Implicit Intergroup Bias: Cognitive, Affective, and Motivational Underpinnings

David M. Amodio & Saaid A. Mendoza

New York University

In press, in B. Gawronski and B. K. Payne (Eds.) *Handbook of Implicit Social Cognition*. New York: Guilford.

Correspondence should be addressed to:

David M. Amodio
Department of Psychology
New York University
New York, NY 10003
Phone: (212) 998-3875
Email: david.amodio@nyu.edu

Implicit Intergroup Bias: Cognitive, Affective, and Motivational Underpinnings

Research on implicit race bias has led the surge in implicit social cognition research over the past 20 years, in part because it gives a distinctly social psychological face to an abstract cognitive construct. The domain of intergroup bias provides a unique context for the study of implicit social cognition that emphasizes the roles of cognition, affect, and motivation in coordinating social behavior. Furthermore, it connects the intrapersonal mechanisms of social cognition to dyadic, group, and societal-level processes, thereby linking implicit cognition to social behavior. It is for this reason that studies of implicit race bias have been particularly influential in the development of theory and research in the field of implicit social cognition.

Theories of intergroup relations have also benefited profoundly from implicit social cognition research. Several intergroup phenomena that had previously eluded theoretical explication, such as modern forms of racism, have been largely explained by models of implicit social cognition. In this way, theories and methods of implicit social cognition have contributed to our understanding of how prejudices and stereotypes are represented and expressed in behavior, and how such behaviors are affected by intergroup dynamics. Hence, research on implicit race bias has sustained the interest of the field because, on one hand, it provides critical social context for the study of implicit processes, while on the other hand, it has provided an expanded theoretical approach to social behavior in intergroup relations.

In this chapter, we review major findings and theoretical perspectives in the area of implicit intergroup bias. The structure of this chapter follows from the two major types of questions addressed by research on implicit racial bias: How are implicit biases represented in the mind? And how are implicit biases expressed in behavior? We begin our discussion of these questions with a brief review of the field's theoretical origins and description of some key terms used in the literature. Next, we review major findings in the contemporary literature on implicit

race bias, focusing on how implicit biases are expressed in behavior, and how these expressions may be changed. We then describe two major theoretical approaches to accounting for the phenomenon of implicit racial bias, and conclude with a discussion of some remaining questions and controversies in the field. Our goal is to orient the reader to the basic findings in the literature on implicit race bias, and to provoke thought on the larger theoretical issues and pressing challenges in this area of research. Finally, although we focus primarily on implicit biases regarding African Americans (the main historical target of intergroup discrimination in America), the processes discussed in this chapter refer to general mechanisms of cognition, affect, and motivation, and so the major themes we discuss should apply broadly to implicit cognitive processes concerning other social groups.

Origins of research on implicit race bias

Early interest in implicit racial bias grew out of concerns that self-report questionnaires did not always capture people's true attitudes toward members of racial outgroups. Although the mismatch of word and deed toward a social outgroup is a phenomenon that likely spans the ages, it has gained the attention of social scientists only recently with the emergence of social psychology (Allport, 1954; LaPierre, 1934). An early experimental demonstration of this phenomenon by Rankin and Campbell (1955) showed that, although White participants reported similarly positive attitudes toward the White and Black experimenters in their study, their physiological responses revealed greater autonomic arousal when they were touched by the Black experimenter (ostensibly to check their pulse), compared with the White experimenter. This early report of an implicit racial outgroup bias was followed by a series of studies showing a similar pattern of divergence between implicit and explicit responses (Crosby, Bromely, & Saxe, 1980).

Why did the subjects' self-reported attitudes not match their physiological reaction to race? Some researchers suggested that post-civil rights era norms proscribing prejudice led respondents to conceal their biases (Crosby et al., 1980; Rankin & Campbell, 1955; Sigall & Page, 1971). Others proposed that people were simply unaware of their biases (Devine, 1989). The bottom line was that much of people's intergroup behavior was not accounted for by their self-reported attitudes and beliefs. This discordance between self-reports and behavior raised a number of profound questions for social psychologists and prejudice researchers alike. Were people's "true" racial attitudes residing somewhere in the unconscious, hidden from introspection? To others, it was a slightly different question: To what extent do explicit vs. implicit forms of bias predict behaviors in different situations? At a more practical level, these developments highlighted the need for new methods capable of assessing implicit forms of bias – an endeavor that has had major implications for theoretical developments in this area of research.

Like most great ideas in science, contemporary ideas about automatic and implicit processes emerged in the minds of several different scientists working in different areas of psychology in the 1970s and 1980s. In particular, research on how concepts are learned and stored within semantically-related categories suggested that the categorical processing of social information may operate automatically (e.g., Meyer & Schvaneveldt, 1971, 1976; though noted years earlier by Allport, 1954). Interest in category processing led to methodological innovations such as the sequential semantic priming technique, which allowed scientists to assess the strength of implicit associations without having to rely people's deliberative responses, such as with self-reports (Meyer & Schvaneveldt, 1971; Neely, 1977). In a different literature, memory researchers had discovered dissociations between episodic (explicit) and procedural (implicit) forms of memory (Cohen & Squire, 1980; Graf & Schacter, 1985; Jacoby & Witherspoon, 1982), which suggested dissociable underlying systems for implicit and explicit processes. In yet

another literature, research on human factors examined the degree to which a choice or motor response involved automatic (parallel) vs. controlled (serial) processing (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). Finally, Langer's (1975; Langer & Abelson, 1972) research on the role of "mindlessness" in social behavior demonstrated how automatic responses could be triggered and implemented by situational cues with little conscious intervention. Although these incipient perspectives on implicit cognition (and implicit social cognition) had roots in much earlier theorizing (Allport, 1954; Freud, 1933; James, 1890), they represented a new age of theoretical and methodological sophistication. Together, these converging lines of research set the stage for the emergence of implicit social cognition, which in turn provided the backdrop for modern research on implicit intergroup bias.

Social psychologists applied these early advances in implicit and automatic forms of cognition to questions about person memory, social judgments, and social behavior (Bargh & Pietromonaco, 1982; Smith & Miller, 1979; Srull & Wyer, 1980), including questions about social stereotypes (Gaertner & McLaughlin, 1983; Taylor, Fiske, Etcoff, & Ruderman, 1978). A seminal series of studies by Gaertner and McLaughlin (1983) first demonstrated the implicit priming of racial stereotypes, such that participants categorized African American stereotype words more quickly when they were paired with the group label "NEGRO" than the label "WHITE" (see also Dovidio, Evans, & Tyler, 1986; Perdue, Dovidio, Gurtman, & Tyler, 1990; Perdue & Gurtman, 1990). On the basis of these findings, researchers posited that stereotypic beliefs were represented in the mind in a semantic network. Interestingly, however, the degree of bias on priming tasks was often unrelated to subjects' self-reported racial attitudes and beliefs.

As evidence for implicit racial associations accumulated, researchers puzzled over their theoretical significance and struggled with the fact that implicit assessments were typically not correlated with self-reported attitudes and beliefs. Devine's (1989) landmark paper on the

automatic and controlled components of stereotyping and prejudice provided an important theoretical solution to this puzzle. In it, she proposed that reaction-time assessments reflected automatic processing of passively-learned stereotypic associations, whereas self-report measures typically reflected intentionally-endorsed beliefs. In a set of three studies, Devine (1989) demonstrated that high- and low-prejudice subjects held similar knowledge of African American stereotypes, and that regardless of their explicit beliefs about Blacks' civil rights, subliminal priming of the stereotyped category would cause people to judge new individuals in a stereotype-consistent fashion. However, when subjects were aware that their responses could be influenced by race, they controlled their responses to reflect their explicit beliefs rather than their automatic stereotyping tendencies. That is, low-prejudice subjects chose not to endorse racial stereotypes, whereas high-prejudice subjects did. These findings supported the idea that shared cultural knowledge of stereotypes predisposed all members of a culture to automatic stereotyping tendencies, but that low-prejudice individuals will replace these tendencies with belief-based egalitarian responses when they have sufficient cognitive resources.

With the theoretical scaffolding of Devine's (1989) dissociation model in place, researchers began to develop new methods for assessing one's degree of implicit racial bias (Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, McGhee, & Schwartz, 1998). Much of this work focused on the circumstances in which implicit and explicit measures of racial bias did or did not correspond (Blair, 2001; Nosek et al., 2007). Other research examined the extent to which implicit measures predicted bias in social behavior, such as in anticipated or actual interracial interactions (Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Fazio et al., 1995; McConnell & Leibold, 2001). In general, the domain of intergroup bias has provided a unique context for studying implicit processes because it examines these processes as they relate to social behavior, interpersonal interactions, and group dynamics. Hence, the findings from

implicit race bias research have addressed questions of intergroup bias while advancing our understanding of more general aspects of implicit social cognition.

Definitions and usage

Before proceeding with our review, it is worthwhile to define our terms. In particular, the terms "implicit" and "explicit" have been used to refer to a range of constructs, and they are sometimes confused with the constructs of automaticity and control. Similarly, the term "implicit" is often ascribed to different experimental tasks, yet it is sometimes unclear just how a task might be implicit. To clarify such issues at the outset, we provide our definitions of key terms (although we acknowledge that other researchers may prefer alternative definitions).

Implicit vs. explicit. In line with the literature on learning and memory that forms the foundation of modern implicit social cognition, we use the terms *implicit* and *explicit* to refer to one's level of awareness of a particular psychological process (Jacoby & Witherspoon, 1982; Schacter, 1987; Squire, 1986). That is, an *explicit* process can be consciously detected and reported (regardless of whether it was triggered spontaneously). Any process that is not explicit is referred to as *implicit*. Hence, "implicit" describes a process that cannot be directly inferred through introspective awareness (Greenwald & Banaji, 1995; Wilson, Lindsey, & Schooler, 2000).

Automatic vs. controlled. The terms implicit and explicit are distinguishable from automatic and controlled. In line with classic work on automaticity and control, we define *control* as referring to an intentional regulative process and *automatic* as referring to an unintentional process (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). Controlled processes are typically goal-directed, whereas automatic processes may be triggered spontaneously by external cues (see Bargh, 1994, for a more detailed analysis of automaticity). The regulative nature of control refers to the process of overriding a prepotent tendency or

favoring one particular response over another. Control does not relate to content per se, such as an explicit belief, but rather to the deliberate adjudication of an endorsed response over a different, undesired response. Importantly, the automatic/controlled nature of response is independent of its implicit/explicit nature (although some features of automaticity and implicitness may tend to coincide).

"*Implicit" tasks*. A particular task may be designed to assess the expression of automatic (vs. controlled) or implicit (vs. explicit) processes in behavior. Responses on such tasks constitute an observable behavior from which an implicit or automatic process may be inferred, even if the response itself is explicit or involves control, as in the Implicit Association Test (IAT, Greenwald et al., 1998). Because a behavioral response reflects a combination of automatic and controlled processes, and because the response can usually be perceived explicitly, it would be inaccurate to describe any particular task or behavior as "implicit" (Jacoby, 1991; Payne, 2001). Rather, it is the influence of an underlying association on behavior that may be implicit, and this influence is the critical inference made from such task responses (Payne, 2008). This issue will come up later in this chapter as we discuss interpretations of behavioral tasks used to infer implicit forms of racial bias.

In this chapter, we will use the colloquialism of "implicit task" or "implicit measure" to describe a behavior-based procedure for inferring a pattern of implicitly-biased behavior. When changes in performance are observed, it is important to refer to it as a change in the "expression" of implicit bias rather than a change in an underlying bias per se, given that a change in behavior may or may not reflect a change in underlying mental structures. For this reason, it is difficult to evaluate claims of malleability versus change in responses on implicit tasks.

Implications of usage. At a broader level of analysis, the distinction between implicit/explicit and automatic/controlled processes has important implications for the

psychological questions under investigation. The terms *implicit* and *explicit* describe the property of awareness, and thus these terms are particularly relevant to questions about attribution, mental representation, self-reflection, and person perception, but not as relevant to issues concerning action. By contrast, the terms *automatic* and *controlled* describe a property of an action, which has particular relevance to questions about goals, motivation, and behavior, but with less direct relevance to mental representation and person perception. Indeed, a difference in emphasis can be seen in the research literature, where some research is focused on indentifying and characterizing the mental representation of implicit bias (Gawronski & Bodenhausen, 2006; Sherman, 1996), and other research focuses on the role of implicit bias in behavior (Amodio & Devine, 2006; Dovidio, Gaertner, & Kawakami, 2002; Payne, 2005). Thus, precision in the use of these terms is necessary because they refer to different psychological questions.

The phenomenon of implicit race bias

The seminal work of Gaertner and McLaughlin (1983) and Devine (1989) prompted an explosion of studies on the basic phenomenon of implicit race bias. Much of this work has been descriptive. That is, the idea that people could possess unconscious intergroup biases was novel and fascinating, and as a result, much attention turned to documenting this phenomenon using an array of "implicit" tasks (Blair, 2001). Throughout this work, the chief defining characteristic of implicit racial biases was a dissociation with explicit measures of racial attitudes and beliefs (e.g., Devine, 1989; Gaertner & McLaughlin, 1983; see also Greenwald & Banaji, 1995, Wilson, Lindsey, & Schooler, 2000). In this section, we provide a selective review of the major types of implicit bias phenomenon that have been studied in the literature.

Implicit stereotyping

Initial studies of implicit bias examined racial stereotypes, inspired by questions about the changing nature of stereotypes over time (Karlins, Coffman, & Walters, 1969). In the first

demonstration of implicit stereotyping, described above, Gaertner and McLaughlin (1983) found that African American stereotypic words were categorized more quickly when primed by labels of the social group. They interpreted this effect as evidence that the prime and target words were included within a common semantic network, and used the degree of stereotype-consistent response facilitation to estimate a particular subjects' degree of bias.

As personal computers became more common in the laboratory, researchers increasingly used sequentially-primed lexical decision tasks, in which a prime word quickly preceded the presentation of the target word on the computer screen, and responses were made on the computer keyboard (e.g., Macrae, Bodenhausen, & Milne, 1995; Macrae, Stangor, & Milne, 1994; Spencer, Fein, Wolf, Fong, & Dunn, 1998; Wittenbrink, Judd, & Park, 1997). For example, Wittenbrink et al. (1997) used a primed lexical decision task to examine positive and negative stereotypes of Black and White Americans (see Wentura & Degner, this volume). As with Gaertner and McLaughlin (1983), the logic was that if the prime and target were represented in the same mental category, activation of the prime should enhance accessibility of the target, thereby speeding one's lexical judgment. The authors found that the Black prime significantly speeded the categorization of negative African American stereotype words relative to all other targets, whereas the White prime speeded categorization of White positive stereotype words. An advantage of the lexical decision paradigm is that it appears to provide a relatively straightforward assessment of the strength of semantic associations.

Dovidio and his colleagues (e.g., Dovidio et al., 1986; 1997) took a slightly different approach to assessing stereotype associations. In the general version of their paradigm, primes consisting of White or Black faces or group labels are presented very quickly, and then replaced by a target stimulus. Target stimuli consist of trait adjectives that could apply to either a person or a non-social object (e.g., a house), and subjects are told to categorize each target adjective

according to whether it "could ever be true" or "is never true" of people (or of houses, in other blocks of trials). This task is notable because the instructions place subjects in the mindset of making social judgments, which may be more in line with real-life social situations than the relatively decontextualized word/nonword judgments made in basic lexical decision paradigms.

Several other variations of the semantic priming paradigm have been used to assess implicit stereotypes. Examples have included a primed word-pronunciation task (Kawakami, Dion, & Dovidio, 1998); primed word fragment completion (Gilbert & Hixon, 1991; Spencer et al., 1998); stereotype-naming Stroop task (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000); and the IAT (see below for more detail; Amodio & Devine, 2006; Rudman, Greenwald, & McGhee, 2001). A related set of paradigms have examined "weapons bias," whereby a White American subject is quicker to correctly identify a gun and more likely to misidentify a hand tool as a gun when primed by a Black face than a White face and (Lambert et al., 2003; Payne, 2001, 2005; Payne, Lambert, & Jacoby, 2002). A variation on the weapons identification task is the *Shooter Task*, in which subjects must quickly choose to "shoot" or "not shoot" male targets who are holding guns or innocuous objects (Correll, Park, Judd, & Wittenbrink, 2002, 2007). White and Black Americans alike tend to show a similar pattern of bias, such that they are quicker to "shoot" armed Blacks than Whites, and more likely to erroneously shoot unarmed Blacks than unarmed Whites.

The basic logic behind the range of implicit stereotyping tasks is the same, in that they assume that the racial prime activates elements of the stereotype in one's mind, and that the heightened accessibility of the stereotype facilitates the processing of a stereotype-related target (while inhibiting the processing of stereotype-incongruent targets). Across studies and paradigms of American research subjects, a relatively consistent pattern of findings has demonstrated an association of Black people with negative African American stereotypes. This association is

considered to be implicit because responses on the task are either too fast for conscious deliberation or, in some cases, the group prime is presented so quickly that it cannot be consciously perceived. Stereotype-facilitated responses in these studies may also be considered automatic because they are initiated without awareness or intent, and they appear to operate in the absence of intentional control. Finally, implicit responses tend to be uncorrelated with explicit racial attitudes and endorsed racial stereotypes, yet they are sometimes associated with knowledge of stereotypic beliefs held by one's society (Correll et al., 2002; Devine, 1989).

Implicit evaluative bias

Whereas implicit stereotyping research emerged from the traditional literature on intergroup stereotyping and prejudice, interest in implicit racial evaluations emerged primarily from the attitudes literature in social psychology. According to the traditional tripartite model of attitudes, an attitude (or *evaluation*) is a favorable/unfavorable assessment of an object that reflects cognitive, affective, and behavioral processes (Eagly & Chaiken, 1998). Importantly, the cognitive component may refer to a semantic association between the object and the concept of good (much like a stereotypic association), whereas the affective component refers to the aroused affective response associated with the object. It is notable that in social psychology, attitudes research has focused primarily on the cognitive component of attitudes and evaluations, in both its theoretical models and its measures (Breckler, 1984). This is especially true in the implicit social cognition literature, in which measures of implicit attitudes typically rely on semantic judgments, with little attention given to the measurement of high-arousal affective responses. For this reason, our review of implicit racial evaluation focuses on measures that appear to tap into the cognitive (or semantic) component of an attitude. Implicit *affective* forms of racial bias are then addressed in the following section.

According to representational accounts, an implicit racial evaluation reflects a semantic association between an attitude object (e.g., a member of a racial group) and general concepts of good vs. bad (Fazio, 2007), or, alternatively, the net valence of semantic associations with the attitude object (Gawronski & Bodenhausen, 2006). In both cases, the primed activation of an attitude object should increase the accessibility of associated good/bad concepts. Building on the principle of evaluative networks, Fazio and his colleagues developed a sequential priming technique to measure the degree to which an attitude object facilitates responses to positive vs. negative words (Fazio, Chen, McDonel, & Sherman, 1982; Fazio, Powell, & Herr, 1983; see Wentura & Degner, this volume). Faster categorizations of positive words compared with negative words following the presentation of the attitude object would suggest an implicit positive evaluation, or attitude. It is notable that alternative theoretical accounts have been proposed to explain evaluative priming effects on such measures (for a review, see Klauer & Musch, 2003). The representational account is that the prime increases accessibility of the target, via a semantic network, which speeds the mental processing of the target. By this account, priming of the negative attitude object "spider" would raise the accessibility of all negative attitude objects in one's mind, making it easier to then process a negative target word than a positive word. An alternative explanation is that the prime activates a valence-congruent response, which is in line with a valence-consistent target word but inconsistent with a valence inconsistent word. By this account, priming of the word "spider" would set a negative categorization response in motion. The categorization of a negative target word would be facilitated because the congruent response was already activated. The difference between these accounts concerns whether priming effects occur at the level of mental representations or actions.

To measure implicit responses to racial groups as the attitude objects, Fazio et al. (1995) designed a computerized priming task in which Black or White faces were presented as primes

for 315 ms, followed by a black screen (135 ms), and then either a positive or negative adjectives were presented as the target. Subjects were instructed to categorize target words as *good* or *bad* as quickly as possible, via button press. Responses on this task were considered to be implicit because the short stimulus onset asynchrony (450 ms) made it difficult to deliberate on the association between the prime and target. Fazio et al. (1995) found a pattern of race-biased responses among both White and Black subjects in their studies. White subjects responded most quickly to positive adjectives following White face primes, showing an implicit pro-ingroup bias. Black subjects responded most quickly to negative targets following White face primes, showing an implicit anti-outgroup bias. Importantly, among White subjects, the magnitude of bias was uncorrelated with responses on the Modern Racism Scale (MRS; McConahay, 1986), an explicit measure of prejudiced beliefs. However, Fazio et al. (1995, Study 4) found that among participants reporting low motivation to control prejudice, stronger implicit bias was correlated with more prejudiced racial attitudes.

Since its introduction, the IAT has become a very popular method for assessing implicit evaluations (Greenwald et al., 1998). The IAT is a dual categorization task in which participants categorize words as pleasant or unpleasant, and faces as either Black or White, by pressing one of two keys on the computer keyboard (see Teige-Mocigemba, Klauer, & Sherman, this volume). On "bias-compatible" blocks of the IAT, participants must classify White faces and positive words with one response key, and Black faces and negative words with the other. A person with a strong anti-Black or pro-White bias should find these trials easy and perform them quickly. On "bias-incompatible" blocks, these pairings are reversed, such that White faces and negative words are classified with one key, and Black faces and positive words are classified with the other. A person with an anti-Black or pro-White bias should find these trials to be difficult and perform them more slowly. Evaluative bias is characterized by faster responses on compatible

blocks than incompatible blocks. The "IAT effect" – the difference in response latencies for incompatible minus compatible blocks – reflects two processes: (a) the ease with which bias-consistent responses are made (i.e., the strength of an automatic association) and (b) the difficulty with which a bias-inconsistent response is made (i.e., the extent to which controlled processing is needed). Thus, the IAT effect represents a combination of automatic and controlled processing (see also Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005). However, because the source of automaticity and the need for control is presumably unconscious and unintentional, the IAT effect is considered to represent an implicit bias.

Payne, Cheng, Govorun, and Stewart's (2005) Affect Misattribution Procedure (AMP) assesses implicit evaluative bias through self-reported judgments, in contrast to the more typical reaction-time based procedures. In the AMP, subjects view a prime picture of the attitude object (e.g., a Black vs. White face), which they are typically instructed to ignore. Next, an unfamiliar target picture (e.g., a Chinese pictograph) is presented. Participants must then evaluate the target picture as pleasant or unpleasant, in a forced dichotomous choice. Payne et al. (2005) observed that, across trials, target pictures were evaluated more negatively following Black face primes than White face primes. The AMP is unique because it assesses implicit evaluative bias using a self-report format, which lends itself to much higher inter-item reliability scores than reaction-time assessments. The task may be considered "implicit" because subjects are unaware of exactly how their response to the prime might influence their evaluation of the target. As such, Payne et al. (2005) have used the AMP to underscore the theoretical point that "implicit" refers to awareness of how a bias influences a response, rather than to the experience of bias or to the response itself.

It is notable that the measures of evaluative bias reviewed here are sometimes described as assessing "affect." However, it remains whether such measures are able to pick up on the

affective component of an evaluation, which is typically marked by some degree of autonomic arousal, in addition to cognitive aspects of an evaluation. Additional research will be needed to determine the extent to which such measures of evaluative race bias are driven by aroused affective reactions or cognitive associations pertaining to emotional appraisals. This distinction becomes critical when considering the underlying neurocognitive mechanisms of bias measured by the task, described in the *Memory Systems Model of Implicit Bias* section below.

Implicit affective bias. Just as research on explicit intergroup bias suggests a distinction between cognitive and affective forms of bias (e.g., Dovidio, Brigham, Johnson, & Gaertner, 1996; Judd & Park, 1993), researchers have attempted to distinguish between semantic (or conceptual) and affective forms of implicit bias (Amodio & Devine, 2006; Wittenbrink, Judd, & Park, 1997, 2001). However, affect has been a difficult construct to capture in modern social cognition research, particularly when responses are measured using self-report or behavioral assessments involved semantic judgments (Breckler, 1984). That is, cognitive and affective processes typically operate in concert, and the degree to which each contributes to a response is very difficult to determine. Yet, as noted above, a key defining characteristic of an affective response is autonomic arousal. To the extent that word categorizations on a priming task occur with little arousal, it is difficult to interpret them as "affective." What, then, is the role of affect in implicit bias?

In light of these issues, Amodio, Harmon-Jones, and Devine (2003) sought to examine affective processes associated with implicit race bias in a way that could be dissociated from semantically-driven evaluations. Recent advances in the neuroscience literature suggested that subcortical brain structures such as the amygdala were specifically involved in affective responses to threatening stimuli. Importantly, the brain regions involved in this type of affective response were different from those involved in semantic processing. Amodio et al. (2003)

proposed that a neuroscience approach could be used to identify affective forms of implicit bias that were independent of semantically-driven evaluative and stereotypic associations. To this end, they used an emotion-modulated startle-eyeblink assessment of amygdala activity (Lang, Bradley, & Cuthbert, 1990).

When a person is startled (e.g., by a loud noise), they show a whole-body startle reflex. One component of this reflex is the defensive eyeblink. This blink response is larger when a person is in an aversive state just prior to being startled, but smaller when in an appetitive state just prior to being startled – an effect mediated by amygdala inputs to the reflexive blink circuit (Davis, 1992). Thus, a magnified blink reflects an aroused *aversive* response (and greater amygdala activity) to a stimulus preceding the startling event, whereas an attenuated blink reflects an aroused *appetitive* state (and lower amygdala activity). Amodio et al. (2003) chose to use the startle eyeblink measure because it could assess changes in amygdala activity within a few hundred milliseconds after the presentations of an ingroup vs. outgroup face. By comparison, fMRI methods at that time could only measure slow shifts in brain activity across long blocks of trials (e.g., Phelps et al., 2000). Furthermore, the startle eyeblink method measures amygdala activity associated specifically with an aroused affective state, given that the startle reflex is modulated via the central nucleus of the amygdala, which activates autonomic responses (LeDoux, 2000). By contrast, current fMRI methods cannot distinguish between the activity of amygdala subnuclei, and thus cannot clearly assess a response related to aroused affect.

Amodio et al. (2003) observed larger startle eyeblink amplitudes to Black vs. White faces, indicating a negative affective response to Blacks among White participants, on average. The degree of affective bias was unrelated to self-reported racial attitudes (assessed by the Attitudes Toward Blacks scale, Brigham, 1993). These findings provided evidence of a rapidly-

activated and implicit form of affective bias. This general pattern of biased amygdala activity toward outgroups has been conceptually replicated in several studies (e.g., Cunningham et al., 2004; Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Wheeler & Fiske, 2005). However, as noted above, it is unclear whether fMRI assessments of amygdala activity passive face-viewing tasks can probe affective responses as effectively as the startle eyeblink procedure.

Although neural assessments allow researchers to probe the psychological mechanisms of affective bias, several researchers have used peripheral physiology measures to index intergroup affect (see Guglielmi, 1999, for a review). Following the tradition of Rankin and Campbell (1955), Vanman, Paul, Ito, and Miller (1997) used electromyography (EMG) to measure subtle changes in facial muscles associated with frowning and smiling at ingroup vs. outgroup faces. Although White participants who reported either high or low levels of prejudice on the MRS provided equally high ratings of perceived friendliness for White and Black people in the pictures, facial EMG measures revealed more negativity toward Black faces among the high-prejudice participants (Study 3). Mendes, Blascovich, and their colleagues have measured patterns of cardiovascular responding in intergroup interactions, and have observed greater threat-related patterns of activity toward outgroup members that may be characteristic of implicit affective responses (Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell, 2001; Mendes, Blascovich, Lickel, & Hunter, 2002). Finally, it is notable that research using event-related potential methods suggest that racial and gender categorizations may be made in as little as 200 ms following face presentation (Ito & Urland, 2003), but it is unclear whether this effect reflects an affective or semantic process. Interest in the affective component of implicit race bias has grown in recent years, and we expect to see major advances in the future as researchers develop new methods for assessing affective responses.

Effects of implicit bias on behavior

A large accumulation of research findings attest to the existence of an implicit racial bias. But does the phenomenon of implicit race bias have any real significance for social behavior? One can argue that implicit bias is only a problem to the extent that it influences behavior and leads to discrimination. Whereas most research has focused on documenting intrapersonal forms of implicit bias and exploring the conditions under which it does or does not correspond with explicit measures, attention has increasingly turned toward understanding how such biases may be expressed in behavior (Dasgupta, 2004).

In early studies of implicit bias, the focus was on behavioral expressions. For example, Devine (1989) showed that stereotypes, when implicitly activated, could color judgments of a race-unspecified target person. Fazio et al. (1995) went a step further by examining White subjects' behavior toward Black female experimenter. Subjects with stronger evaluative bias on the computerized priming task showed more uncomfortable nonverbal behaviors during the interaction. However, neither implicit bias nor nonverbal discomfort was associated with explicit racial beliefs or judgments. Similar results were obtained by Dovidio et al. (1997; 2002), who showed that a subliminally-primed measure of implicit evaluative bias predicted more anxious and less friendly nonverbal behaviors during an interracial interaction, but that these responses were unrelated to explicit racial attitudes. Other studies have shown that implicit evaluative bias predicts greater personal distance from an outgroup member (Amodio & Devine, 2006; McConnell & Liebold, 2001) and that greater implicit stereotyping is associated with a reluctance to engage with an outgroup member in an interaction (Sekaquaptewa, Espinoza, Thompson, Vargas, & von Hippel, 2003).

Expressions of bias: Hostility or anxiety?

Implicit race bias is often thought of as the nonconscious analog of overt antipathy, and therefore one might expect implicit bias to be expressed in hostile acts toward outgroup

members. However, studies of implicit bias effects on behavior have not shown evidence for the antipathy hypothesis. Rather, implicit evaluative bias tends to be expressed as anxiety and discomfort (Fazio et al., 1995; Dovidio et al., 1997, 2002; Trawalter & Shapiro, this volume). More recent work suggests that this discomfort is often interpreted as unfriendliness by one's interaction partner, which may then perpetuate into the reciprocation of hostility (Pearson et al., 2008; West, Shelton, & Trail, in press). Other research suggests that when individuals with high levels of implicit evaluative bias become aware that they possess a negative outgroup bias, they tend to exert stronger regulatory efforts to counteract any implicit biases, acting with greater care and increased friendliness (Monteith, Voils, & Ashburn-Nardo, 2001; Shelton, Richeson, Salvatore, & Trawalter, 2005). Hence, the way in which implicit evaluations and stereotypes are expressed in behavior is often complex and therefore very difficult to study.

When considering the expression of implicit bias as discomfort vs. antipathy, it is useful to consider that negative implicit associations with racial outgroups could reflect several different types of reactions (see Olson & Fazio, 2004). For example, an outgroup face may be a source of anxiety to a research participant, rather than a target of antipathy. This anxiety could stem from perceptions of threat from outgroup members or from the concern of appearing racist on the task (as in Frantz, Cuddy, Burnett, Ray, & Hart, 2004). Outgroup faces may even automatically trigger egalitarian responses, such as sympathy, yet still produce a negative bias due to the oppression and maltreatment that is associated with low status groups (Uhlmann, Brescoll, & Paluck, 2006). Indeed, most research subjects are university students who tend to hold progressive egalitarian values. For these subjects, then, implicit bias stemming from any source (threat, anxiety, or sympathy) should correspond to uncomfortable feelings during the interaction. In this case, measures of implicit bias would predict discomfort, inhibition, and

avoidance behavior rather than hostility. More research is needed to determine the situations in which implicit bias may be expressed as discomfort vs. hostility.

Expressions of implicit stereotyping vs. implicit evaluative bias

Whereas past research has dissociated the effects of implicit and explicit forms of race bias on different types of behaviors, Amodio and Devine (2006) examined the differential effects of evaluative vs. stereotyping forms of implicit bias on behavior. On the basis of neuroscience models of learning and memory, they proposed that implicit evaluative bias was largely driven by affective systems, which are expressed through basic-level behavioral channels such as nonverbal behaviors and anxiety-related responses. By contrast, they proposed that implicit stereotypes are driven by semantic memory systems, which are expressed primarily in higher-level judgments and goals, such as trait impressions and plans for interacting with an outgroup member. In their studies, White subjects completed measures assessing implicit evaluative associations (pleasant/unpleasant associations unrelated to stereotype content) or stereotypic associations (in which evaluative content was controlled). Indeed, these measures of implicit evaluative bias and implicit stereotyping were independent. More importantly, they were uniquely predictive of these different classes of behavior toward a Black student. For example, more negative implicit evaluation scores predicted further seating distance from a Black study partner, whereas implicit stereotyping predicted subjects' expectations that their Black partner would succeed on measures of academic ability (vs. non-academic abilities). Amodio and Devine (2006) suggested that a consideration of the distinct affective and semantic systems underlying different forms of implicit bias would permit a more refined model of how implicit biases may be expressed in behavior.

Understanding how implicit biases are expressed in behavior is arguably the most important question in implicit race bias research today. Although this topic has received

disproportionately little attention in the past (in part because of the challenges associated studying real intergroup social behavior), researchers are increasingly focusing on this issue. In the end, theories of how racial biases are represented inside the head matter only to the extent that they influence behavior (Amodio & Devine, 2005). Therefore, a better understanding of how implicit bias is expressed in social behavior will be critical for validating the theoretical models of implicit intergroup bias that are dominant in the extant literature.

Moderators of implicit bias

A major goal of intergroup bias researchers is to develop methods for reducing prejudice. The discovery of implicit forms of racial bias raised a new and formidable challenge to this goal – the automaticity of implicit bias seemed to imply that its application was inevitable. Indeed, some theorists opined provocatively that resistance to implicit racial biases was futile; that such biases were a necessary consequence of the mind's reliance on categorical processing to deal with the overwhelming complexities of the social world (Bargh, 1999). But other researchers pointed to humans' profound capacity for self-regulation (Devine & Monteith, 1999), and emerging research on the malleability of implicit task responses suggested that implicit race bias could indeed be moderated by a range of personal and situational factors (e.g., Dasgupta & Greenwald, 2001; Rudman et al., 2001; Lowery, Hardin, & Sinclair, 2001; for review, see Blair, 2002). These initial findings of implicit bias malleability served as a call to arms for intergroup bias researchers interested in reducing expressions of prejudice and stereotyping.

Here, we provide a brief review of the theory and methods pertaining to changes in implicit bias. The literature on implicit bias malleability is complex, with several different methodological approaches and theoretical explanations. At the level of measurement, changes in implicit bias are (virtually) always indicated by a change in behavioral responses on an implicit bias task. Thus, at a descriptive level of analysis, the evidence for change is always seen

in the expression of a behavior. Theoretically, a change in behavior may be due to several different processes. For this reason, our discussion of change in implicit bias considers research on a range of underlying processes. In our discussion, we note how particular demonstrations of implicit bias change may be interpreted as evidence for a variety of mechanisms, even though an author's preferred interpretation may favor one specific mechanism. In this way, we illustrate the complexity of psychological processes that may underlie a change in observable task behavior.

Changes in representations. The Holy Grail of implicit race bias research is to change the underlying associations that form the basis of implicit bias. Change in performance on implicit bias tasks is sometimes interpreted as a change in the underlying representation of racial associations. However, this interpretation is difficult, if not impossible, to test directly using behavioral or physiological measures, and therefore such explanations remain hypothetical. For example, Olson and Fazio (2006) had subjects view pairings of Black faces with positive images, and White faces with negative images. After this training, subjects were quicker to identify negative words primed by White faces, which effectively reduced the effect of race on task performance. Did this task change subjects' representations of White people? Or did it train them to expect a negative image whenever a White face was primed?

In another line of research, Kawakami, Phills, Steele, and Dovidio (2007) trained a subset of subjects to move a joystick in an "approach" direction when they saw a Black face. Subjects in this condition exhibited less bias in their later performance on a behavioral measure of implicit bias, compared with those who did not engage in approach training. What explains the change? Did approach training change the underlying representation? Did it train subjects to adopt an approach motivation when they saw a Black face? Did it create a cue to engage greater control when a Black face was encountered? Or create a situational cue that Black people are

approachable and thus safe? As discussed by Kawakami et al. (2007), the exact mechanism underlying the change in performance is difficult to specify.

An elegant set of studies by Rydell and McConnell (2006; McConnell, Rydell, Strain, & Mackie, 2008) demonstrated a dissociation between the acquisition and change of implicit vs. explicit attitudes. On the basis of dual-processes models positing that implicit systems change slowly whereas explicit systems change quickly (Sloman, 1996; Smith & DeCoster, 2000), they predicted that implicit biases would change after repeated trials, whereas explicit biases would change after a single instance of new counter-attitudinal information. Indeed, this is what was observed across several studies. This research elucidated the distinct processing dynamics of implicit vs. explicit systems. However, the mechanism underlying the observed change in implicit responses remains difficult to determine. Did it involve a change in representation? A change in accessibility? Implicit goal activation (e.g., Bargh, Gollwitzer, Lee-Chai, Barndollar, & Troetschel, 2001)? Although the effects observed in these studies may be interpreted as changes in underlying representations, it is difficult to rule out other explanations when behavioral assessments of implicit bias are used.

Goal effects. The goal to engage in a positive interaction can have a major influence on the expression of implicit bias (Lowery et al., 2001; Richeson & Shelton, 2003; Shelton et al., 2005). Exposure to positive exemplars of a stigmatized outgroup can also motivate a respondent to view members of that group in a more positive light, thereby reducing the expression of bias (Dasgupta & Greenwald, 2001; Govan & Williams, 2004). Exposure to egalitarian messages may activate prosocial goals in the context of an intergroup interaction (Sinclair, Lowery, Hardin, & Colangelo, 2005). The goal to perceive a person according to their race vs. their gender has also been shown to moderate whether race- or gender-based stereotypes are applied to trait judgments and behaviors (Macrae, Bodenhausen, & Milne, 1995; Pratto & Bargh, 1991).

Perspective taking also constitutes a goal process, whereby a perceiver is motivated to empathize with a stigmatized social group member. Following this logic, Galinsky and Moskowitz (2000) showed that perspective taking can also reduce the expression of implicit bias.

Goal strategies may be used explicitly to focus an individual on situational cues or critical aspects of the task at hand, which serves to reduce the influence of extraneous factors, such as race, on one's behavior. For example, Mendoza, Gollwitzer, and Amodio (2009) used *implementation intentions* – specific if-then plans that link a situational cue to a specific action – to enhance subjects' accuracy when performing an implicit stereotyping task. By giving subjects a strategy that increased performance accuracy and filtered out the influence of race, the implementation intentions effectively reduced the expression of implicit race bias. Similarly, Stewart and Payne (2008) gave subjects if-then plans to think counter-stereotypical thoughts, which interrupted the influence of implicit racial biases on task performance. Hence, strategies that promote goal-directed action may shield an individual from the influence of race and limit the effect of implicit racial biases on task performance.

Situational effects. Elements of a situation can activate thoughts, emotions, or goals that moderate perceptions of and responses to outgroup members. Several studies have shown that viewing a Black man in the context of a dark alley elicits more biased responses than a church context (Barden, Maddux, Petty, & Brewer, 2004; Wittenbrink et al., 2001). Interacting with a positive exemplar of a stigmatized outgroup in a safe setting (e.g., a classroom) has also been shown to lead to reduced expressions of negative racial evaluations (Lowery et al., 2001). However, it remains unclear whether situational moderators alter expressions of bias by changing the way an individual perceives race-related stimuli, by changing the activated representations of a racial outgroup, by activating an alternative response goal, or by cuing a

more controlled mode of response. Most likely, the effects are driven by a combination of these processes.

Controlled processing. Performance on implicit tasks is driven by a combination of automatic and controlled processes (Amodio, 2008; Payne, 2001, 2005; Sherman et al., 2008). Indeed, simply following task instructions to categorize a word or complete a word fragment requires a high degree of control. Furthermore, research using event-related potentials to assess control-related brain activity has shown that controlled processing can be triggered implicitly when racial-concepts are activated in an unfolding response (Amodio et al., 2004; Amodio, Devine, & Harmon-Jones, 2008). Thus, control need not be deliberative, and therefore it is difficult to determine when changes in performance on an implicit task are due to spontaneously-engaged control or some other hypothesized process, such as a change in underlying representations (Amodio et al., 2008; Payne, 2005). When racial issues are made salient in an initial task, subjects may become more vigilant to cues that indicate the need for more careful and controlled responding (Amodio, Harmon-Jones, & Devine, 2007; Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002). That is, highly-controlled performance on an implicit task increases task accuracy, which may thus preclude biases from emerging in behavior. Sensitivity to cues for control may be long-lasting, and thus they may constitute a form of sustained implicit bias change.

Individual differences. Although similar patterns of implicit race bias are usually observed across members of a culture, some research has identified consistent individual differences. These include internal and external motivations to respond without prejudice (Amodio et al., 2003; Devine et al., 2002) and chronic egalitarianism (Moskowitz, Gollwitzer, Wasel, & Schaal, 1999; Moskowitz, Salomon, & Taylor, 2000). Yet again, it remains unclear exactly why some individuals show less implicit race bias on behavioral and physiological

measures than others. Do they lack biased associations in their mental representations (Devine et al., 2002)? Are they more resistant to forming biased associations in the first place (Livingston & Drwecki, 2007)? Are they more sensitive to cues for responding without bias, and thus more adept at control (Amodio et al., 2008; Monteith et al., 2002; Moskowitz et al., 1999)? Again, our understanding of the mechanisms underlying these effects is limited by our methodological reliance on behavioral expressions and often ambiguous physiological measures.

Evaluating studies of implicit bias malleability. As should be evident from our discussion, it is exceedingly difficult to make strong inferences about the cause of an observed change in behavioral performance on an implicit measure of race bias. That is, the necessary reliance on behavior is a major limiting factor, without a clear solution. As a result of this limitation, inferences about the mechanisms underlying changes in implicit task responses are often ambiguous. However, physiological or neuroimaging measures may be used in conjunction with behavioral assessments of implicit bias to provide some insight into the possible mechanisms. Neuroimaging methods, such as event-related potentials and functional magnetic resonance imaging, offer clues about the involvement of neural systems associated with general forms of controlled processing, attention, and affect. But processing distinctions that are central to sociocognitive theories, such as between representations, accessibility, and associative conceptual links, relate to complex patterns of brain activity that cannot be directly inferred using neuroimaging measures (at least not at the present time). Given the limitations in assessing changes in implicit bias described here, it may be useful to remain open to alternative mechanisms, and to focus interpretations on expressions of bias rather than on presumed underlying changes that may ultimately be untestable.

Theoretical accounts of implicit racial bias

Implicit processes are like the *dark matter* of social cognition. We have strong reason to believe they exist, given that so much of our behavior is unexplained by explicit beliefs and intentions. But because implicit processes are defined by the absence of awareness, they excel at eluding concrete description (Fazio & Olson, 2003). Without a concrete description of an implicit process, it is difficult to build a cogent explanatory model. It is notable that several theoretical accounts have been proposed to explain the operation of particular tasks designed to assess implicit bias (Brendl, Markman, & Messner, 2001; Conrey et al., 2005; Gawronski, LeBel, & Peters, 2007; Greenwald et al., 2002; Karpinski & Hilton, 2001). By contrast, few theoretical models have been articulated to delineate the specific psychological mechanisms that comprise an implicit process, beyond the basic notion that it reflects an association stored in memory (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Wilson et al., 2000). Importantly, for the present purposes, there are no models of implicit bias that pertain uniquely to racial bias. That is, most researchers assume that implicit racial bias is a specific case of a general implicit process, and therefore general models of implicit social cognition are applied. In this section, we present two general theoretical approaches for implicit social cognition that have been influential to models of implicit race bias.

Representational approaches

Research on implicit race bias originated from cognitive theories of mental representation, and as such, these theories reflect the dominant model of implicit bias. Representational models address the question of how information about social groups is stored and activated in the mind, and how it contributes to the mental processes of person perception and attribution. Inspired by computer-based models of the mind, representational models assume that information is stored in a network of concepts, as in associational models, or a network of smaller informational units that underlie the representation of concepts, as in connectionist

models (e.g., Sherman, 1996; Smith & Branscombe, 1987; Smith & DeCoster, 2000). For example, implicit stereotypes may be represented in an associational network of attributes related to the concept of "African American" (Figure 1). Different connections may have different weights, which determine the degree to which the activation of one concept activates others (Bodenhausen & Macrae, 1998; Fazio, 1990; Gawronski & Bodenhausen, 2006; Smith & DeCoster, 2000).

Associative models of implicit bias assume that components of the network may represent an evaluation (e.g., "good" or "bad"), a trait attribute (e.g., "lazy" or "intelligent"), or, according to some models, a generalized affective disposition such as a positive or negative feeling (Gawronski & Bodenhausen, 2006). Accordingly, implicit racial attitudes are represented by the relative strength of association between a racial group and positive vs. negative concepts in the informational network. Stereotypes, by comparison, refer to the specific set of trait attributes that are linked to a particular social group (Park & Judd, 2005). Some models differ on the specific point of whether implicit racial attitudes per se are represented in a semantic network (e.g., Fazio et al., 1995). Others posit that the network represents implicit affective and semantic associations, but that the evaluation, or attitude, is propositional in nature and more likely to operate in explicit processes (Gawronski & Bodenhausen, 2006; Gawronski, Peters, Brochu, & Strack, 2008).

Representational models of implicit bias have been extremely influential in the field of social cognition, and they have generated a large amount of research. A major advantage to representational models of implicit social cognition is that they are amenable to formal theoretical modeling. They are also intuitively appealing. Indeed, representational models are built in accordance with the way we store information in other systems, such as computers or libraries. However, it is important to note that a representational model is hypothetical and

abstracted inductively through experimentation, and thus it does not necessarily reflect the way that information is actually represented or how the mind actually operates.

The advantages of representational models are balanced by some important limitations. These include a general disconnect with the non-cognitive systems (e.g., emotion, attention), inconsistencies with functional neuroanatomy, and a lack of connection to actual behavior. For example, several influential dual-process models posit that implicit associations are learned through a slow, associative process in memory (Smith & Decoster, 2000). However, affective associations learned through a classical conditioning occurs rapidly, often after a single exposure to an association (LeDoux, 2000). Therefore, traditional representational models may provide a good account for how semantic associations with social groups (i.e., stereotypes or evaluative associations) are learned and stored, but they do not provide an adequate account of affective forms of bias.

Another critical limitation of representational models is that few, if any, specify a connection between mental processes and behavior, and thus they are silent regarding the mechanism through which implicit racial bias is expressed in behavior. That is, representational models do not address how basic emotional processes, such as autonomically-aroused states like anxiety, fear, anger, or compassion, influence the activation and expression of implicit biases. Some theorists have attempted to address this issue by proposing cognitive representations of affect, which are then assumed to interact in a network with cognitive representations of bias (Gawronski & Bodenhausen, 2006; Storbeck & Clore, 2008). This approach typically focuses on how affect shapes cognitive representations. However, the approach of treating emotions as cognitive structures may not fully capture the nature of a true emotional state or the process through which it influences behavioral expressions of racial bias. Similarly, Strack and Deutch (2004) proposed a model through which cognitive and motivational systems influence

"behavioral schemata" (i.e., a representation of behavior), but the mechanisms through which schemata translate into actual behavioral responses remain unclear. For these reasons, representational models are limited in their ability to account for emotional aspects of implicit intergroup processes and their behavioral expression. We should note, however, that these limitations refer to broad and long-standing questions about the cognition-affect interface with which the field has grappled. Although these are general issues, we see them as critical to the understanding of implicit racial bias effects.

Memory systems model of implicit bias

Although representational models have dominated research on implicit social cognition, alternative approaches have recently emerged from research integrating models of learning and memory from the human and non-human neuroscience literatures. Amodio's Memory Systems Model (MSM) of implicit bias applies an integrative social neuroscience approach to address questions of how implicit racial biases are learned, stored, and expressed in behavior (Amodio, 2008; Amodio & Devine, 2006; Amodio et al., 2003; see also Carlston's Associated Systems Theory, 1994). Past theory and research has demonstrated multiple forms of implicit learning and memory, each associated with distinct neural substrates (Figure 2; Squire & Zola, 1996; Poldrack & Packard, 2003). This model departs from traditional representational models of implicit processing derived from dual-process accounts of automaticity and control, which assume that implicit processes reflect a single system of associations characterized by a uniform processing mode. The MSM posits that different implicit systems learn according to different parameters, and that they influence emotions, perceptions, cognition, and behavior via different neural and neurochemical circuits. A large body of behavioral, neuroimaging, and brain-lesion research now supports the MSM view (Poldrack & Foerde, 2008).

In an effort to better understand the mechanisms of implicit bias and their expression in behavior, Amodio and colleagues have applied the MSM approach to the study of intergroup bias (Amodio, 2008; Amodio et al., 2003; Amodio & Devine, 2006). They noted that affective forms of implicit bias correspond to affective forms of learning and memory, such as classical fear conditioning, which are supported by the amygdala and its associated subcortical circuitry. By contrast, implicit stereotyping reflects semantic associations, which involve conceptual forms of learning and memory, linked to regions of the neocortex such as the left prefrontal cortex (e.g., Brodman areas 45/47) and the medial temporal lobe (Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997). Although most research to date has focused on comparing affective and semantic systems underlying implicit evaluative bias and stereotyping, other aspects of implicit bias likely involve additional systems, such as those associated with habit learning and reward.

The MSM is useful because it generates hypotheses for how different forms of implicit bias should influence judgments and behavior. For example, if implicit affective bias reflects a system that governs the activation of autonomic arousal and triggers avoidance behaviors in response to threat, then measures of implicit evaluations should predict basic inhibition and avoidance behavior. If, by contrast, implicit stereotyping reflects the operation of semantic memory systems, which have stronger connections to neural regions involved in judgment formation and goal representation, then implicit stereotypes should influence impressions of outgroup members and goal-driven behaviors. This distinction has been supported by studies of behavior (Amodio & Devine, 2006; Amodio & Hamilton, 2009) and neural activity (Potanina, Pfeifer, & Amodio, 2009). It is notable that, according to the MSM, an implicit evaluation may reflect a combination of affective and semantic associations. In line with classic models of attitudes, an evaluation may be driven by a combination of affective and cognitive (i.e., semantic) processes (Eagly & Chaiken, 1998). Behavior-based measures of implicit bias, such as

the IAT, are unable to parse the specific contributions of affect and cognition. Nevertheless, Amodio and Devine's (2006) findings suggest that, barring abnormal brain function (Phelps, Cannistraci, & Cunningham, 2003), measures of implicit evaluative bias may reflect affective processes.

The MSM also generates specific hypotheses for how affective and semantic forms of implicit bias may be learned and unlearned. For example, classically-conditioned associations are learned rapidly, often after a single experience. Once learned, they are tenacious and may never be fully extinguished (Bouton, 1994). By contrast, semantic associations are learned slowly, after repeated and highly-probably pairings between two stimuli (Reber & Squire, 1994). Semantic associations are presumably unlearned in a similarly slow fashion, after repeated non-pairings. It is notable that these predictions are different than those suggested by representational models, which assume that implicit associations are learned and unlearned slowly (Smith & Decoster, 2000; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007). Amodio (2008) has suggested that past social cognitive models correspond well to the implicit semantic memory system, but do not account for affective forms of implicit bias. Thus, the MSM is not inconsistent with representational models per se, but suggests that representational models pertain to a subset of the range of implicit processes relevant to race bias. A major advantage of the MSM approach is that it posits a model of implicit bias that is integrated with perceptual, emotional, motivational, and cognitive systems, and it delineates pathways from different memory systems to behavior. This model will become increasingly useful as researchers turn more attention to understanding how implicit biases are expressed in social behavior.

Although our discussion has focused on the implications of the MSM approach for issues of racial bias, the MSM describes general, basic-level processes that apply to all attitude objects, social and non-social alike. It will be interesting for future work to consider the MSM's

predictions for behavior toward groups that are perceived with varying degrees of affect (cf. the Stereotype Content Model; Fiske, Cuddy, Glick, & Xu, 2002). More broadly, we expect that integrative approaches such as the MSM will become more common as the field of psychology becomes increasingly interdisciplinary.

Remaining questions and controversies

Although an enormous amount of research has been conducted on implicit bias, many important questions remain. In this section, we touch on two such issues. The first concerns the meaning of responses on an implicit measure – how should responses on implicit tasks be interpreted? The second concerns the broader controversy of whether bias on an implicit task should be considered a mark of true prejudice.

Issues in implicit measurement

Measurements of implicit bias have a mystique about them. How do they work? How can they measure our hidden thoughts? This mystique has cultivated a view that implicit tasks provide a direct probe into the unconscious mind, such that responses on an implicit bias task provide a pure representation of our mental processes and contents. To be clear, "implicit" tasks measure behavior, or in some cases, physiological responses. The logic is that if a concept is cued (e.g., by a racial prime), then its effect on a subsequent behavior may be observed. For example, Bargh, Chen and Burrows (1996) primed subjects with subliminal pictures of Black faces and measured the extent to which it led to more hostile behavior toward an experimenter. Similarly, Devine (1989) primed African American concepts and measured the degree of stereotyping applied in later judgments of a story character. In both cases, the object of interest is cued (e.g., Black people), and its effect on behavior is measured. In this same way, an "implicit task" primes the object of interest and then measures its effect on a behavioral response (e.g., speed to respond to a target). The main difference is that, in an implicit task, the behavioral

outcome is contained within the task, and the measurement is repeated across several trials. In this sense, an implicit task may be thought of as a "behavioral assay," or a circumscribed index of how the actual behavioral effect would occur in a social situation.

We suggest that a useful distinction between explicit and implicit tasks is that an explicit measure assesses the reporting of a belief, or proposition, whereas an implicit measure assesses a behavioral or physiological response. Considered this way, the critical difference between implicit and explicit measures is the channel of expression through which the response is made, rather than a hypothetical process. As noted previously, an implicit task does not provide a pure measure of implicit or automatic processes (Amodio et al., 2008; Payne, 2001, 2005), but rather a combination of processes that are expressed through behavioral channels. Similarly, explicit measures may also assess a combination of underlying processes, though they may be particularly sensitive to explicit beliefs. For this reason, it is useful (and practical) to interpret implicit task responses as behavioral expressions rather than as pure implicit processes.

Is implicit bias really prejudice?

To be clear, prejudice and discrimination remain strong and pervasive in American society. Controversy and debate surrounding the meaning of implicit race bias measures does not question the existence of prejudice in America. Indeed, the finding that most Americans show more favoritism toward Whites than Blacks on measures such as the IAT cannot be dismissed or explained away – it truly reflects that at some level of processing, people in America tend to have racist tendencies, and these tendencies are often expressed in behavior (Jost et al., in press). This is not controversial. What is controversial concerns a more subtle issue about how implicit racial bias relates to conscious beliefs and overt behavior. Setting aside the issue of whether research on implicit bias reveals a real form of prejudice in American society (it does), this section addresses some of the finer points in evaluating the meaning of implicit bias.

In his seminal paper on the measurement of implicit racial evaluations, Fazio et al. (1995) dubbed their sequential evaluative priming task the "Bona Fide Pipeline." This name was a reference to Jones and Sigall's (1971) "Bogus Pipeline" – a fake physiological contraption that purported to assess subjects' true racial attitudes. When connected to the bogus pipeline, Jones and Sigall's (1971) subjects reported higher levels of prejudice than control subjects, with the belief that any attempt to conceal their true attitudes would expose them as liars. Fazio et al.'s (1995) sequential priming method purported to be a direct conduit to one's true attitude, obviating the need for "bogus" procedures used in the past. Similarly, when the IAT was introduced, it was heralded as a measure of one's "true" attitude (Banaji, 2001). Given that the vast majority of Americans, including non-Whites and egalitarians, showed an Anti-Black bias on the IAT, this view was quite threatening to many people (e.g., Arkes & Tetlock, 2004). In essence, it pointed a finger at most people and accused them of bigotry. Several researchers voiced the concern that laypeople completing the IAT online on web sites would be misled into believing that they were unconscious bigots (e.g., Blanton & Jaccard, 2006).

The "true attitude" view contrasted with Devine's (1989) theory that automatic tendencies reflected passive learning in a historically racist culture, but that one's *true* belief could only be expressed with the aid of controlled processing (see also Amodio et al., 2003, 2008; Devine et al., 2002). Indeed, several researchers have made a specific point to avoid using the term "prejudice" to describe implicit processes because prejudice is a complex construct that is associated with a wide range of attitudes, beliefs, and behaviors, particularly as the term is used colloquially (see Payne & Cameron, this volume). We ascribe to this principle of usage; the reader may have noticed the absence of the term "implicit prejudice" in the present chapter.

A compromise position was proposed by Wilson et al. (2000), who argued that implicit and explicit measures assess different attitudes that exist in different modes of psychological

processing. According to the dual-attitudes approach, an individual may simultaneously possess negative implicit attitudes and positive explicit attitudes toward an outgroup. This approach acknowledges ownership of associations that exist within one's mind, even if they were formed without one's intention and contradict one's explicit beliefs. Importantly, both Devine (1989) and Wilson et al. (2000) argue that implicit attitudes and stereotypes can be overridden with controlled processing, and thus the responsibility for the expression of implicit race bias ultimately resides with the individual.

In the end, the question "is implicit bias prejudice?" is too complex for a simple yes or no answer. The discussion of whether implicit bias constitutes prejudice corresponds to legal distinctions concerning punishment based on intent vs. harm (Heider, 1958). If a person is held accountable based on their intent, then implicit bias is not prejudice. If their intent is irrelevant, but rather harm (i.e., the expression of implicit bias as discrimination) is the key issue, then implicit bias may constitute prejudice. We will leave this debate to the legal scholars (e.g., Lane, Kang, & Banaji, 2007). We hasten to add, however, that from a social psychological point of view, the question of "true prejudice" is not the critical question. That is, the goal of research on implicit bias is not to identify whether a person is prejudiced, but to understand the mechanisms of the social mind as they relate to intergroup processes and social behavior.

Conclusion

Implicit social cognition continues to represent the latest great frontier of social psychology. Although recent advances have already shed light on the psychological mechanisms that operate in the unconscious regions of the mind, they have likely just scratched the surface. Research on implicit race bias has made unique contributions to the study of implicit social cognition. As a domain of study, it stands as an exemplar for the interplay of implicit and explicit attitudes and beliefs in the context of social relationships, goals, and group structures. At the

same time, implicit race bias research has revealed a new dimension of intergroup processes that inform broader theories of intergroup relations. In this way, the field of implicit race bias has come to represent an important link between intrapersonal and interpersonal processes in the social psychological theory and research. In this chapter, we highlighted major extant findings from the field and discussed some of the current debates and controversies that drive much contemporary investigation. Continuously evolving, this field stands poised to contribute new insights into the expression of implicit processes in behavior, further connecting research on social cognition with broader social psychological questions about the individual in society.

References

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

Amodio, D. M. (2008). The social neuroscience of intergroup relations. *European Review of Social Psychology*, *19*, 1-54.

Amodio, D. M., & Devine, P. G. (2005). Changing prejudice: The effects of persuasion on implicit and explicit forms of race bias. In T. C. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd Ed., pp. 249-280). Thousand Oaks, CA: Sage Publications.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*, 652-661.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science*, *18*, 524-530.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, *94*, 60-74.

Amodio, D. M., & Hamilton, H. (2009). *Intergroup anxiety effects on implicit evaluative race bias vs. implicit stereotyping*. Unpublished manuscript, New York University.

Amodio, D. M., Harmon-Jones, E., Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink responses and self-report. *Journal of Personality and Social Psychology*, *84*, 738-753.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, *15*, 88-93.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15,* 257-278.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology*, *87*, 5-22.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1-40). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bargh, J. A. (1999). The cognitive monster: The case against controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 361-382). New York: Guilford Press.

Bargh, J. A., Chen, M., & 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.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Troetschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. Journal of Personality and Social Psychology, 81, 1014-1027.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness

on impression formation. *Journal of Personality and Social Psychology, 43,* 437-449.
Blair, I. (2001). Implicit stereotypes and prejudice. In G. Moskowitz (Ed.), *Cognitive social psychology: On the tenure and future of social cognition* (pp. 359-374). Nahwah, NJ: Erlbaum.
Blair, I. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, 242-260.
Blanton, H., & Jaccard, J. (2006). Tests of multiplicative models in psychology: A case study using the unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *113*, 155-169.
Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology, 80*, 253-267.
Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. Wyer, Jr. (Ed.), *Stereotype activation and inhibition: Advances in social cognition* (Vol. 11, pp. 1-52). Mahwah, NJ: Erlbaum.
Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*, 219-231.
Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, *47*, 1191-1205.
Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 760-773.
Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology*, *23*, 1933–1967.
Carlston, D. E. (1994). Associated Systems Theory: A systematic approach to the cognitive representation of persons and events. In R. S. Wyer (Ed.), *Associated systems theory: Advances in social cognition* (Vol. 7, pp. 1-78). Hillsdale, NJ: Erlbaum.
Cohen, N. J., & Squire, L. R. (1980). Preserved learning and pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science, 210*, 207-210.
Conrey, F. R, Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89, 469-487.
Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314-1329.
Correll, J., Park, B., Judd, C. M., Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology, 37,* 1077-1345.
Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin, 87*, 546-563.
Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of Black and White faces. *Psychological Science*, *15*, 806-813.
Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research*, *17*, 143-169.
Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*, 801-814.

Davis, M. (1992): The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience*, *15*, 353–375.

Devine, P. G. (1989). Prejudice and stereotypes: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5-18.

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 339-360). New York: Guilford Press.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E, & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*, 835-848.

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice and discrimination: Another look. In C. N. McCrae, C. Stangor, & M. Hewstone (Eds.), Stereotypes and Stereotyping (pp. 276-319). New York: Guilford.

Dovidio, J. F., & Gaertner, S. L. (1986). Prejudice, discrimination, and racism: Historical trends and contemporary approaches. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 1-34). New York: Academic Press.

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

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62-68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*, 510-540.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316-326.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 269-322). New York: McGraw-Hill.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE Model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-107). New York: Academic Press.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25,* 603-637.

Fazio, R. H., Chen, J., McDonel, E. C., & Sherman, S. J. (1982). Attitude accessibility, attitude-behavior consistency and the strength of the object-evaluation association. *Journal of Experimental Social Psychology, 18*, 339-357.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*, 316-326.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology, 54*, 297-327.

Fazio, R. H., Powell, M. C., & Herr, P. M. (1983). Toward a process model of the attitude-behavior relation: Accessing one's attitude upon mere observation of the attitude object. *Journal of Personality and Social Psychology, 44*, 723-735.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Status and Competition. *Journal of Personality and Social Psychology, 82*, 878-902.

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30,* 1611-1624.

Freud, S. (1961). Civilization and its discontents (J. Strachey, Ed. & Trans.). Oxford, England: Hogarth. (Original work published 1930).

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly, 46,* 23-30.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708-724.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692-731.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34,* 648-665.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*, 509-517.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology*, *40*, 357-365.

Graf, P., & Schacter, D.L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 501-518.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition. *Psychological Review, 102*, 4-27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3-25.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition. The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Guglielmi, R. S. (1999). Psychophysiological assessment of prejudice: Past research, current status, and future directions. *Personality and Social Psychology Review*, *3*, 123-157.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley & Sons.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513-541.

Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology*, *36*, 300-324.

James, W. (1980). *The principles of psychology*. Oxford: Holt.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*, 349-364.

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D., Dasgupta, N., Glaser, J. & Hardin, C.D. (in press). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no

manager should ignore. *Research in Organizational Behavior.*
Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*, 109-128.
Karlins, M., Coffman, T. L. & Walters, G. (1969). On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology,* 13, 1–16.
Karpinski, A., & Hilton, J. K. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 774-788.
Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, *24*, 407-416.
Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training on the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*, 871-888.
Kawakami, K., Phills, C., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*, 957-971.
Klauer, K. C., & Musch, J. (2003). Affective priming: Findings and theories. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 7-49). Mahwah, NJ: Erlbaum.
Lambert, A. J., Payne, B. K., Shaffer, L. M, Jacoby, L. L., Chasteen, A., & Khan, S. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology, 84*, 277-295.
Lane, K. A., Kang, J., & Banaji, M. R. (2007). Implicit social cognition and law. *Annual Review of Law and Social Sciences*, *3*, 427-451.
Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*, 377-395.
Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311-328.
Langer, E. J., & Abelson, R. P. (1972). The semantics of asking a favor: How to succeed in getting help without really dying. *Journal of Personality and Social Psychology*, *24*, 26-32.
LaPierre, R. T., (1934). Attitudes vs. actions. *Social Forces*, *13*, 230-237.
LeDoux, J.E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155-184.
Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*, 720-722.
Livingston, R. W. & Drwecki, B. B. (2007). Why are some individuals not racially biased? Susceptibility to affective conditioning predicts nonprejudice toward Blacks. *Psychological Science*, *18*, 816-823.
Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.
Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology*, *69*, 397-407.
Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*, 808-817.

Macrae, C. N., Stangor, C., & Milne, A. B. (1994). Activating social stereotypes: A functional analysis. *Journal of Experimental Social Psychology*, *30*, 370-389.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-126). New York: Academic Press.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*, 435-442.

McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, *94*, 792-807.

Mendes, W. B., Blascovich, J., Lickel, B., & Hunter, S. (2002). Challenge and threat during interactions with White and Black men. *Personality and Social Psychology Bulletin*, *28*, 939-952.

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2009). *Reducing the Expression of Implicit Race Bias: Reflexive Control through Implementation Intentions*. Manuscript submitted for publication.

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

Meyer, D. E., & Schvaneveldt, R. W. (1976). Meaning, memory, structure, and mental processes. *Science*, *192*, 27-33.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice reduction efforts. *Journal of Personality and Social Psychology*, *65*, 469-485.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*, 1029-1050.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting and reacting to implicit racial biases. *Social Cognition, 19*, 395-417.

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*, 167-184.

Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Implicit control of stereotype activation through the preconscious operation of egalitarian goals. *Social Cognition*, *18*, 151-177.

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18,* 36-88.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*, 653-667.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421-433.

Park, B., & Judd, C. M. (2005). Rethinking the link between categorization and prejudice within the social cognition perspective. *Personality and Social Psychology Review, 9*, 108-130.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*, 181-192.

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*, 488-503.

Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass*, *2*, 1-20.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277-293.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology, 38*, 384-396.

Pearson, A. R., West, T. V., Dovidio, J. F., Powers, S. R., Buck, R., & Henning, R. (2008). The fragility of intergroup relations: Divergent effects of temporal delay in audio-visual feedback in intergroup and intragroup interaction. *Psychological Science*, *19*, 1272-1279.

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, *59*, 475-486.

Perdue, C. W. & Gurtman, M. B. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology*, 26, 199-216.

Phelps, E.A., Cannistraci, C.J., & Cunningham, W.A. (2003). Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia*, *41*, 203-208.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby J. C., Gore, J. C., Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 729-738.

Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*, 180-183.

Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology*, *41*, 141-156.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*, 197-205.

Poldrack, R. A. & Packard, M. G. (2003). Competition between memory systems: Converging evidence from animal and human studies. *Neuropsychologia*, *41*, 245-251.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium.* Hillsdale, NJ: Erlbaum.

Potanina, P. V., Pfeifer, J. H., & Amodio, D. M. (2009). *Stereotyping and evaluation in intergroup bias: fMRI evidence for a multiple-memory systems model.* Manuscript submitted for publication.

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27*, 26-47.

Rankin, R. E., & Campbell, D. T. (1955). Galvanic skin response to negro and white experimenters. *Journal of Abnormal and Social Psychology*, *51*, 30-33.

Reber, P. J., & Squire, L. R. (1994). Parallel brain systems for learning with and without awareness. *Learning & Memory*, *1*, 217-229.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*, 995-1008.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, 37, 867-878.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality & Social Psychology*, *81*, 856-868.

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

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., & von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, *39*, 75-82.

Shelton, J. N., Richeson, J. A., Salvatore, J., & Trawalter, S. (2005). Ironic effects of racial bias during interracial interactions. *Psychological Science*, *16*, 397-402.

Sherman, J. W. (1996). Development and mental representation of stereotypes. *Journal of Personality and Social Psychology*, *70*, 1126-1141.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*, 314-335.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology*, *18*, 247-255.

Sinclair, S., Lowery. B., Hardin, C., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583-592.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Smith, E. R., & Branscombe, N. R. (1987). Procedurally-mediated social inferences: The case of category accessibility effects. *Journal of Experimental Social Psychology*, *23*, 361–382.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108-131.

Smith, E. R., & Miller, F. D. (1979). Attributional information processing: A response time model of causal subtraction. *Journal of Personality and Social Psychology*, *37*, 1723-1731.

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. A. (1998). Automatic

activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin*, *24*, 1139-1152.

Squire, L. R. (1986). Mechanisms of memory. *Science*, *232*, 1612-1619.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the USA*, *93*, 13515-13522.

Srull, T. K. & Wyer, R. S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology*, *38*, 841-856.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*, 1332-1345.

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorization and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*, 778-793.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences*, *94*, 14792-14797.

Uhlmann, E. L. Brescoll, V. L., & Paluck, E. L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice. *Journal of Experimental Social Psychology*, 42, 491-499.

Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, *73*, 941-959.

West, T.V., Shelton, J.N., & Trail, T.E. (in press). Relational anxiety in interracial interactions, *Psychological Science*.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*, 56-63.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*, 101-126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*, 262–274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality & Social Psychology*, *81*, 815–827.

Figure Captions

Figure 1. Example of a network representation of implicit racial associations. Stereotype associations are depicted by weighted associative links to the concept of African Americans. Evaluative (and sometimes affective) associations are represented as the weighted valence of each link in a network. Alternatively, some theorists represent evaluation in terms of links to general concepts of "positive" and "negative."

Figure 2. Diagram of dissociable memory systems and their putative neural substrates, illustrating qualitatively different forms of implicit learning and memory processes (adapted from Squire & Zola, 1996).

Figure 1

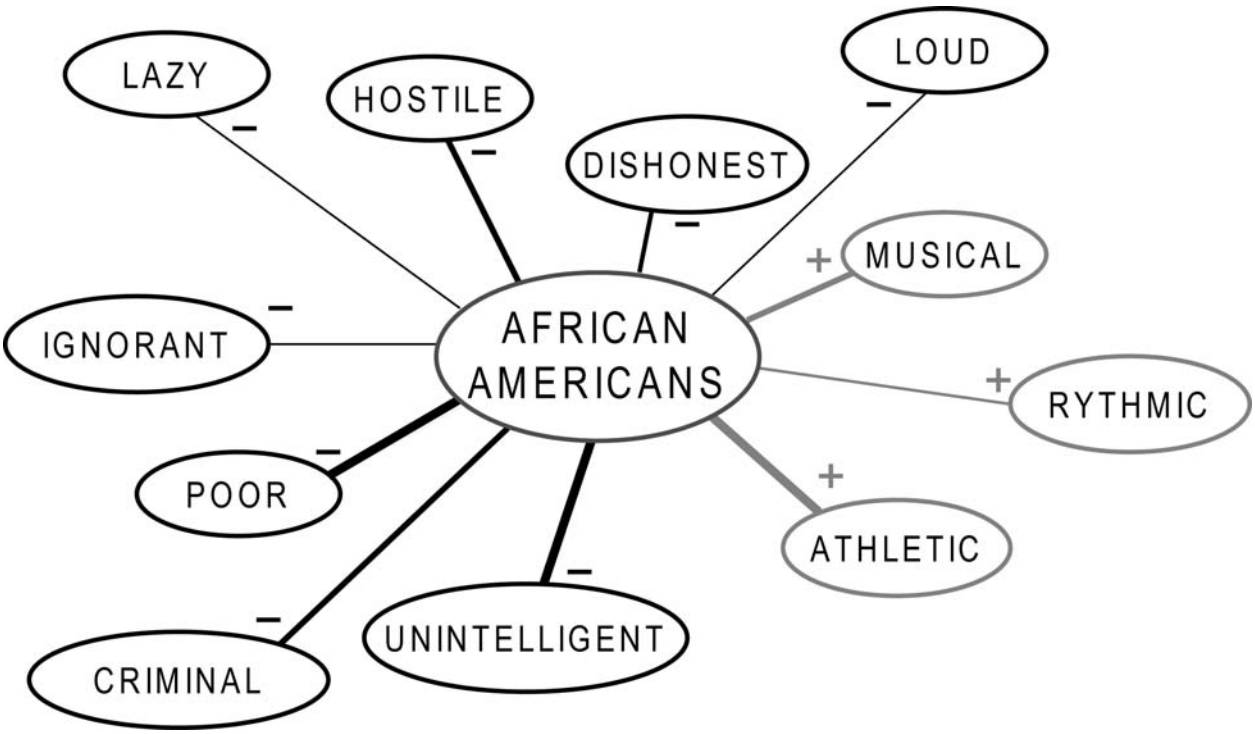
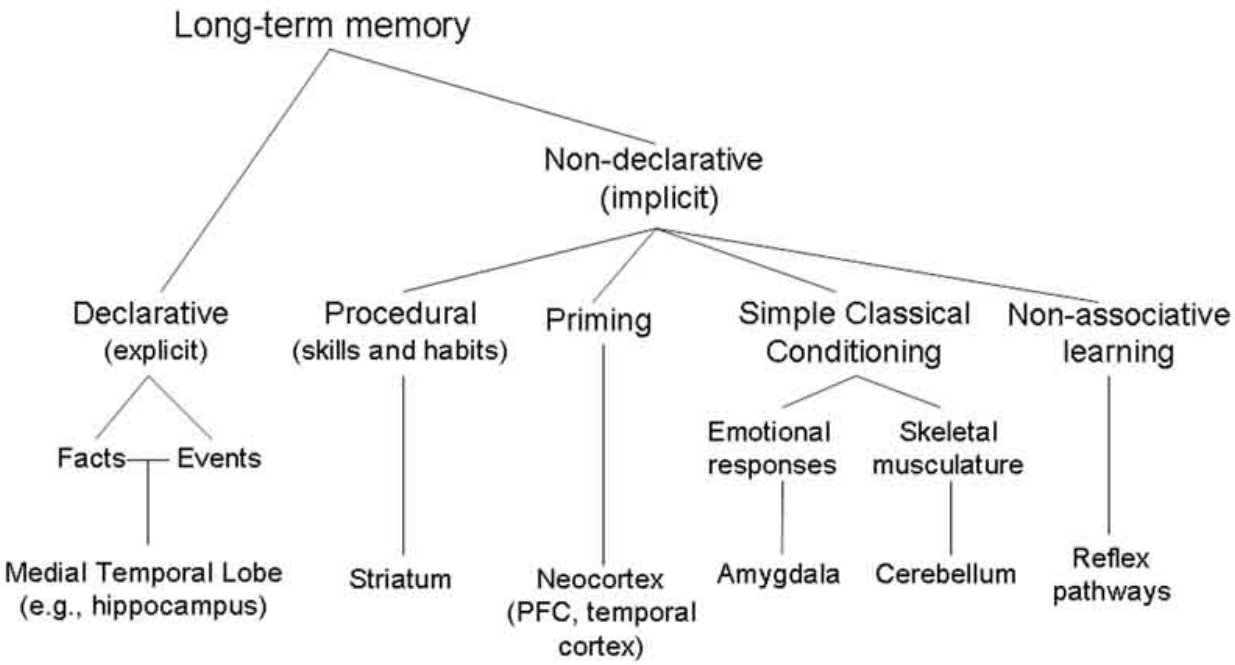

LAZY
HOSTILE
DISHONEST
LOUD
MUSICAL
IGNORANT
AFRICAN AMERICANS
RYTHMIC
POOR
ATHLETIC
UNINTELLIGENT
CRIMINAL
–
+

Figure 2


Long-term memory
Non-declarative (implicit)
Declarative (explicit)
Procedural (skills and habits)
Priming
Simple Classical Conditioning
Non-associative learning
Facts
Events
Emotional responses
Skeletal musculature
Medial Temporal Lobe (e.g., hippocampus)
Striatum
Neocortex (PFC, temporal cortex)
Amygdala
Cerebellum
Reflex pathways