable in the narrow sense that "reasonable" people in the same circumstances would exhibit the same behavior (Hart, 1976), which is nevertheless (a) inconsistent with logic, practice, intention, and assessment of similar outcomes in other domains, and (b) irrational in the classical, axiomatic sense of subjective expected utility theory (Arrow, 1963; Savage, 1972).

Overview of Experiments

Memory is regarded as the petri dish in which one can view the movements of particular forms of stereotypes and prejudice. Such a use of memory to study stereotypes is not new (see Hamilton, 1981), but research on implicit memory has allowed the even more prominent use of memory to examine stereotypes (see Banaji and Greenwald, 1994; Greenwald and Banaji, 1995). Implicit memory measures are those that do not require conscious recollection of a prior episode (Schacter, 1987). In line with such a definition, Greenwald and Banaji (1995) defined implicit stereotypes to be "the introspectively unidentified (or inaccurately identified) traces of past experience that mediate attributions of qualities to members of a social category" (p. 15).

Of the many potential connections between belief and memory, our focus has been on *beliefs about social groups as revealed through implicit or automatic memory*. Beliefs that serve as the building blocks of this research program concern characteristics of social groups that are widely endorsed, at least in the United States at the turn of the twenty-first century. Such beliefs may or may not be accurate descriptions of social groups, and while the limit on their accuracy is itself of interest, that is not an issue of primary concern here. For the purpose of this research, we work with beliefs that are simply endorsed with wide consensus, and we track their unwitting use in the judgment of individual members of social groups. Examples of the type of beliefs about social groups whose effects we might observe are that men are more likely to be famous than women; that the elderly are less alert mentally than younger adults; that women are emotional, nurturing, and submissive, but men are aggressive, competent, and strong; that the poor are more likely to be black than white; that women have less talent for leadership than men; that East Asians are intelligent but passive; and that gay men are feminine.

In one line of research, we examined how a subset of such beliefs about gender and race are expressed in judgments of individual members of these groups. Experimentally, individuals were created to be equally associated or dissociated with characteristics such as fame, criminality, aggression, or dependence, and situations were designed to measure the extent to which beliefs about the social groups of these individuals were unconsciously used in judging them. By creating conditions under which conscious awareness or control was reduced, we found that implicit memory reveals the implicit operation of stereotypes and prejudice.

In these experiments we have also discovered that implicit expressions of beliefs and attitudes are unrelated to explicit versions of the same beliefs and attitudes. Here college students are the theoretically appropriate population for study, for we are keenly interested in individuals known to consciously hold egalitarian beliefs and nonprejudicial attitudes. Showing the operation of implicit stereotypes and prejudices in individuals who consciously disavow their presence captures a dissociation in which we are interested, that between conscious and unconscious social judgment. Our interest is driven by a relatively straightforward concern. Dissociations between implicit and explicit beliefs are fundamentally important in understanding their nature, the relationship of each to the other, and the consequences of each. Further, it is necessary to acknowledge that there are indeed consequences of implicit and unintended expressions of beliefs because of their power to reward and punish on the basis of group membership.

Verifications of beliefs about social groups are so pervasive, frequent, and fleeting as to be quite unnoticed in the course of everyday life. A glimpse of John revving a tow truck, Jane walking with a brood of kindergartners, Tyrese slam-dunking a basketball, Mary Cheng playing a violin—these actions are ordinary enough that they may not evoke conscious contemplation of their meaning or cause. But they are automatically added to the cumulative mental record of social experience, with each episode strengthening a particular association—that between the psychological attributes signified by the act and the social group to which the actor belongs. Such exposures strengthen the belief that males are strong, that females are nurturant, that young black men are athletic, that Asian Americans have musical talent.

The consequences of such learning interest us here, for it is the well-learned and automatically activated associations between psychological qualities and social groups that can short-circuit the consciously espoused goal of assuring mental due process in social judgment, that is, that a person be judged not by color of skin, but by content of character. If this is indeed an espoused goal, then the findings from recent experiments in social psychology suggest that in social interaction such a goal of fair and equal treatment is largely ephemeral (see Bargh, Chen, and Burrows, 1996). A substantial and growing literature shows that exposure to specific behaviors automatically leads to inferences about the abstract psychological qualities that underlie the behaviors (Uleman, 1987), and that generalization from observation or experience with single individuals to other members of the group may be swift and unconscious (Henderson-King and Nisbett, 1996; Lewicki, 1986). Further, such beliefs can actually produce behavior consistent with the activated belief, demonstrating that implicit stereotypes have self-fulfilling consequences (Chen and Bargh, 1997). Beliefs about qualities such as intelligence and ability, toughness and softness, and the capacity for good and harm are implicitly and lawfully applied to individual members of groups, whether consciously endorsed or not, and whether consciously deemed morally appropriate or not.

In some experiments we have analyzed the computational circumstances that give rise to implicit stereotyping and prejudice. That the speed of mental computation can provide insight into social processes is not a new idea, although explicit appreciation of its value is not easily available in discussions of social cognition (Banaji, 1995). We have examined the automatic activation of beliefs and attitudes about social groups by measuring the time required to produce them under conditions of varying cognitive constraints. We use the time it takes to produce a decision about a person (in the presence of associated or dissociated beliefs or evaluative information) as a measure of the strength of the stereotype or attitude (cf. Fazio, Sanbonmatsu, Powell, and Kardes, 1986). For instance, gender-stereotypic words such as "mechanic" or "secretary" are presented for approximately 250 milliseconds, then replaced by a male or female name (John, Jane). Participants must judge whether the second item is a male or female name. The speed of judging male and female names in the presence of stereotypic items can reveal the strength of such stereotypes and provide an individual difference measure of the strength of such beliefs. By varying the constraints under which the judgments are made, we can speak more directly about the mechanisms that limit the application of conscious intention and deliberate strategy. Such constraints serve theoretically to vary the bounds on the rationality of participants' output.

To describe the subterranean connections between belief and memory, we proceed by summarizing the research procedures and major findings obtained in a single laboratory, with some experiments demonstrating the effects of implicit beliefs without conscious *awareness,* and others showing evidence of their operating outside conscious *control.* Together they serve as the empirical basis of our claims about the belief-memory relationship, and allow consideration of broader questions regarding the moral implications of implicit beliefs and preferences.

The Experiments

Selective Application of Activated Beliefs

Some of our research revealed how exposure to behaviors that automatically activate beliefs about social groups can implicitly and selectively influence judgment. In particular, we examined the consequences for targets who were judged during a state of temporary activation of stereotypes in the perceiver.

In one series of studies, we activated abstract knowledge about beliefs associated with women and men, such as *dependence* and *aggressiveness,* by presenting sentences that captured relevant behaviors which participants had to unscramble (T never goes alone; P kicked the dog). In a later session, one that participants believed to be unrelated to the prior task, we obtained the judgments of two individuals, Donna and Donald, who performed identical actions (Banaji, Hardin, and Rothman, 1993). After exposure to behaviors about dependence (T never goes alone), a female but not a male target was expected to suffer from the perception of greater dependence, while after exposure to behaviors about aggression (P kicked the dog), a male but not a female target was expected to be punished because of the perception of greater aggressiveness. Even more strongly than predicted, results revealed that the temporary activation of a belief did not influence person judgment when the target did not belong to the stereotyped group—say, when a male target was judged after exposure to dependence-related information and when a female target was judged after exposure to aggression-related information. Targets were judged more harshly only when the activated stereotyped belief and the targets' group membership were stereotypically matched; that is, when a female target was judged after exposure to behaviors depicting dependence, and a male target was judged after exposure to behaviors depicting aggression. In other words, activated beliefs about social groups are differentially sticky with regard to whom they are applied to in social judgments.

Temporary exposure to behaviors that activate a belief associated with social groups appears to shift the judgment of targets who merely belong to the social group associated with the activated belief. Interestingly, these data reveal both the ease with which shifts in judgment are possible and their limits. Belief activation appears to be necessary for producing stereotyped judgments, for no change from baseline was observed in the no-activation condition. In other words, men and women were not judged to be differentially aggressive or dependent when a belief was not activated prior to judgment.[7]

The conclusion about the relationship between belief and memory is obvious: judgments of individuals are shifted in a more extreme direction when they follow activation of a belief about the target's social group. Participants are not aware of such shifts and would perhaps even deny that such influences on their judgment are possible. From other experiments we know that individuals who employ stereotypes implicitly have no prior intention of doing so, and would consider such use to be unjust. Yet such ordinary and implicit influences occur and threaten the goal of fair and equal interpersonal treatment (Chen and Bargh, 1997).

Unequal Standards for Judgment

We tested whether the established link between gender and fame (for instance, that men as a group are more likely to be considered famous than women) would increase the assignment of (false) fame to nonfamous men when compared with equally nonfamous women. Participants in the research were exposed to a list of names, famous and nonfamous, male and female. Later they were presented with the same names in addition to new (previously unseen) names with similar characteristics. The task was to identify whether each name was the name of a famous person or not. Jacoby, Kelley, Brown, and Jasechko (1989) predicted and found that when faced with this task, people are poised to make a particular error detected in the form of false alarms on nonfamous names. Unable to separate varying sources of familiarity for a name (that is, familiarity from recent exposure versus familiarity from fame), participants were twice as likely to incorrectly judge a familiar (nonfamous) name "famous" than an unfamiliar (nonfamous) name.

The rationale for our experiments was that a belief about greater male fame ought to predispose participants toward greater incorrect identification of male than female names when implicit memory for nonfamous names still lingered as a function of prior exposure. Indeed, in four experiments we found a greater propensity for false alarms (identification of nonfamous names as famous) for familiarized male than female names, with no difference found on unfamiliarized names (Banaji and Greenwald, 1995). Using signal detection statistics, we found that this effect was located in the component of bias (β) and revealed itself in a more lenient criterion for judging male than female fame. In other words, an accurate belief about the differential fame of two social groups translated into differential standards for judging individuals equally (un)deserving of fame. Here ordinary conditions of familiarity were sufficient to justify greater assignment of fame to men than to women. Again, the ordinary and implicit nature of social judgment is the basis of a threat to fair and equal social treatment.

False Memories Created by Race Beliefs

We have also used a variant of the gender-fame task to examine errors that may occur under a more stark set of conditions (Walsh, Banaji, and Greenwald, 1995). Male names were varied in race between European American and African American (Frank Smith and Adam McCarthy or Tyrone Washington and Darnell Jones).[8] We suggested to the participants that they might have memory for these names, some of which were those of criminals. In fact, none of the names were names of criminals. Participants were told that some of the names on the list might seem familiar because they had appeared in the media. The task was to identify each name as criminal or noncriminal. In five experiments, we found that on average subjects "remembered" 1.7 times as many black than white names as criminals.

This finding was obtained with various proportions of white and black names (85:10; 50:50), and within and between subject designs. Additional experiments varied instructions that pointed to the race of the targets with varying degrees of explicitness, and in one case even informed participants that "people who are racist identify more black than white names; please do not use the race of the name in making your judgment." Such experiments have continued to reveal the bias. Our assessment is that participants who show the race bias believe their judgment to be based on a genuine memory for each identified name, black or white, and also believe that their assessment is not influenced by the race of the name. Yet belief produces a memory and, consistent with research on false memory (Roediger and McDermott, 1995), it leads in these circumstances to the greater false identification of black men as criminals.

More than any other experimental result from our laboratory, this finding has provoked a "rationality" defense of our participants' behavior. We report a relevant experimental variation here, simultaneously pointing out that the term "rational" as used in informal questions about this research has generally been innocent of any rigorous definition.[9] In the absence of other knowledge about the individual in question, it is rational, the argument goes, to use existing knowledge about the link between race and crime in completing the task at hand. In other words, the rational choice, it is argued, lies in using group membership in judgment.

If identifying proportionally more black than white names as criminal is indeed rational, then conditions allowing fuller access to cognitive resources should produce even greater rational behavior (that is, the greater identification of criminals on the basis of group membership) than conditions under which such access is restricted. In a separate study we did not find this to be the case. Subjects were asked to identify criminal names in a 2 × 2 design. One factor varied an instruction regarding racism: the control group was given no instruction, and the experimental group was alerted that "people have been found to associate criminality with African Americans more than with whites, Asians, or other ethnic groups. This is true for people who believe they have race prejudices (people who are racist) as well as for those of us who believe we are not prejudiced. Please try not to be influenced by the race of the name in making your judgments." The other factor varied the time available to complete the task. Participants were either self-paced throughout or were informed that they had only one minute to complete the task. The task actually took about one minute to complete, so the time-pressure instruction created only the expectation of a constraint.

When no instruction to avoid using the race of the name was offered, both the self-paced and the time-pressure conditions produced the familiar race bias. In other words, having greater access to cognitive resources in the self-paced condition failed to produce the putatively more rational response of greater criminal identification of black than white names. When the instruction to avoid using race of the name was in place, participants equalized their identification of black and white names as criminals in the self-paced condition, but still continued to show race bias in the time-pressure condition. Those who assume that a race bias is rational must concede that rationality apparently gives way to what would be considered the less rational response (of no race bias) when instructions to be fair and a perceived time constraint are present. That an instruction to achieve a prescribed goal (in this case of unbiased responding) and the availability of sufficient resources can change behavior (a race bias in criminal judgment) illustrates a fundamental characteristic of boundedly rational behavior: the domain dependence of these bounds and their malleability within specific problem contexts.

For decades, civil rights legislation has been premised on the assumption that to discriminate on the basis of group status (race, religion, sex) in decisions about individuals is unacceptable. When the decision is a judgment of criminality, when such judgments do not reflect an explicit prejudice on the part of the actor, and when the decision is based on cultural knowledge that is widely shared however dubious its origin, the consequences are deeply disturbing. Participants in our experiments are neither racist in the accepted sense of the term nor inclined to cause harm intentionally to the individuals they identify. In fact, explicit measures of racism and belief in the fairness of the criminal justice system show participants to be consciously egalitarian and fair-minded. Nonetheless, their behavior reveals the influence of beliefs on (false) memory about vital attributes of a person's character.

It is important that performance on explicit measures is not related to the magnitude of the race bias. Implicit and explicit stereotypes may be quite dissociated, as seen in these experiments, although they may come to be associated under other circumstances. We are still far from understanding the nature of the association between explicit and implicit attitudes and beliefs, and it is clear from more recent data (Lepore and Brown, 1997) that their relationship is by no means a simple one (see Blair, forthcoming).

In summary, the experiments on personality, fame, and criminality judgments demonstrate that ordinary conditions of judgment reveal the complex interaction of stereotypic beliefs and memory. In the gender-fame studies, participants were seduced by the familiarity of names to give males the greater benefit of fame implicitly, even when no such privilege was earned. In the personality experiments, harsher assessments of aggression and dependence were applied along gender-stereotypic lines, even when the differentially judged men and women had performed identical actions. In the race-crime studies, the costly judgment of criminality was levied disproportionately on nonguilty black men because of a false memory generated with surprising ease.

In spite of their ordinary and implicit operation, such judgments are not without consequence, for they clearly reveal the inequitable distribution of punishment and reward along lines of group membership. The impact of such judgments is seen to be even greater when their ordinary and implicit character reveals the ubiquity of their influence in everyday social interaction, and the slim opportunities that exist for self-doubt or disbelief about the poverty of the underlying mental due process.

Automatic Activation of Gender

Small differences in time can make large differences in the behavior of complex systems. If the sea of quarks had taken 10^{-30} instead of 10^{-35}

seconds to form, the shape and form of the universe would have been vastly different—if one had formed at all. The significance of small amounts of time, here on the order of milliseconds, is visible in many activities involving skill, such as music, cooking, and baseball. For example, expert opinion about the difference between a "not bad" and a "good" judgment on a catcher's release time is 9/100ths of a second (Will, 1990).[10] In the equally skilled game of social perception, differences in response latencies can reveal how the interaction of specific social experiences and a boundedly rational cognitive architecture jointly shape thought and behavior.

In some experiments we have taken small differences in the time to complete a social computation as an indicator of the strength of social beliefs, that is, the association between social groups and the qualities ascribed or denied to them. Time to respond to associations between social groups and physical or psychological qualities has allowed us to measure a particular component of unconscious thought: the lack of control over expressions of stereotypic beliefs and prejudicial attitudes (Banaji and Hardin, 1996; Blair and Banaji, 1996).

Our assumption is a simple and powerful one—that the speed of response to one stimulus in the context of another stimulus (related or unrelated) is an indicator of the underlying strength of association (semantic or evaluative) between the two (Meyer and Schvaneveldt, 1971; Neely, 1977). Such automatic responses capture thoughts and feelings that are deployed without conscious control, and our procedure has served well in exploring the strength of automatically activated beliefs by measuring their association in memory. Using a variation of a standard semantic priming technique, we presented gender-stereotypic words (emotional, aggressive, skirt, cigar) for short durations of approximately 300 milliseconds followed by male and female first names (Ann, Lisa, David, George). The speed of rapidly judging names to be either male or female was taken to be a measure of the strength of association in memory between these social groups and associated concepts (see also Dovidio, Evans, and Tyler, 1986).

A central feature of unconscious processes is the notion of control. A growing literature demonstrates that social actors' ability to control and modify their beliefs, judgments, and behavior is constrained by variables such as the awareness of inappropriate influences on judgments and behavior, the availability of cognitive resources to make spontaneous corrections, and the knowledge of suitable strategies to implement such corrections. The greater the degree of conscious deliberation that can be exerted over an action, a thought, or a feeling, the greater the assumed control over it.

Among the most fundamentally learned social categories is that of gender. Children show evidence of knowledge about gender and its associations at an early age (Fagot and Leinbach, 1989; Martin and Little, 1990). From our experiments we have solid evidence of the ability to classify gender-related information into female-male categories: first names (Jane, John), traits (nurturant, competitive), occupations (nurse, doctor), kinship (sister, brother), and verbal or pictorial representations of physical attributes (lipstick, cigar). Presenting prime-target pairs for approximately 300 milliseconds, we have shown that feminine primes strongly facilitate judgments of female over male names and that, analogously, masculine primes strongly facilitate judgments of male compared to female names. In other words, prime-target pairings whose gender association is congruent facilitate judgment when compared with pairings that do not share the property of gender (Banaji and Hardin, 1996; Blair and Banaji, 1996).

Implicit Attitudes

In a more recently developed task, Greenwald, McGhee, and Schwartz (1998) used an interference task, namely control, to capture the same process of unconscious judgment. The procedure, called the *implicit association test* (IAT), was devised to measure the strength of attitudes by assessing the extent to which two concepts (for instance, black–good/white–bad versus black–bad/white–good) are associated. The task requires participants to classify items from two categories (*black* names and *unpleasant* words) on a computer key while at the same time classifying items from two contrasting categories (*white* names and *pleasant* words) on a different key. Response latencies to perform this task are compared with trials in which the opposite categories are paired, that is, when black names/pleasant words are assigned to a single response key and white names/unpleasant words are assigned to a contrasting key. The underlying assumption is that if two concepts are evaluatively congruent (black-bad and white-good), trials that involve such pairings should be relatively easier than pairings that associate incongruent or less congruent concepts (black-good and white-bad). The difference in response latencies in the two types of pairings provides a measure of automatic attitude toward the group black compared with the group white.

In fact, Greenwald and colleagues (1998) showed that the IAT task is a quite powerful indicator of automatic attitudes toward nonsocial categories such as insects versus flowers and weapons versus musical instruments, with the vast majority of participants showing favorable attitudes toward flowers and instruments. In measuring social attitudes, this group found that independent of explicitly expressed attitudes toward social groups, white and Asian participants showed negative attitudes toward black Americans, and Korean Americans and Japanese Americans showed greater implicit liking for their respective ingroups compared with the outgroup.

The implicit association test permits measurement of attitudes and beliefs in a wide range of categories. We are currently conducting experiments that measure (a) automatic liking of male and female leaders with an interest in predicting voting behavior (Carpenter and Banaji, 1997); (b) automatic gender identity, gender attitude, and their relationship to each other (Lemm and Banaji, 1998); (c) automatic gender attitudes toward science and math versus language and arts, links between academic orientation and self-concept, and the developmental course of such preferences (Nosek, Banaji, and Greenwald, 1998); (d) the relationship between automatic self-esteem, group esteem, and group identity (Rosier, Banaji, and Greenwald, 1998); and (e) dissociated attitudes toward a single object (Mitchell, Nosek, and Banaji, 1998).

Taken together, the experiments on uncontrollable beliefs and attitudes demonstrate the difficulty in curbing unconscious associations between social groups and activation of stereotypic beliefs and prejudicial feelings toward them. In the automatic gender-stereotyping studies described earlier, we found that the absence of sufficient cognitive resources and a well-defined strategy disallowed conscious attempts at correction. The same was true in the experiments using the IAT. Subjective awareness of inability to perform as fast in the incompatible condition as in the compatible condition often accompanies inability to control automatic preferences and beliefs among participants, who include in their number the experimenters themselves.

The Bounded Rationality of Implicit Social Beliefs

The gender-fame experiments, the race-criminality experiments, and the experiments to measure automatic preferences apparently pose a tension between two positions that we refer to as *guilt by association* and *guilt by behavior.*[11]

On the one hand, it has been argued that the use of knowledge about social groups to make decisions about individual humans is appropriate and defensible. (We refer to this as the guilt-by-association position.) For instance, in a rousing defense of the accuracy of stereotyped judgments, McCauley, Jussim, and Lee (1995) say: "In this case [when no individuating information is available], the stereotype of the group is likely to dominate the evaluation of the stereotyped target (*as normatively it should*)" (p. 301; emphasis added). Often proponents of guilt-by-association decisions compare them to selection decisions about inanimate objects such as computers or restaurants. If, for example, the task is to pick the better of two working models of a mechanical gadget such as a computer, it would be quite appropriate to pick the manufacturer with the lower failure rate. Likewise, if the task is to identify criminals, the guilt-by-association position holds that the greater identification of black than white names in the race-criminality experiments is rational and defensible on grounds of base-rate information.

On the other hand, many personal and social codes of ethics hold that judgments about individuals should be based on an individual's own behavior without attention to group membership. According to this guilt-by-behavior position, it is implausible or incorrect to infer that the parents of murderers are more likely to be murderers because they belong to the same social group—family. Or that because police officers are convicted of crimes at a higher rate than the general population (Uviller, 1996), Officer X is a criminal.

This belief that guilt by association is morally repugnant is so fundamental that it has occupied a central place in all codes of justice from Ptahotep (Ptahotep, 2300 B.C.) to Hammurabi to Asoka (259

B.C.; see Nikam and McKeon, 1958) to the Assize of Clarendon (1166; see Plucknett, 1956) to all modern constitutions (with a small number of European exceptions in this century).[12]

The guilt-by-association position perhaps rests on a particular confusion between what most individuals are likely to do in a given situation and what is considered rational. The ordinary expression of a stereotype takes the form, Many Xs have property A, x is an X, therefore x has property A. Such routine inferences are exactly that—they are routine, and hence perhaps mistakenly assumed to be rational. As every schoolchild knows, behavior that is routine and seemingly reasonable need not be rational. For a decision to be rational, it must conform to the axioms of rationality in the sense of Savage (1972) or Arrow (1963).[13] For the axioms to be valid descriptors of behavior, the existence of a global utility function that captures all possible choices over time is a necessary and sufficient condition. Rational choice consists of decisions that always maximize this global utility function.

It can be shown that the requirements of these axioms are quite constraining (Debreu, 1971). In general, preferences cannot always be articulated, and when they are, they are not always consistent or necessarily stable over time. Additionally, most tasks cannot be readily represented within the confines of a global utility function, and information about the consequences of actions that allow rational choice is not always available. In the remainder of the chapter, our use of the word "rational" is restricted to this classic, axiomatic sense.[14]

Rational behavior, with this axiomatic import, has an all-or-none flavor. Behavior is either rational or irrational and, by definition, rational behavior is always correct. Our arguments about the bounded rationality of our subjects' behavior rely on the fundamental premise that *behavior that seems reasonable can be irrational and therefore incorrect.* Judging a single individual on the basis of information about her or his group can seem statistically justifiable, but cannot be justified by an appeal to rationality. We will demonstrate exactly how participants' behavior is not classically rational and how the departure from rationality can be explained by understanding the information-processing constraints that drive the behavior.

In addition, we suggest that the guilt-by-association position is based on a particular fallacy: two identical decision processes are seen to be equally acceptable even when their outputs have differential moral consequences—incalculable moral consequences can follow misjudgments of humans, whereas no difficulties accrue to an unselected computer. We argue that decision processes should be compared not on the basis of structural similarity alone, but also by taking a consequentialist approach attendant on the benefit or harm produced by the decision.

We devote the remainder of the chapter to assessing the kinds of judgments that the preceding experiments have highlighted. Specifically, our assessment will be based on two standards of good judgment (see Hastie and Rasinski, 1988). According to the first standard, judgments are considered to be correct, appropriate, justified, and ultimately defensible if they fit a coherent theory such as the axioms of probability or rationality. As an example, decisions in the classic "Linda problem" are considered to be incorrect when they fail to meet the conjunction relation—that is, P(A) and P(B) are both always greater than P(A, B). By the second standard, judgments are considered to be good if they fit the data, that is, are empirically verified. For example, a weather forecaster's performance can be assessed by comparing predictions to actual weather. We show that the judgments under scrutiny in this chapter may be considered vulnerable to both standards. In other words, judgments of the sort produced in these experiments are not consistent with either theory-fitting or data-fitting standards.

We further demonstrate how participants' behavior is not classically rational in that it adheres to other accepted criteria of incorrect or problematic judgments. Such behavior is shown to conform to the computational characteristics that exemplify boundedly rational behavior.

Not Classically Rational

Following nearly fifty years of research in psychology, we demonstrate that the behavior of participants in our experiments shows no evidence of classic rationality. Table 5.1 gives a partial list of the many utility functions that participants might choose (if they were rational). Inspection suggests why all of them are unlikely descriptors of behavior. Not only do the utility functions require computations that are

Table 5.1 Possible utility functions for participants in race-criminality experiments

1. Minimize $[(\text{black names/white names})_{\text{sample}} - (\text{black names/white names})_{\text{population}}]$
2. Minimize $[(\text{black names/white names})_{\text{sample}} - (\text{black names/white names})_{\text{arrested}}]$
3. Minimize $[(\text{black names/white names})_{\text{sample}} - (\text{black names/white names})_{\text{convicted}}]$
4. Minimize $[(\text{black names/white names})_{\text{sample}} - (\text{black names/white names})_{\text{incarcerated}}]$
5. Minimize $[(\text{criminal proportion})_{\text{sample}} - (\text{criminal proportion})_{\text{population}}]$
6. Minimize $[(\text{criminal proportion})_{\text{sample}} - (\text{criminal proportion})_{\text{arrested}}]$
7. Minimize $[(\text{criminal proportion})_{\text{sample}} - (\text{criminal proportion})_{\text{convicted}}]$
8. Minimize $[(\text{criminal proportion})_{\text{sample}} - (\text{criminal proportion})_{\text{incarcerated}}]$

Note: Utility functions 1 through 4 are race-conscious utility functions. Utility functions 5 through 8 are race neutral. All the utility functions require awareness of the properties of names in the general population (absolute and relative numbers of criminals and noncriminals, and so on). Each utility function also requires a participant to decide how many names to circle based on these ratios, using criteria that are extrinsic to the problem representation (such as which of the particular names to select, given the numerical outcome of a utility function).

too complex to be performed by subjects without a calculator, they also require data that even subjects keenly aware of the domain are unlikely to have (relative frequency of blacks and whites in America as a whole, of blacks and whites convicted of crimes, of arrested blacks and whites, of incarcerated blacks and whites, of black and white names in news reports, number of Type I and Type II errors in news reports, and the like). We do not dwell on this argument, its conclusions fortunately being in tune with decades of research showing that human behavior is not classically or axiomatically rational (March and Simon, 1958; Newell and Simon, 1972; Simon, 1955, 1976, 1983; Tversky and Kahneman, 1974).

In addition, specific findings from the experiments challenge the consistency of preference structures demanded by classic rationality, with participants' utility functions being malleable in a wide variety of experimental circumstances. In the gender-personality experiments, extremity of ratings (of aggression and dependence) shifted as a function of prior exposure to behaviors related to the personality concepts. In the gender-fame studies, false identifications increased after prior exposure to names. Similarly, in the race-criminality studies, the rate of misidentification was influenced by experimental manipulations of instruction, time pressure (Walsh, Banaji, and Greenwald, 1998), and mood state (Park and Banaji, 1998). Neither argument, based on plausibility and on experimental data, challenges the idea that the observed behaviors are reasonable, but they do not permit the assessment that the behaviors are rational. Bounded rationality is a more appropriate characterization of the behaviors we have encountered here.

Other Standards of Judgment

Disciplines vary in their methods for determining error. To show that using knowledge about a group (however correct it may be) to make judgments about individual members is best characterized as erroneous, we broadly define four criteria: universality of social practice, logic, intention, and analogy.

Social Practice. Across time and culture social practice has universally recognized the moral discomfort inherent in category-based social judgments. Our oldest and most remote example is the apocryphal story of how the sixth-century philosopher Sankara, Hinduism's most rigorous thinker, reached his epiphany into nondualism as the direct result of a category-based social judgment. Leaving the river after his ritual sacred bath, he (a Brahmin) brusquely ordered a man, obviously an untouchable, to step aside so as to avoid any possibility of physical contact. Sankara's shame at the discovery of his prejudice, when it turned out that the untouchable was a deep thinker, influenced the development of the important philosophy of *Advaita* (Iyer, 1964).

Less remotely, concerns with category-based social judgments have been a part of the American political psyche. In the last century, Justice Harlan's dissent in *Plessy* v. *Ferguson* (1897), among the most-cited opinions of the Supreme Court, states eloquently that category-based judgments involving race are immoral and cannot be the basis of public policy. In his dissent he wrote:

> Our constitution is color-blind, and neither knows nor tolerates classes among citizens . . . The law regards man as man, and takes no account of his surroundings or of his color when his civil rights as guaranteed by the supreme law of the land are in-

> volved. It is therefore to be regretted that this high tribunal, the final expositor of the fundamental law of the land, has reached the conclusion that it is competent for a state to regulate the enjoyment by citizens of their civil rights solely upon the basis of race. In my opinion, the judgment this day rendered will, in time, prove to be quite as pernicious as the decision made by this tribunal in the Dred Scott Case.

American history since has revealed the majority opinion's moral bankruptcy, but we cite Justice Harlan here to ask whether what appeared distasteful in 1897 for public policy might seem unacceptable in 1997 for interpersonal and intergroup social judgments.[15]

In the first half of this century, Walter Lippmann (1922) and Gordon Allport (1954) both emphasized the ordinary cognitive foundations of category-based judgments, yet their writings clearly reveal their recognition of the failures inherent in such judgments. Most poignantly, Gunnar Myrdal (1944) noted that Americans experience a moral dilemma, "*an ever-raging conflict between, on the one hand, the valuations preserved on the general plane which we shall call the 'American Creed,' where the American thinks, talks, and acts under the influence of high national and Christian precepts, and on the other hand, . . . group prejudice against particular persons or types of people . . . dominate his outlook*" (p. xlvii; emphasis in original). A half-century later, Devine's work strikingly shows the continued existence of the moral dilemma in the form of heightened guilt among American students confronting their prejudice (Devine, Monteith, Zuwerink, and Elliot, 1991; Zuwerink, Devine, Monteith, and Cook, 1996). It is surprising that with the backdrop of a history such as this from Harlan to Devine, McCauley and colleagues (1995) believe that to make category-based judgments is "normatively as it should be."

Logic. The inference, logically considered, that black-sounding names are more likely to be names of criminals is fallacious. When stating a stereotype in the form of a logical proposition, the appropriate logical quantifier is "some," "several," "many," or "a few," but almost never "all." The type of logical deduction revealed by experimental participants is of the following kind: "Some members of the set X have characteristic α. Object #<22310> is a member of the set X. Therefore object #<22310> has characteristic α." This deduction violates an elementary rule of Aristotelian logic, treating the proposition "Some members of the set X have characteristic α," as though it were the same as "All members of the set X have characteristic α."

Confusing the logical quantifier "some" with the logical quantifier "all" is the kind of error known in logic as a confinement law error (Kalish and Montague, 1964). Psychologists have labeled such errors in syllogistic reasoning as the atmosphere effect (Woodworth and Sells, 1935). Premises containing "some" create an atmosphere for accepting inferences that actually deserve the answer ". . . can't say—*no specific* conclusion follows from the premises. If a person accepts a specific conclusion for an invalid syllogism, that is an error in reasoning, and such errors frequently conform to predictions based on the atmosphere hypothesis" (Bourne, Dominowski, and Loftus, 1979, p. 277). To defend inferences based on stereotypes as accurate (Jussim, McCauley, and Lee, 1995) is thus to challenge a hitherto uncontested rule of Aristotelian logic.

Intention. Here we focus on a different argument, the thesis being that, under many circumstances, an outcome is considered incorrect if it is inconsistent with that which is intended.[16] Intending to draw a cube and having a cylinder emerge instead is obviously an error. Intending to drive on the right side of the road but ending up on the left is likewise an error. In a similar way, intending to feel and behave in line with one's values, but failing to do can be considered an error. In fact, recognizing the inconsistencies between "ought" and "actual" is apparently what accounts for the discomfort expressed when a mismatch between desired feelings and behaviors versus actual feelings and behavior is highlighted (Devine et al. 1991).

Unawareness of the discrepancy between intention and behavior as well as the discomfort that accompanies awareness of such discrepancies cannot justify the characterization of these errors as anything but errors. How a society should choose to deal with such errors and their consequences is a separate question and one that is beyond the scope of this chapter. Our purpose is to emphasize that conclusions about decisionmaking that are disturbing ought not to be mischaracterized as benign or correct.

Analogy. A final argument for considering the experimental results as representing error can be made by analogy. In other areas where criteria of incorrectness similar to those in our experiments are met, the behavior is routinely classified as an error. For example, the long history of research on perceptual illusions (errors) contains many examples of identical objects that nevertheless violate all perceptual experience that they are so. When two objects that are identical in shape and size (such as tabletops in Shepard's parallelogram illusion, 1990, p. 48) are perceived to be dissimilar, we regard the resulting misperception as a remarkable error. Explanations concerning the origin of the perceptual error do not produce a desire to recategorize the error as reflecting a correct judgment.

Likewise, when two behaviors are identical (one performed by Donna, the other by Donald) but are not judged to be so, we must regard the resulting misperception as an error. Interest in false memory led Roediger and McDermott (1995) to replicate an earlier finding that presenting lists of words related to a concept (for example, sleep-related words such as "dream" or "pillow") can produce a false memory for the word "sleep," an item which never appeared on the list.

As the label "false memory" itself suggests, the obtained misidentifications are considered to be false by definition, and scientists who study memory do not become confused about whether to regard such false memories as revealing error. Likewise, mistakenly "remembering" a person who is not a criminal as a criminal is an error by definition, and the apparent confusion it creates about whether to regard it as an error or not may most charitably be understood as reflecting a desire to avoid confronting the seamy side of decisionmaking that accompanies such social judgments.

Computational Characteristics of Bounded Rationality

We have seen how participants' behavior fails to meet conventional criteria for classic rationality and criteria for correctness. Now we use a computational approach to show how the ordinary and implicit cognitive processes that underlie stereotyping can collectively produce behavior that is boundedly rational. We do so by discussing how errors can arise when representations of problems in one domain are mistakenly applied to superficially similar but substantively distinct problems in another domain. We noted previously that the behavior is boundedly rational by demonstrating how computational constraints can be significant determinants; for instance, relaxing a particular constraint such as time can produce a change in the behavior.

In studies of human problem solving, an established finding concerns the inability to represent superficially similar but substantively different problems as distinct. The assumption of similarity can lead to the use of inappropriate methods to solve the problem at hand.[17] For example, Bhaskar and Simon (1977) showed that a problem was misclassified as a thermodynamics problem rather than a physics problem because a copy of steam tables (sometimes necessary for solving thermodynamics problems, but never necessary for solving physics problems) was made available. This caused a lengthy detour, resulting in approximately thirty minutes as opposed to five minutes to solve the problem. Similarly, Hinsley, Hayes, and Simon (1977) reported that subjects skilled at algebra word problems spent large amounts of time on nonsense problems worded similarly to real problems, even though a superficial examination would have revealed the nonsensical character of the alleged problem. Such experiments have confirmed that humans sometimes represent problems inappropriately by failing to see real differences between them. Such misconstrual is a central feature of boundedly rational behavior, and here we describe how it may be implicated in the errors of stereotyping we have observed.

When confronting the task of criminal identification, what is the alternative representation that participants might use? As mentioned before, a task that has informally but frequently been raised as analogous is the task of identifying poorly performing models of a mechanical object such as a computer. It has been proposed that just as it is reasonable to select a functioning computer over a dysfunctional one based on manufacturer history, it is also reasonable to identify more black than white names as criminals.

Signal detection theory offers a useful way to represent a problem solver's decisionmaking on both of these tasks. The goal of the hypothetical computer identification task is to identify dysfunctional computers (poorly performing models); the goal of participants in our experiments is to identify dysfunctional humans (criminals). In one case, the information that is supplied is the manufacturer's name, in the other case it is the person's race. For the computer identification task, the immediate goal is to identify as many poorly performing computers as possible. *Hits* represent poor models correctly identified as such, *misses* represent poor models identified as good ones, *false alarms* represent good models incorrectly identified as poor models, and *correct rejections* represent identification of good models as good. *In the computer selection task, misses have a high cost.* That is, incorrectly labeling a poor model as good can result in one's ending up with a computer that breaks down often. A false alarm, which simply implies that one may have rejected a good computer, is essentially costless. This is because the problem solver's basic objective, which presumably is to "acquire a good model by ruling out all poor models from the candidate set," is not frustrated by false alarms. That is, a false alarm cannot lead to the selection of a poor computer.

Such a representation is consistent with the task performed by our subjects. Analogous to the computer task, hits represent criminals being correctly identified as criminals, and correct rejections involve correctly identifying noncriminals as noncriminals. False alarms represent noncriminals misidentified as criminals, and misses represent the misidentification of criminals as noncriminals. In the criminal selection task, as in the computer selection task, misses are costly (incorrectly labeling a criminal as noncriminal can lead to a criminal's going free). The similarity between the two tasks ends here. *In the criminal identification task, false alarms have an incalculably high cost.*

This difference in the false-alarm costs of the two tasks is sufficiently significant that a representation for one task is inappropriate for the other. For example, the computer selection task involves minimizing misses while producing an unlimited number of false alarms. In contrast, the criminal identification task requires that both misses and false alarms are to be managed because of the high costs of both, and especially those of false alarms. The two tasks are quite distinct. Yet, as the false alarms (on both black and white names) suggest, participants did not use a rule strict enough to prevent *any* false alarms, the only correct outcome. Although the two problems of computer and criminal identification may seem to be the same, the failure to recognize their difference is no different from the misidentification of the physics problem as a thermodynamics problem; only the consequences are graver.

Why is the cost of false alarms high in the criminal identification task? Our society, and most liberal societies, generally proceed on the principle that it is important not to declare the innocent to be guilty. When we consider the many possible objectives of criminal punishment—deterrence, incapacitation, just punishment, and rehabilitation (U.S. Sentencing Commission, 1996, p. 1)—we see that the possible innocence of the punished frustrates every social objective. A decision procedure that ignores the cost of false alarms, however plausibly or excusably, violates basic and almost universally accepted concepts of justice and fairness. Such cost is only more profound when we consider that the incorrect application of guilt is selectively leveled against particular social groups. Thus, the computational path from ordinary cognition to ordinary prejudice ultimately reveals the extraordinary moral burden imposed by human bounded rationality.

Conclusion

In the past, stereotypic beliefs and prejudicial attitudes were largely conceived of as conscious and were treated as outside the interpretive scope of ordinary cognition. Decades of research in social psychology have refuted both myths. It is now evident that the computational and unconscious character of stereotypes and prejudice does not require appeal to the operation of unique processes or unique persons. Rather, as the sample data presented here show, stereotyped beliefs and prejudicial attitudes multifariously reveal their presence through ordinary biases rooted in memory. Social psychology's refutation of these myths has come at a price—the perception that demonstrating the ordinary computational nature of stereotyping and prejudice dissociates it from its moral impact. We have argued to the contrary, that the bounded rationality of human social cognition reveals the hitherto unrecognized but deeply moral quality intrinsic to theories of human judgment and decisionmaking.

References

Abelson, R. P. (1986). Beliefs are like possessions. *Journal for the Theory of Social Behavior, 16,* 223–250.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D., and Sanford, R. N. (1950). *The authoritarian personality.* New York: Harper.

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

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Armour, J. D. (1997). Hype and reality in affirmative action. *University of Colorado Law Review, 68,* 1173–1210.

Arrow, K. J. (1963). *Social choice and individual values.* New Haven: Yale University Press.

Ashmore, R. D., and Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum.

Banaji, M. R. (1995). The significance of an 8 millisecond effect. Paper presented to the Society of Experimental Social Psychology, Washington, DC.

Banaji, M. R. (1997). Introductory comments. *Journal of Experimental Social Psychology, 33,* 449–450.

Banaji, M. R., and Greenwald, A. G. (1994). Implicit stereotyping and unconscious prejudice. In M. P. Zanna and J. M. Olson (Eds.), *The psychology of prejudice: The Ontario symposium,* Vol. 7 (pp. 55–76). Hillsdale, NJ: Lawrence Erlbaum.

Banaji, M. R., and Greenwald, A. G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68,* 181–198.

Banaji, M. R., and Hardin, C. (1996). Automatic stereotyping. *Psychological Science, 7,* 136–141.

Banaji, M. R., Hardin, C., and Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology, 65,* 272–281.

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

Bhaskar, R., and Simon, H. A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. *Cognitive Science, 2,* 192–215.

Billig, M. (1996). *Arguing and thinking: A rhetorical approach to social psychology* (2nd ed.). Cambridge: Cambridge University Press.

Blair, I. V. (Forthcoming). Implicit stereotypes and prejudice. In G. Moskowitz (Ed.), *Future directions in social cognition.*

Blair, I., and Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70,* 1142–63.

Bourne, L. E., Jr., Dominowski, R. L., and Loftus, E. F. (1979). *Cognitive processes.* Englewood Cliffs, NJ: Prentice-Hall.

Carpenter, S. J., and Banaji, M. R. (1997). Implicit attitudes toward female leaders. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Chen, M., and Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology, 33,* 541–560.

Debreu, G. (1971). *A theory of value.* New York: Wiley.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18.

Devine, P. G., Monteith, M., Zuwerink, J. R., and Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60,* 817–830.

Dovidio, J. F., Evans, N. E., and Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology, 22,* 22–37.

Fagot, B. I., and Leinbach, M. D. (1989). The young child's gender schema: Environmental input, internal organization. *Child Development, 60,* 663–672.

Fazio, R. H., Jackson, J. R., Dunton, B. C., and Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–27.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., and Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *The handbook of social psychology* (4th ed.), (pp. 357–411). New York: McGraw-Hill.

Gewirth, A. (1996). *The community of rights.* Chicago: University of Chicago Press.

Greenwald, A. G., and Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 1–27.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–80.

Hamilton, D. L. (1981). *Cognitive processes in stereotyping and intergroup behavior.* Hillsdale, NJ: Lawrence Erlbaum.

Hart, H. L. A. (1976). *The concept of law.* New York: Oxford University Press.

Hastie, R., and Rasinski, K. A. (1988). The concept of accuracy in social judgment. In D. Bar-Tal and A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 193–208). New York: Cambridge University Press.

Henderson-King, E. I., and Nisbett, R. E. (1996). Anti-black prejudice as a function of exposure to the negative behavior of a single black person. *Journal of Personality and Social Psychology, 71,* 654–664.

Hinsley, D., Hayes, J. R., and Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In P. Carpenter and M. A. Just (Eds.), *Cognitive processes in comprehension* (pp. 89–106). Hillsdale, NJ: Lawrence Erlbaum.

Iyer, M. K. V. (1964). *Advaita Vedanta according to Sankara.* New York: Asia Publishing House.

Jacoby, L. L., Kelley, C. M., Brown, J., and Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56,* 326–338.

Jussim, L. J., McCauley, C. R., and Lee, Y-T. (1995). Why study stereotype accuracy and inaccuracy? In Y. T. Lee, L. J. Jussim, and C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 3–27). Washington, DC: American Psychological Association.

Kalish, D., and Montague, R. (1964). *Symbolic logic: Techniques of formal reasoning.* New York: Harcourt Brace Jovanovich.

Kotovsky, K., and Fallside, D. (1989). Representation and transfer in problem solving. In D. Klahr and K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert Simon.* Hillsdale, NJ: Lawrence Erlbaum.

Lemm, K. M, and Banaji, M. R. (1998). Implicit and explicit gender identity and attitudes toward gender. Paper presented at the annual meeting of the Midwestern Psychological Society, Chicago.

Lepore, L., and Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Social and Personality Psychology, 72,* 275–287.

Lewicki, P. (1986). *Nonconscious social information processing.* New York: Academic Press.

Lippmann, W. (1922). *Public opinion.* New York: Harcourt Brace.

Lombardi, W. J., Higgins, E. T., and Bargh, J. A. (1987). The role of consciousness in priming effects on categorization: Assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin, 13,* 411–429.

March, J. G., and Simon, H. A. (1958). *Organizations.* New York: Wiley.

Martin, C. L., and Little, J. K. (1990). The relation of gender understanding to children's sex-typed preferences and gender stereotypes. *Child Development, 61,* 1427–39.

McCauley, C. R., Jussim, L. J., and Lee, Y-T. (1995). Stereotype accuracy: Toward appreciating group differences. In Y-T. Lee, L. J. Jussim, and C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 293–312). Washington, DC: American Psychological Association.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26,* 446–456.

Meyer, D., and Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.

Mitchell, J., Nosek, B., and Banaji, M. R. (1998). Dissociated attitudes. Paper presented at the annual meeting of the American Psychological Society, Washington, DC.

Myrdal, G. (1944). *An American dilemma: The Negro problem in modern democracy.* New York: Harper.

Neely, J. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106,* 226–254.

Newell, A., and Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nikam, N. A., and McKeon, R. (1958). *Asoka, king of Magadha.* Chicago: University of Chicago Press.

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

Nosek, B., Banaji, M. R., and Greenwald, A. G. (1998). Gender differences in implicit attitude and self-concept toward mathematics and science. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Nozick, R. (1993). *The nature of rationality.* Princeton: Princeton University Press.

Park, J., and Banaji, M. R. (1998). The influence of positive mood on implicit stereotyping. Paper presented at the annual meeting of the American Psychological Society, Washington, DC.

Plessy v. Ferguson, 163, U.S. 537 (1897).

Plucknett, T. F. T. (1956). *A concise history of the common law.* Boston: Little, Brown.

Ptahotep (2300 B.C.). *The instruction of Ptahotep (6th Dynasty) 2300–2150 B.C.*

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Belknap Press, Harvard University Press.

Roediger, H. L., and McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Rosier, M. D., Banaji, M. R., and Greenwald, A. G. (1998). The implicit association test, group membership and self esteem. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Savage, R. (1972). *The foundations of statistics.* New York: Dover.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.

Shepard, R. N. (1990). *Mind sights.* New York: Freeman.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69,* 99–118.

Simon, H. A. (1976). *Administrative behavior.* 3rd ed. New York: Macmillan.

Simon, H. A. (1983). *Reason in human affairs.* Palo Alto, CA: Stanford University Press.

Simon, H. A., and Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrika, 29,* 111–138.

Strack, F., Schwarz, N., Bless, H., Kubler, A., and Wanke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology, 23,* 53–62.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–31.

Uleman, J. S. (1987). Consciousness and control: The case of spontaneous trait inferences. *Personality and Social Psychology Bulletin, 13,* 337–354.

U.S. Sentencing Commission. (1996). *Federal sentencing guidelines.* St. Paul, MN: West Publishing Company.

Uviller, H. R. (1996). *Virtual justice: The flawed prosecution of crime in America.* New Haven: Yale University Press.

Walsh, W., Banaji, M. R., and Greenwald, A. G. (1995). *A failure to eliminate race bias in judgments of criminals.* New York: American Psychological Society.

Walsh, W., Banaji, M. R., and Greenwald, A. G. (1998). The misidentification of black men as criminals. Manuscript, Yale University.

Wegener, D. T., and Petty, R. E. (In press). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology.*

Will, G. (1990). *Men at work.* New York: G. K. Hall.

Woodworth, R. S., and Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451–460.

Zuwerink, J. R., Devine, P. G., Monteith, M. J., Cook, D. A. (1996). Prejudice toward blacks: With and without compunction? *Basic and Applied Social Psychology, 18,* 131–150.

Notes

1. This chapter is concerned with beliefs about social groups, namely stereotypes. As a result, a natural theoretical connection is extended to the construct, prejudice. Following convention, by *stereotype* we refer to the cognitive component, or beliefs about social groups (for example, Politicians are crooks) and by *prejudice* to the affective component, or attitudes about social groups (I dislike politicians).
2. For a resistance to the view that stereotyping and prejudice are not acts of ordinary cognition, but in some sense reflect special processes, see Billig (1996, pp. 158–170).
3. Paraphrased from Max Klinger of *MASH.*
4. The experiments reported in this chapter involve a continuing collaboration with Tony Greenwald, and research with several students past and present, mostly notably Irene Blair, Siri Carpenter, Buju Dasgupta, Jack Glaser, Aiden Gregg, Curtis Hardin, John Jost, Kristi Lemm, Jason Mitchell, Brian Nosek, Jai Park, Marshall Rosier, Alex Rothman, and Wendi Walsh.
5. Bounded rationality and the information-processing approach to psychology have been the defining, paradigm-shifting concepts of modern psychological and social science. As is all too common with such sweeping transformations, it is entirely possible that a new generation of readers does not have a clear sense of the meaning of the term *bounded rationality.* When we declare behavior to be boundedly rational, we deem it to have the following characteristics (March and Simon, 1958, p. 169): (a) behavior is satisficing rather than optimizing; (b) alternatives for actions are explored through sequential processes; (c) these sequential processes largely use specialized, domain-specific knowledge, rather than general, domain-independent problem-solving strategies; (d) each sequential process is restricted in the scope of problems it can deal with; and (e) the collections of processes are largely independent of one another, so that the memory and problem-solving system is best viewed as a collection of loosely coupled, "nearly decomposable" units (Simon and Ando, 1961).
6. See Ashmore and Del Boca (1981) and Banaji and Greenwald (1994) for comments about the historical transformation in definitions of stereotypes, from treating them as exaggerations and incorrect judgments to focusing on the application of group knowledge (accurate or inaccurate) to judgments of individuals.
7. Data from other investigators suggest that the stereotyping effects obtained in our studies may have been removed or reversed in the presence of awareness of the activating or priming event (see Lombardi, Higgins, and Bargh, 1987; Strack, Schwarz, Bless, Kubler, and Wanke, 1993; Wegener and Petty, 1997).

8. Names used in these experiments were generated by the experimenters and by research participants. Each name was then judged by a new group of participants for the likelihood that it was a European-American or African-American name (on a five-point scale). Selected names in each category were those judged to be high in the likelihood of being African American (or European American) *and* low in likelihood of being European American (or African American).
9. We will shortly discuss the question of why our participants do not meet axiomatic criteria of rationality.
10. We thank John Jost for bringing this source to our attention.
11. It should be obvious that positive judgments that confer benefits on recipients (instead of guilt), if they are differentially administered as a function of group membership, have a similar discriminatory effect.
12. Guilt by association is to be carefully distinguished from punishment by association. Even in T'ang China, when family members of traitors were executed it was not assumed that they were guilty. The punishment was most likely for reasons of deterrence and retribution.
13. These axioms are as follows:

    1. Given any pair of outcomes A and B, it is always true that A *is preferred to* B ($A \geq B$), B *is preferred to* A ($B > A$), or one is *indifferent* between A and B ($A \ngtr B$ and $B \ngtr A$).
    2. Preferences are transitive (if $A > B$ and $B > C$, then $A > C$).
    3. If action a_1 leads to A, and action a_2 leads to B, and $A > B$, then $a_1 > a_2$.

14. Such an axiomatic characterization of rationality is generally agreed to be somewhat sterile, a perception that has spawned alternative conceptions of rationality that are descriptively plausible (Anderson, 1990; Gewirth, 1996; Nozick, 1993; Rawls, 1971). Such accounts have generally been more successful in capturing a commonsense conception of rationality, simultaneously discarding some of the more implausible, unpersuasive aspects of classic axiomatic rationality (global consistency and utility functions, universally specified preference structures, and the like).

    In spite of its sterility, the classic axiomatic conception of rationality provides a syntactic framework that can house many of the alternatives that have been suggested. Even when the requirement of global consistency is abandoned, every alternative descriptive conception of rationality relies on representing it as adaptive behavior (see Anderson, 1990, p. 28). This takes the form, typically, of representing behavior as constrained optimization. (By "constrained optimization" we imply the conventional sense used in mathematics, economics, or operations research: the objective is to maximize or minimize the numerical value of some mathematical functions, while each of a number of other mathematical functions—the constraints—are required not to fall below or exceed some other specified value. Much of the success of modern social science derives from the generality of this representation.)

We will see that every alternative conception of rationality considered to be less sterile and more descriptively plausible can be cast in the form of utility functions that must be maximized, subject to certain constraints. Within the stereotyping context, we show that the utility functions and constraints that might represent rationality in any of these other more descriptively plausible senses require the participants in our experiments to either (a) use knowledge that they are unlikely to have, or (b) make assumptions that cannot meet their own expressed moral beliefs and standards. Demonstrating the difficulty of using classic axiomatic rationality as a valid descriptor of participants' behavior serves therefore as a vehicle for simultaneously demonstrating the same difficulty with these more substantive and plausible conceptions of rationality.

15. We point out the applicability of judicial decisions to discussions about interpersonal decisions. On the other hand, legal scholars (see Armour, 1997) have applied the evidence about biases in interpersonal judgment to matters of public policy such as affirmative action.
16. Not all failures to implement intended behavior are errors. To pick an extreme example, intending to kill someone and not being able to follow through would hardly be seen as an error.
17. Conversely, sometimes superficially dissimilar problems that are substantively the same are not recognized as such (Kotovsky and Fallside, 1989).

Acknowledgments

The research and writing of this chapter were supported in part by NSF Grant SBR-9422241, a triennial leave from Yale University, and fellowships from the John Simon Guggenheim Foundation and Cattell Fund to M. R. Banaji. We thank John Bargh, Wil Cunningham, Nilanjana Dasgupta, Robyn Dawes, Jack Glaser, Richard Hackman, Reid Hastie, Jason Mitchell, Brian Nosek, Peter Salovey, Daniel Schacter, Yaacov Schul, Wendi Walsh, and Bernd Wittenbrink for their comments on a previous version.