Banaji, M.R. & Dasgupta N. (1998).The Consciousness of Social Beliefs: A Program of Research on Stereotyping and Prejudice. In V.Y. Yzerbyt, G. Lories, & B. Dardenne (eds), MetaCognition: Cognitive and Social Dimensions(157-170). London: Sage Publications.

9

The Consciousness of Social Beliefs: A Program of Research on Stereotyping and Prejudice

Mahzarin R. Banaji and Nilanjana Dasgupta

In the mid-1990s, two important volumes on metacognition appeared. A collection of core readings containing classic and contemporary articles (Nelson, 1992) was followed by a volume of recent theoretical and empirical contributions (Metcalfe & Shimamura, 1994). Together they showed the prominence that the study of metacognition has come to occupy in psychology, and are testimony to the unique advances that are possible through an explicit effort to examine self-reflective processes. Through this research, the use of terms such as monitoring, control, feeling of knowing, and consciousness made previously marginalized constructs legitimate targets of scientific analysis. In so doing, the study of metacognition has expanded the realm of research questions that future generations of psychologists will be permitted to ask about cognition.

The present volume is unusual in its inclusion of social psychological perspectives on metacognition, and in this regard stands in contrast even to its two immediate predecessors. The gathering of social psychological perspectives is more than a simple addition to ongoing analyses, for social psychology has historically been engaged in the study of processes that assume self-reflection. Whether it be the study of attitudes, beliefs, or self-related processes, metacognitive processes have been centrally implicated in theory and research. To study an individual's beliefs about a social group, or attitudes toward political events, or assessments of self-worth, fundamentally requires an assumption that such knowledge exists at levels of consciousness to which access is possible. In addition, the seeming disparities between attitudes and action, between intention and behavior, between the proffered and real causes of behavior, have made metacognitive processes of natural interest. The inclusion of social psychology's core concerns in ongoing analyses of metacognition influences the nature of the theoretical questions that are asked and the target domains that are studied.

The joint focus on social and cognitive perspectives highlights an interesting divergence in the manner in which the histories of the two fields have unfolded with regard to the study of mental processes more generally. In cognitive psychology, the understanding is that metacognitive processes

have been ignored, and only through explicit argument have they been included in the fray of legitimate questions. The remnants of a displeasure with introspection practiced at the turn of the century and the behaviorist interlude are cited as historical reasons that kept the study of metacognition at bay (Nelson, 1992; Tulving, 1994). In social psychology, where the dominant method routinely required self-reports of mental processes such as feelings, opinions, beliefs, intentions, and values, the output of conscious, self-aware entities reflecting on the contents of their consciousness was hardly questioned. In fact, it is only rarely that the problematic aspects of a social psychology that has been so constructed have been questioned (Greenwald & Banaji, 1995; Nisbett & Wilson, 1977). The meeting point of two fields with differing priorities but many of the same fundamental concerns is bound to be an interesting one.

Our interest focuses on the ways in which characteristic features of consciousness (such as awareness, intentionality, and control) shape beliefs, attitudes, and behavior. For the past several years, we have been engaged in a program of research specifically concerned with beliefs and attitudes toward social groups and their members. Although the target domain may be most easily labeled as the study of stereotyping and prejudice, the core issues concern questions of consciousness. The domain of stereotyping and prejudice has unique features when viewed through the metacognitive lens. The most obvious concern is with how humans make use of knowledge that is known about a category (Many Xs are Y) in judgments of instances (X_1 is Y). This domain also tackles the disparity between knowledge that is inherent in a culture as a whole (Xs are Y) and an individual's own endorsement of that belief (Xs are not Y). To what extent do judgments reflect culturally held beliefs versus ones that are consciously endorsed by the individual? Are individuals able to control and shape their judgments in accordance with their conscious intentions? And finally, in contemporary societies that agree on the negative social consequences of stereotyping and prejudice, this domain offers an opportunity to examine the similarities and differences in the actions of those who hold consciously favorable attitudes from those who hold consciously unfavorable ones. Do such groups also vary in their implicit or unconsciously expressed beliefs?

Although several attempts have been made to offer a classification of the questions regarding consciousness, in this context we will work with one suggested by Johnson-Laird (1983). To answer the question "What should a theory of consciousness explain?" Johnson-Laird proposed a tractable set of problems that a theory of consciousness must solve. Four such problems were generated, with the goal of making the study of consciousness amenable to uniquely psychological (rather than philosophical) inquiry: awareness, control, self-awareness, and intentionality. In this chapter, we use the issues and data generated by the target domain of implicit social beliefs as a relatively unique platform to analyze questions of consciousness. In particular, we study beliefs about social groups (e.g. gender, race) that are spontaneously used by participants, but without awareness of their

usage, without control over their expression, and without intention to use them in judgment of others. The issue of self-awareness will not be addressed here, for our data do not speak directly to this aspect of consciousness. If we venture beyond the data themselves, issues of awareness, control, and intentionality also speak to the troubling and largely philosophical discussion to date regarding the responsibility individuals have for their actions, and the legitimacy of individual rewards and punishments for actions that are attributed to conscious agents. These are essential questions that the study of metacognition raises, but they remain muted if analyses remain focused on metacognitive processes in traditional domains (e.g. test performance, puzzle-solving).

The problem of awareness

For many judgments and decisions humans make, there is a perceived cause of the behavior that is assumed by the actor to be the actual cause. Such causes may often be offered in self-reports to explain or justify actions. Decisions and judgments can be assumed to be guided by higher-order beliefs, such as in the hypothetical statement "It is important to judge X fairly." Participants in our experiments on stereotyping and prejudice, with few exceptions, would endorse such a statement, perhaps even agreeing with more elaborate statements of fairness in the treatment of individuals. Yet, as a growing literature in social psychology demonstrates, there is not sufficient reason to assume that decisions fall into line with self-reports of higher-order beliefs, nor that there is reassuring accuracy in prediction of the actual cause of an action. In this regard, the findings we will highlight will bear some resemblance to the theme of other research on metacognition such as the inability to know what is known (Glenberg, Wilkinson, & Epstein, 1982), the shaky basis of confidence judgments (Loftus, Miller, & Burns, 1978; Shaw, 1996; Wilson & LaFleur, 1995), the difficulty with reality monitoring (Johnson & Raye, 1981), and more generally to research on judgments elicited under conditions of uncertainty.

Consider the task for a subject in one of our experiments based on a method used by Larry Jacoby (Jacoby, Kelley, Brown, & Jasechko, 1989) to study the unconscious influence of the past on the present. The subject is exposed to a list of names, famous and non-famous, male and female. Later, the subject is presented with the same names in addition to new (previously unseen) names with the same characteristics. The task is to identify whether each name represents the name of a famous person or not. Faced with this task, Jacoby et al. (1989) correctly predicted the specific error that subjects are poised to make. Unable to separate the source of familiarity of a name (i.e. the familiarity that accrues to a name from prior exposure versus familiarity that accrues from the actual fame of the name), participants are twice as likely to incorrectly judge a familiar (previously seen) non-famous name to be famous than an unfamiliar (previously unseen) non-famous name

to be famous. A mistaken belief about the source of familiarity leads to an erroneous attribution of familiarity to fame. The source of the bias stems from the often correct logic "This name feels familiar, therefore it must be famous," that nevertheless fails in this ordinary and commonly occurring context.

Our interest being in social groups, the additional variable of name gender was introduced, and the finding across several experiments bore out the hypothesis that the accurate belief of greater male fame would operate through the more likely assignment of fame to non-famous male than female names (Banaji & Greenwald, 1995). A feeling of familiarity with previously exposed names was assumed to interact with a general belief about greater male fame to produce the faulty attribution on familiarized non-famous male names. In this case the belief is true when applied to the population as a whole (i.e. fame is indeed more strongly associated with males as a group than females as a group), but the application of the belief in the individual cases captured in this experimental analog represents an error. The belief in greater male fame can be quite easily verbalized, but in this context, the application of the belief appears to operate without awareness. We know from questions posed to subjects that they remained quite unaware of the source of influence (i.e. gender) that urged judgment of a familiarized male name to be famous than an equally familiarized female name. Such unawareness produced a particular decision effect, as revealed in signal detection analysis, specifically, in the differential criterion for judging familiarized male versus female names: The subjective threshold or criterion, captured by the statistic β for judging male fame was set significantly lower than that for judging female fame.

Being unaware of the source of influence on one's judgment (in this case, being unable to control the effects of prior exposure and being unaware of the role of gender in influencing judgment) is not an uncommon occurrence. These experiments capture the ways in which our beliefs, operating unconsciously, can lead to benefits such as fame being undeservedly bestowed (or not) on unsuspecting targets (see Banaji, Blair, & Glaser, 1997). Here, the problem of awareness is the problem of a self-reflective being whose bounded rationality also leads to errors of consequence. The same fundamental processes that allow effective categorization and generalization also produce judgments that may be inaccurate and inequitable.

In another series of studies, we temporarily activated abstract knowledge about specific constructs such as *dependence* and *aggressiveness* (Banaji, Hardin, & Rothman, 1993) and in a quite different setting obtained judgments of individuals named Donna and Donald who performed identical actions. Following the large literature on construct accessibility effects (Higgins, 1989), we predicted that previous exposure to abstract knowledge about traits would increase their use in person judgment, but only when the gender of the specific target was stereotypically congruent with the previously activated knowledge. Even more strongly than expected, results showed that previously activated abstract knowledge did not influence

person judgment at all when the target did not carry the stereotypic group marker (i.e. 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 in the condition of jointly occurring knowledge activation and the fit of stereotypic group membership (i.e. when a female target was judged after exposure to dependence-related knowledge and a male target was judged after exposure to aggression-related knowledge).

Rather than a specific feeling of familiarity with a particular item of knowledge as in the previous fame experiments, exposure to abstract statements appears to have changed the threshold of judgment such that passers-by who fit the social category associated with the activation were handed a more extreme negative judgment. Had awareness of the influencing agent existed, the judgment outcome would have surely differed. As other research indicates, metacognitive correction processes are often engaged in the presence of awareness of perceived bias. Awareness of prior activation has been found to alleviate bias and sometimes even reverse its direction (Lombardi, Higgins, & Bargh, 1987; Strack, Schwarz, Bless, Kubler, & Wanke, 1993; Wegener & Petty, in press). These data suggest that the effects obtained in the present studies may have been removed or reversed in the presence of awareness.

In ongoing research (Walsh, Banaji, & Greenwald, 1995), we have used a variant of the gender-fame task to examine errors that may occur under even more striking cognitive circumstances. Subjects are asked to make a judgment on names that also vary in social category – in this case, however, the judgment is one of criminality, and the names vary in race (black, white, Asian). Importantly, a different basis for familiarity is provided that involves no prior exposure to names. Unlike the fame studies where previous familiarity with names was necessary to create uncertainty about the cause of later perceptual fluency, and unlike the trait judgment studies in which trait knowledge was activated in an unrelated context prior to judgment, in these studies we merely suggested that memory for names may exist. Subjects were told that some of the names on the list might be familiar to them because they had appeared in the media as names of criminals. In multiple experiments, we have shown that this instruction alone can produce one and a half times more black than white identifications with the producers of this error being persuaded that their judgment was based on genuine memory for criminal names.

Among the surprising aspects of this research has been the difficulty in removing the race bias in spite of specific instructions to do so, including alerting subjects that racist individuals are more likely to identify black compared with white names. Beliefs about social groups, whether they are descriptors of the group or not, are in obvious error when applied to the individual case in which they are undeserved, as many decades of civil rights legislation remind us. The participants in our experiments are neither racist in the accepted sense, nor are they intentionally inclined to cause harm to

the individuals they identified as criminals. In fact, explicit measures of racism and belief in the fairness of the criminal justice system show participants to be egalitarian and even to be progressive moral agents. However, such beliefs are not correlated with the bias observed on the criminal name identification task. Performance on these two tasks are guided by different types of knowledge. These data reveal that the mere suggestion of name familiarity (in the absence of actual familiarity) is sufficient to produce misidentifications with potentially serious consequences.

Together, these experiments reveal that awareness of the source of influence on judgment is not always or easily possible, and that such conditions are ideal to study the unconscious influence of social beliefs and memory on judgment (Greenwald & Banaji, 1995). In the fame studies, it was difficult for participants to undertake the metacognitive exercise of knowing the source of felt familiarity of a name. In the trait judgment studies, the influence of the prior event was even better hidden from awareness, perhaps even leading perceivers to the belief that their judgment reflected properties of the target itself. Finally, in the race-crime studies, knowledge about the link between race and crime at the group level was sufficient to cause individual misidentifications in the absence of any episodic memory basis at all.

Another line of research further informs about the ways in which social judgments may be influenced by metacognitive processes (i.e. subjective willingness to judge others) without perceivers' awareness of the origin of that influence (Leyens, Yzerbyt, & Schadron, 1992; Yzerbyt, Schadron, Leyens, & Rocher, 1994). In this work Yzerbyt and colleagues examined the conditions under which subjective feelings of confidence propel biased judgments of persons in the absence of awareness of the source of subjective confidence. Similar to the race-crime studies described previously, this work also documents the ease with which metacognitive processes such as feelings of confidence or familiarity can be (falsely) induced and erroneously applied to judgments of individuals.

In a series of studies, Yzerbyt et al. exposed subjects to audio information about an individual member of a known social category (e.g. librarian, comedian). A feeling of confidence and subjective readiness to judge was induced in half of the subjects by misinforming them that they had received diagnostic information about the target in a previous dichotic listening task. The mere suggestion that relevant individuating information had been received was shown to evoke more extreme stereotypical judgments of the target librarian or comedian compared with a control condition. In addition, subjects who received the false familiarity suggestion exhibited greater confidence in the accuracy of their judgments despite their inability to recall specific information that had ostensibly been received. The process described by Yzerbyt and colleagues is similar to that of the race-crime studies in which the baseline condition produced incorrect identifications based on a simple suggestion that there might be some memory for names of criminals.

While few published studies have directly investigated the effect of illusory confidence or willingness to judge on stereotyping (with the exception of Yzerbyt et al., 1994), the findings of several other studies may be understood as being consistent with such an interpretation (Banaji, Hardin, & Rothman, 1993; Beckett & Park, 1995; Bodenhausen & Wyer, 1985; Darley & Gross, 1983; Johnson, Whitestone, Jackson, & Gatto, 1995; Landy & Sigall, 1979; Ugwuegbu, 1979). In all these studies, exposure to non-diagnostic information evoked in perceivers a greater willingness to render a stereotypic judgment. In contrast, experimental conditions in which only social category information was available was not sufficient to evoke the same response. As yet, it is unclear what conditions exactly lead to the increased use of social beliefs in the absence of any additional activation (such as in the race-crime case) versus the conditions that require specific if subtle prior activation to produce stereotyping (e.g. Banaji, Hardin, & Rothman, 1993; Yzerbyt, et al. 1994). In summary, data from several studies when interpreted in terms of the illusory confidence framework suggest that metacognitive decisions about the social judgeability of targets, albeit implicit and perhaps necessarily so, produces increased stereotype usage.

The problem of control

Most central to a cognitive and social view 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 is the assumed control over it. The term "automatic" has come to capture most commonly those psychological processes that operate outside conscious control. In a well-established procedure to measure control, the assumption is a simple one – that the speed of response to a stimulus in the context of another is an indicator of the underlying strength of association (e.g. semantic or evaluative) between the pair. Thus, relatively fast responses are assumed to tap thoughts and feelings that are deployed without conscious deliberation. This assumption has served the field well, and the cooperation of microcomputers has significantly speeded up psychology's understanding of automatic processes. The most common measure of control remains response latency (measured in milliseconds), although other measures such as approach and avoidance techniques involving motor tasks may become tractable measures of automaticity in the future (Chen & Bargh, 1996).

In our program of research, the issue of control has been cast in the form of the automaticity of judgments elicited by social group knowledge. Among the most fundamental of social groups is that of gender. Very early,

children learn to associate attributes differentially with being female and male (Fagot, 1985; Fagot & Leinbach, 1989; Martin & Little, 1990), and we assume that such learning would be shown to occur even earlier than documented if non-verbal measures of such associations were obtained. In our experiments, we have obtained evidence of people's ability to classify gender-related information from a variety of domains into female–male categories. First names are an obvious choice, but so are other attributes such as traits (e.g. emotional, aggressive), occupations (e.g. secretary, mechanic), kinship terms (e.g. aunt, uncle), and verbal and pictorial representations of objects (e.g. skirt, cigar). Using a task routinely employed to study semantic memory, we have shown that feminine primes reliably facilitate judgments of female names and that masculine primes reliably facilitate judgments of male names (Banaji & Hardin, 1996). In other words, the congruence between the gender of prime and target automatically facilitates and intereferes with the judgment.

Having ascertained that this is the case, we sought to show the robustness of this learning by giving participants information that could assist in circumventing the spontaneous behavior pattern (Blair & Banaji, 1996). We created two conditions varying the stimulus onset asynchrony (SOA), such that prime-target pairs appeared in quick succession (350 milliseconds) or were relatively slower (2000 milliseconds). In each condition, half the participants were told to expect either stereotypic or counterstereotypic pairings. When stereotypic pairings between prime and target were expected (i.e. male prime – male target; female prime – female target), the pattern of data was expected to mimic the previously obtained one in the baseline condition of no instruction. The condition of greater interest in understanding the role of control was one in which instructions prepared subjects to expect counterstereotypic pairings and armed them with a strategy to respond more quickly to such pairings than stereotypical ones.

The assumption is that in the counterstereotypic condition, the judgment should be relatively easy when both sufficient resources (e.g. 2000 milliseconds SOA) and a suitable strategy to counteract biases are available to control spontaneous responses to gender-congruent pairings. In contrast, when sufficient resources are not available (e.g. 350 milliseconds SOA) nor an effective strategy easily identifiable, gender knowledge automatically evoked from words (even those whose primary meaning is not gender relevant, e.g. mechanic or sewing), should not allow control over automatic responses to gender-congruent pairings. Results showed support for these predictions, expressed in the form of a four-way interaction between SOA, strategy prime gender, and target gender. These studies have shown that a higher-order goal can be effective but only under conditions that allow control. It is not our understanding that such conditions are a common occurrence in everyday life.

In a more recently developed task, Greenwald, McGhee, and Schwartz (1998) have used a different interference task to examine a similar issue. The procedure, called the Implicit Association Test (IAT) was devised to

measure strength of attitudes through a comparison of theoretically predicted compatible and incompatible responses. Imagine the following experimental scenario. You are asked to classify two types of stimuli on a computer keyboard, using two different keys (A and B) to do so. Let us assume that the categories to be classified were names of *flowers* (daffodil, rose) on key A and *insects* (fly, cockroach) on key B. As you might imagine, the task is an easy one to perform, i.e. producing overall high speed and a low error rate. Suppose that you were then trained to classify a different set of two categories, *positive* (cake, baby) or *negative* (devil, vomit) words. As you might imagine, this task too should be easily performed, again yielding fast response latencies and a low error rate.

Now, suppose that the task were to become more complex, with the judgment requiring a decision about either of the two levels of both classification tasks in a joint task, i.e. the stimulus could be an item from any one of the four categories: Insects, flowers, pleasant words, unpleasant words. Responses to the items however, still use only two keys: Insect names and negative words use key A, whereas flower names and positive words use key B. Now, response time should fall, and error rates should increase. The data of interest are obtained by comparing the latencies on this joint task with performance on the alternative joint task, insects and positive words on key A and flowers and negative words on key B. The first joint task is an evaluatively compatible one (positive words and flowers versus negative words and insects), thus classification latencies are expected to be much faster for this task than the second, evaluatively incompatible task (positive words and insects versus negative words and flowers). The difference in latencies in the compatible and incompatible conditions is taken as a measure of the relative favorability toward flowers compared with insects.

The task is a generic one, with the ability to readily substitute insects and flowers with other categories as Greenwald et al. (1998) did. They found that subjects were faster to classify black and white names when black names were paired via a key to unpleasant words and white names were paired via a key to pleasant words. They also showed that Korean and Japanese subjects showed opposite patterns of implicit attitudes indicating greater ingroup than outgroup liking. The subjective experience when performing the IAT is quite instructive. The compatible condition (black–negative, white–positive) is palpably easier than the incompatible condition, even among those who consciously hold no negative evaluation of black Americans, for the task does not allow control over this implicit negative attitude. As expected, Greenwald et al. (1998) report a lack of correlation between explicit (semantic differential) measures of attitude and the implicit measure of attitude obtained on the IAT. Their data illustrate the failure to exert conscious control over automatic attitudes despite perceivers' awareness of the presence of prejudice in their spontaneous judgments and their conscious disavowal of such prejudice.

Taken together, these studies demonstrate most obviously and strongly the difficulty in curbing the unconscious operation of social beliefs in

judgments. In the automatic gender stereotyping studies, participants were unable to control automatic activation of stereotyping. So also in the IAT studies, the negative attitudes toward social groups were revealed in the inability to control automatically activated preferences.

Higher-level social beliefs (theories about how beliefs ought to operate, how they ought to be controlled or tempered, etc.) can produce control, but only over those expressions that lie more squarely within conscious thought. Such higher-order beliefs, captured on more explicit measures cannot exert control over automatic versions of beliefs toward the same object. Both consciously controlled and relatively automatic beliefs have obvious impact on behavior and influence the shape of interpersonal interaction, but it is unclear at this point how deep and extensive is the contribution of each form of social expression. Our indulgence of implicit processes reflects their relatively dormant status in psychological research and our view that the influence of implicit processes is pervasive and influential.

The problem of intentionality

It is not common for psychologists to dwell on questions of free-will and responsibility for actions. Yet, it is clear that advances in experimental psychology's analyses of unconscious processes must necessarily inform discussions of these matters, traditionally the subject of philosophical, political, and legal debate. We raise some links here, but with great caution, because there is only speculation to offer about these issues that have received little empirical scrutiny. The notion of responsibility for actions is closely tied to the construct of intention, and this, in turn, is closely linked to the constructs of awareness and control that have recently been experimentally studied. If we challenge the long-standing assumption that accurate awareness of the cause of an action or ability to exert conscious control over the action is possible, the notion of intention also becomes suspect. In the data presented earlier in which awareness and control over stereotypes and prejudice are minimal or nonexistent, it is difficult to assume that any conscious intention to misjudge was operative. In other words, conscious intentions cannot be reliable predictors of implicit judgment, feeling, and action.

Those who express no explicit intention to harm, to be prejudiced, or to be unfair in their social judgments may nevertheless cause harm, act prejudicially, and behave in contradiction to their egalitarian beliefs. Such a dissociation between lack of intention to harm on the one hand and discriminatory impact on the other hand has been the topic of much discussion in the law. For the notion of intention, the implications of unconscious processes are deep, although they do not immediately help resolve the questions that arise. We admittedly raise the link between the data we have examined and the legal standing of the notion of intention speculatively. We do so however, in order to imagine the possibility of a future application of scientific evidence about unconscious social judgment for the law.

The notion of intention has been formally recognized in Anglo-American jurisprudence since the time of Edward I (The Statute of Edward, 1325). In general, a prosecutor must be prepared to prove more than the fact that the defendant performed a prohibited act. The assumption in the law is that the act alone is not criminal unless it be accompanied by a specified mental state. The legal maxim, the act is not guilty until the mind is guilty, applies in almost all of criminal law. The doctrine, in its shortened form is referred to as *mens rea*, or the guilty mind, and a similar set of assumptions underlies civil law as well. Yet, legal positions on matters involving intention have been quite inconsistent. At times, employers' hiring practices are judged to be unlawful if they operate to maintain the effect of prior discrimination regardless of their conscious intention to discriminate. A practice was deemed invalid, e.g. Duke Power Co. was held responsible, when it caused disparate impact on a social group (*Griggs v. Duke Power Co.*). Additionally, in *Griggs* the Supreme Court ruled that the burden of proof lay with the employer (the actor/perceiver in our case) to show that its practices were fair and not discriminatory toward members of differing social groups.

Yet, an examination of American legal case history reveals there are far more legal cases on the other side. Not only was *Griggs* itself overturned, in most civil rights cases from the last two decades, the court has held that discriminatory intent must be proven for the act to be considered unlawful. The most striking of these cases is *Washington v. Davis*, involving the use of a test in which white police officers had a success rate that was four times greater than that of black police officers, and the test was not shown to predict on-the-job performance. Here, the court went so far as to say that no intention to harm meant that no injury had even occurred. A tension resides between the notion of discriminatory *intention* versus discriminatory *impact*, i.e. an emphasis on actor intention versus harm to the target. That is the issue on which the court remains inconsistent and divided. And it remains so on ideological grounds rather than as a result of evidence about the extent to which ordinary social agents, both individual and institutional, can produce harm. This is, of course, an old theme in social psychology but one that has acquired new power to inform because of our recent ability now to identify the cognitive and metacognitive mechanisms by which such acts come to be realized (Banaji, Blair, & Glaser, 1997).

The notion of intention, while clearly connected to the concepts of awareness and control (i.e. without awareness and control it is difficult to imagine an intentional act), is also connected to the concept of goals. That is, intentions usually operate in the service of particular goals, and both have been traditionally assumed to be components of conscious thought. However, recent theorizing offered by Bargh and colleagues examines the extent to which goals and motives may be automatically activated (Bargh & Gollwitzer, 1994). These investigators have demonstrated that socio-behavioral goals (such as achievement motivation) can be automatically activated and influence behavior (e.g. produce higher scores on a test). They argue that goals and motives that are consciously versus

unconsciously activated can have equivalent impact (Chartrand & Bargh, in press). To return to the point about social judgments that have discriminatory impact, such studies and their accompanying logic indicate that it may be quite difficult to separate the impact of actions that are caused by conscious intention from "auto-motive" ones. Perhaps a shift in our thinking about intention is in order, moving away from current legal and lay definitions of the term (intentional: done deliberately, *American Heritage Dictionary*, 1992). At the very least, the debate would need to include a discussion of how we are to treat the distinction between intentions and goals that are consciously expressed and expressible and those that are not, especially if data about the influence of unconscious intentions continue to accumulate.

Research on implicit social judgment processes, in particular the data on influences that lie outside conscious awareness, control, and intention, can transform the study of metacognition by bringing into its purview processes and issues that would not otherwise have been encountered. This research emphasizes the importance of studying metacognitive processes in the context of the social world in which they operate and have their influence.

Ackowledgments

This work was supported in part by a grant from National Science Foundation (SBR–9422241). We are grateful to R. Bhaskar for comments on a previous draft.

References

Banaji, M.R., Blair, I.V., & Glaser, J. (1997). Environments and unconscious processes. In R.S. Wyer, Jr (Ed.), *Advances in social cognition, 10*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

Bargh, J.A. & Gollwitzer, P.M. (1994). Environmental control of goal-directed action: Automatic and strategic contingencies between situations and behavior. *Nebraska Symposium on Motivation, 41*, 71–124.

Beckett, N.E. & Park, B.M. (1995). Use of category versus individuating information: Making base rates salient. *Personality and Social Psychology Bulletin, 21*, 21–31.

Blair, I.V. & Banaji, M.R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.

Bodenhausen, G.V. & Wyer, R.S. (1985). Effects of stereotypes in decision making and information-processing strategies. *Journal of Personality and Social Psychology, 48*, 267–282.

Chartrand, T.L. & Bargh, J.A. (in press). Automatic activation of impression formation and

memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology.*

Chen, M. & Bargh, J.A. (1996). An automatic effect of (all) attitudes on behavior: Preconscious approach and avoidance responses to liked and disliked stimuli. Unpublished manuscript. New York University.

Darley, J.M. & Gross, P.H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33.

Glenberg, A.M., Wilkinson, A.C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory and Cognition, 10,* 597–602.

Greenwald, A.G. & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A.G., McGhee, D.E. & Schwartz, J.K. (1998). Measuring individual differences in inplicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology.*

Griggs v. Duke Power Co. (1971). 401 US 424.

Fagot, B.I. (1985). Changes in thinking about early sex role development. *Developmental Review, 5,* 83–98.

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

Higgins, E.T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J.S. Uleman and J.A. Bargh (Eds), *Unintended thought* (pp. 75–123). New York: Guilford Press.

Jacoby, L.L., Kelley, C.M., Brown, J., & 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.

Johnson, M.K. & Raye, C.L. (1981). Reality monitoring. *Psychological Review, 88,* 67–85.

Johnson, J.D., Whitestone, E., Jackson, L.A., & Gatto, L. (1995). Justice is still not colorblind: Differential racial effects of exposure to inadmissible evidence. *Personality and Social Psychology Bulletin, 21,* 893–898.

Johnson-Laird, P.N. (1983). A computational analysis of consciousness. *Cognition and Brain Theory, 6,* 499–508.

Landy, D. & Sigall, H. (1979). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology, 29,* 299–304.

Leyens, J-P., Yzerbyt, V.Y., & Schadron, G. (1992). Stereotypes and social judgeability. In W. Stroebe & M. Hewstone (Eds), *European Review of Social Psychology* (Vol 3, pp. 91–120). Chichester, England: Wiley.

Loftus, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31.

Lombardi, W.J., Higgins, E.T., & 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.

Martin, C.L. & Little, J.K. (1990). The relation of gender understanding to children's sex-typed preferences and gender stereotypes. *Child Development, 61,* 1427–1439.

Metcalfe, J. & Shimamura, A.P. (1994). *Metacognition.* Cambridge, MA: MIT Press.

Nelson, T.O. (1992). Preface. In T.O. Nelson (Ed.), *Metacognition: Core Readings* (pp. ix–xi). Boston: Allyn & Bacon.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Shaw, J.S. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied, 2,* 126–146.

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

Tulving, E. (1994). Foreword. In J. Metcalfe & A.P. Shimamura (Eds), *Metacognition.* (pp vii–x). Cambridge, MA: MIT Press.

Ugwuegbu, D.C.E. (1979). Racial and evidential factors in juror attribution of legal responsibility. *Journal of Experimental Social Psychology, 15,* 133–146.

Washington v. Davis (1976). 426 US 229.

Walsh, W.A., Banaji, M.R., & Greenwald, A.G. (1995). A failure to eliminate race bias in judgments of criminals. Paper presented at the meetings of the American Psychological Society, New York.

Wegener, D.T. & 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.*

Wilson, T.D. & LaFleur, S.J. (1995). Knowing what you'll do: Effects of analyzing reasons on self-prediction. *Journal of Personality and Social Psychology, 68,* 21–35.

Yzerbyt, V.Y., Schadron, G., Leyens, J-P., & Rocher, S. (1994). Social judgeability: The impact of meta-informational cues on the use of stereotypes. *Journal of Personality and Social Psychology, 66,* 48–55.