19
IMPLICIT BIAS AND RACE

Michael Brownstein

Introduction

In 1928 L. L. Thurstone asked 239 white male students at the University of Chicago to report their preferences between pairs of "races" and "nationalities," such as "Greek vs. Mexican," "American vs. Hindu," and "Negro vs. Turk." Unsurprisingly, one of the most discrepant responses between pairs was in the "American vs. Negro" comparison, with almost all participants strongly preferring "Americans" to "Negros." (The very fact that Thurstone distinguished between "American" and "Negro" or "Hindu" is perhaps also unsurprising, and unsettling. See Devos and Banaji (2005) on associating "white" with "American.") Compare this to Brian Nosek and colleagues' finding in 2007 that in a pool of 700,000 subjects the most frequent answer to the question, "who do you prefer, black people or white people?" was "I have no preference." The disparities between these findings underscores how dramatically explicit (i.e., verbally reported) black-white racism—white people's prejudices and racial attitudes toward black people—has declined over the past 75 years in the United States (see also Judd et al. 1995 and Schuman et al. 1997; for discussion of implicit bias directed toward non-black members of socially stigmatized groups, such as Asians and Latinxs, see Dasgupta (2004)). Despite this, it's clear that racial discrimination persists systematically, pervasively, and brutally in the United States. This presents a puzzle that philosophers, sociologists, political scientists, economists, psychologists, and others have considered. Why do stark racial disparities in housing and hiring, police violence and incarceration, medical treatment and health outcomes, and on and on, persist in places like the United States today, if most people's ~~explicit~~ beliefs about race have changed so much?

One part of the answer is that what people say explicitly—on a questionnaire like Thurstone's or one of its contemporary analogues, such as the Modern Racism Scale (MRS; McConahay 1986)—does not represent the whole of what people feel or think. This is news to almost nobody, of course. What is news to many is that some element of people's thoughts and feelings can be measured without having to ask them directly what they think or feel, using a host of "indirect" measurement techniques, most prominently the "Implicit Association Test" (IAT; Greenwald et al. 1998). The IAT is one of many reaction time measures that asks participants to sort words or pictures into categories as quickly as possible while making as few errors as possible. A person taking the most well-known IAT—the black-white IAT—will be presented with variations of the images in Figures 19.1–19.4.

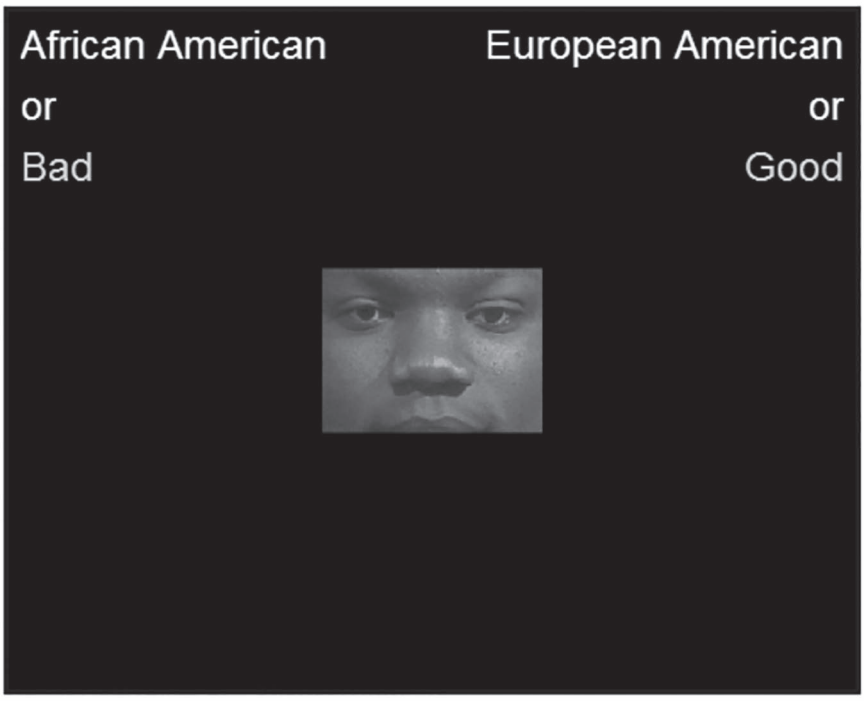

Figure 19.1 Implicit Association Test
Screenshots from the Project Implicit website, accessed on May 24, 2017, https://implicit.harvard.edu/implicit/.
Reprinted with permission.

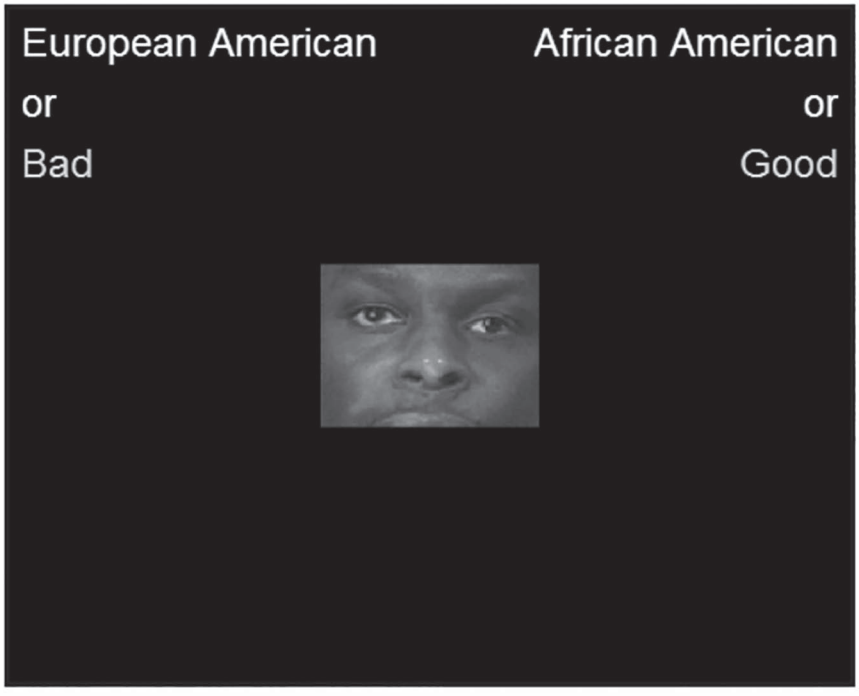

Figure 19.2 Implicit Association Test
Screenshots from the Project Implicit website, accessed on May 24, 2017, https://implicit.harvard.edu/implicit/.
Reprinted with permission.

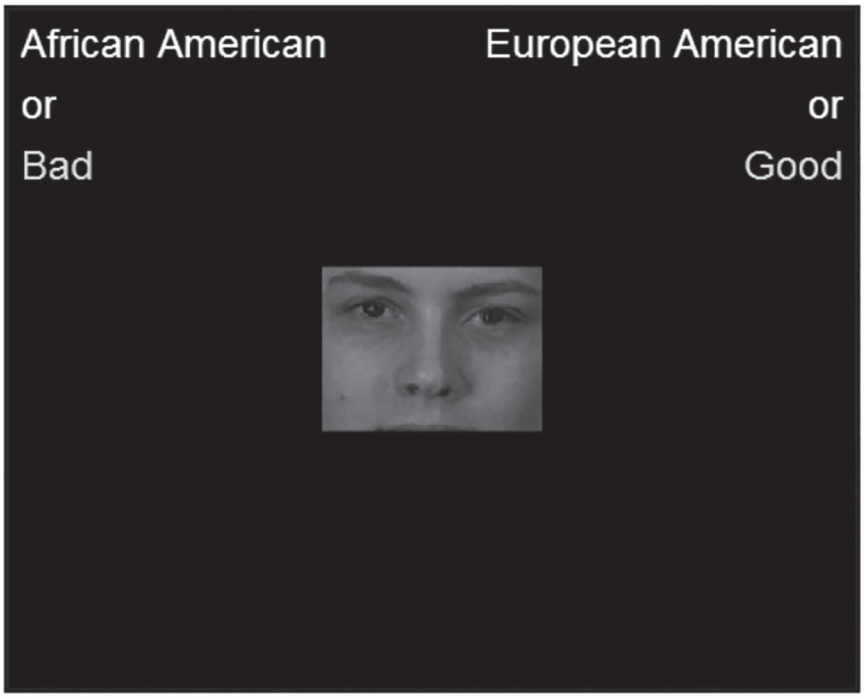

Figure 19.3 Implicit Association Test
Screenshots from the Project Implicit website, accessed on May 24, 2017, https://implicit.harvard.edu/implicit/.
Reprinted with permission.

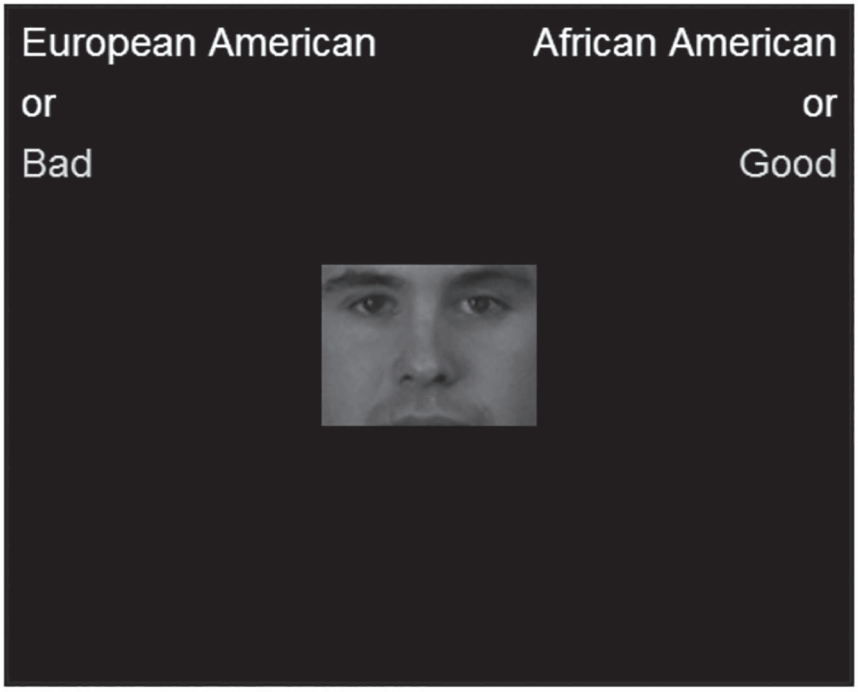

Figure 19.4 Implicit Association Test
Screenshots from the Project Implicit website, accessed on May 24, 2017, https://implicit.harvard.edu/implicit/.
Reprinted with permission.

The goal is to sort the pictures to the left or right. Notice that the categories on the left and right pair a social group label with a positive or negative word. In Figures 19.1 and 19.3, the pairing is "compatible" with widespread negative attitudes toward black people, while in Figures 19.2 and 19.4 the pairing is "incompatible." Most white subjects (over 70%) will be faster and make fewer mistakes on compatible than on incompatible trials (Nosek et al. 2007). Researchers consider this to represent an "implicit preference" for white faces over black faces. Remarkably, while roughly 40% of black participants demonstrate an implicit in-group preference for black faces over white faces, and 20% show no preference, roughly 40% of black participants demonstrate an implicit out-group preference for white faces over black faces (Nosek et al. 2002; Ashburn-Nardo et al. 2003; Dasgupta 2004). This finding has upended the view that in-group favoritism is the primary driver of implicit bias. Rather, it appears that implicit bias is driven by a combination of in-group favoritism and sensitivity to the value society places on particular groups.

The significance of computerized reaction time measures like the IAT could seem small if it weren't for an extensive body of literature showing that IAT scores predict discriminatory behavior. The stronger one's associations of good with white faces and bad with black faces on the black-white IAT, the more likely one is, for example, to judge otherwise equivalent curricula vitae (CV) better if they have white-sounding names on them than black-sounding names (Bertrand et al. 2005). Doctors who demonstrate more implicit bias on the IAT are more likely to attribute equivalent symptoms to coronary artery disease and recommend thrombolysis for white patients compared to black patients (Green et al. 2007). And perhaps most ominously, in light of continued shootings of unarmed black men by police officers, the stronger one's implicit biases as measured by the IAT, the more likely one is to "shoot" unarmed black men in a computer simulation than unarmed white men (Correll et al. 2002; Glaser and Knowles 2008), a pattern that remains when study participants are black as well as when they are police officers (Plant and Peruche 2005). These are just a flavor of the disturbing findings. Overall, the IAT is particularly useful for predicting negative non-verbal and "micro-behaviors" (Valian 1998, 2005; Dovidio et al. 2002; Cortina 2008; Cortina et al. 2011; Brennan 2013) and action undertaken when information is incomplete, decisions need to be made quickly, or agents are stressed or under cognitive load.

The IAT (and other indirect measures) is thought to quantify "implicit biases." Implicit bias is a term of art referring to prejudiced implicit attitudes, and implicit attitudes are, of course, contrasted with explicit attitudes. By and large, explicit attitudes can be thought of as what people report on questionnaires or other "direct" measures. But what is an implicit attitude? There are two questions here. First, what is an attitude? In psychology, attitudes are understood as likings or dislikings; or, more formally, as associations between a concept and an evaluation (Nosek and Banaji 2009). This conceptualization of attitudes is importantly different from the typical usage in philosophy, which is much more expansive (including beliefs, desires, intentions, etc.). Unless otherwise indicated, hereafter I'll discuss attitudes in the psychological sense. The structure of attitudes, in particular implicit attitudes, is a matter of theoretical contention (Gendler 2008a, 2008b; Levy 2012, 2015; Madva 2012, 2016; Mandelbaum 2013, 2016; Beeghly 2014; Brownstein 2016; Brownstein and Saul 2016a, 2016b), but I won't address this topic directly.

The second question is what makes an attitude implicit. Following the general characterization of implicit attitudes found in the empirical literature, most philosophers have focused on two qualities: lack of awareness of one's implicit attitudes and lack of control over them. For example, here is Daniel Kelly and Erica Roedder's characterization in their influential 2008 paper:

> the IAT requires subjects to make snap judgments that must be made quickly, and thus without moderating influence of introspection and deliberation and often without conscious intention. Biases revealed by an IAT are often thought to implicate relatively automatic processes.
>
> (525)

Similarly, Jennifer Saul (2012: 244) describes implicit biases as "unconscious tendencies to automatically associate concepts with one another," and in a similar vein elsewhere I call implicit biases "relatively unconscious and relatively automatic features of prejudiced judgment and social behavior" (Brownstein 2015: 1).

The qualifiers found in these definitions—"relatively" automatic, "relatively" unconscious, "tendencies" to. . . —reflect the fact that research increasingly suggests that there are senses in which people *are* conscious of their implicit attitudes and *can* control them. Implicit attitudes are unlike features of our minds to which we have no direct introspective access, like the fact that visual stimuli from the retina are processed upside-down. We can learn this fact, but we can't know it introspectively. Similarly, implicit attitudes *can* be reshaped and even possibly eliminated using certain "self-regulation" techniques. That is, implicit biases are unlike untrainable reflexes, such as pupillary dilation.

Recent research on awareness of and control over implicit biases calls for further consideration of what makes an attitude implicit. Such consideration has many philosophical ramifications, in addition to simply clarifying the construct called implicit bias. For example, whether we are morally responsible for having implicit biases or for acting in ways influenced by implicit bias depends on what it means for those biases to be implicit (Kelly and Roedder 2008; Smith 2008, 2012; Levy 2012, 2015; Holroyd 2012; Saul 2013; Levy and Mandelbaum 2014; Brownstein 2015; ~~Madva, ms b~~; and the chapters by Faucher, Glasgow, Sie and Voorst Vader-Bours, Washington and Kelly, and Zheng in Brownstein and Saul 2016a, 2016b). Moreover, whether the pervasiveness of implicit bias is cause for skepticism about our ability to make unbiased assessments of CVs, resumes, student papers, and more will also depend on what implicitness entails. A third philosophical question—which has received less attention than these—focuses on what, in broad terms, the empirical literature on implicit bias tells us about contemporary attitudes toward race. This is the question on which I will focus here.

To be clear, my aim is *not* to give an argument about the nature of racism in the modern world or in the United States. Rather, my aim is to give an argument about what the research on implicit bias tells us about attitudes toward race in a broad sense. An important caveat is that the participants in most research to date have been white undergraduates in the US and UK, so in reality my argument will be about what the research on implicit bias tells us about this population. (See Dunham et al. (2006) for an example of research on implicit intergroup attitudes in non-US or UK populations.) Is the dramatic decline in explicit racial prejudice within this population due to actual

changes in people's racial attitudes? Or is it rather due to changes in what is socially acceptable to say and do? Or perhaps attitudes have changed, but largely only in respect of subjects' awareness of them. Here, then, are three possible faces of contemporary black-white racial cognition, as represented by the literature on implicit bias: (1) racial attitudes have changed wholesale over the past century; (2) racial attitudes as such haven't changed, but social norms have; (3) racial attitudes have "split" into conscious and unconscious attitudes.

While each of these is possible, I will not give the first option—that racial attitudes have changed wholesale—any further consideration. There is no doubt that *specific* racial attitudes, or specific components of racial attitudes, have changed. But it is beyond question that racial prejudice persists today.

Research on implicit bias can only reveal an incomplete picture of racial prejudice and discrimination, of course. Psychological research on racial cognition cannot capture many crucial political, economic, and institutional causes of discrimination and prejudice (Anderson 2010; see Madva, forthcoming for discussion). Nevertheless, it can tell us much, and understanding what it tells us is important. I will begin in the next section by describing the two most prominent ways that researchers in the empirical literature have characterized implicit racial attitudes. Different theories of implicit social cognition, I shall try to show, align with different philosophical interpretations of contemporary racial cognition. One theory stems from research on automaticity and considers implicit biases to reflect people's "true" attitudes in the absence of "contamination" by strategic self-presentation considerations (i.e., considerations of how one wants to be received by others). On the interpretation I will call "True Attitudes," which follows from this theory, most people's ownmost attitudes toward socially stigmatized groups (e.g., blacks, women, Latinos, the elderly, members of the LGBTQ community) really are prejudiced, yet in some circumstances people are able to control these prejudiced thoughts and feelings in order to act in accord with social norms. A second theory stems from research on memory and considers implicit biases to be unconscious counterparts to people's conscious attitudes. On the interpretation I call "Driven Underground," people often have conflicting thoughts and feelings about others based on their perceived social group membership, and some of those thoughts and feelings are unavailable to introspection. These two interpretations of implicit bias—"True Attitudes" and "Driven Underground"—have different ramifications for questions about moral responsibility, epistemology, and ethics. However, both conceptions are somewhat flawed, I will argue in the third section. After that I will then argue that what makes attitudes implicit is not a difference in their *content*, but rather a difference in the *processes* that cause implicit attitudes to form and change. A leading process-focused theory of implicit attitudes is Bertram Gawronski and Galen Bodenhausen's Associative-Propositional model of evaluation (APE; 2006, 2011). APE suggests that an attitude is implicit when it is "validity-inapt." I conclude by considering what the research on implicit bias tells us about contemporary racial attitudes if APE's conceptualization of implicitness is correct.

True Attitudes and Driven Underground

Research on implicit social cognition has two distinct roots, one focusing on automaticity and the other focusing on unconsciousness. These manifested in two related streams

of research, the first, focusing on automaticity and led by Russ Fazio, and the second, focusing on unconsciousness and led by Anthony Greenwald and Mazarin Banaji. Fazio and Greenwald/Banaji's research led to two different interpretations of modern racial cognition: "True Attitudes" and "Driven Underground."

(What follows in the next two paragraphs is derived from Brownstein (2016), which is in turn indebted to the cogent history of research on implicit social cognition found in Payne and Gawronski (2010) and Amodio and Devine (2009). Note that the two streams of research I discuss below are not the only significant influences on implicit social cognition research. John McConahay's "Modern Racism Theory" (McConahay et al. 1981; McConahay 1982) argues that explicit prejudice has been funneled into more socially acceptable beliefs about public policy, such as affirmative action and desegregation programs. This is probably true, but does not account for well-document effects of implicit attitudes on socially unacceptable behavior, such as biased review of résumés and CVs. Similarly, Jack Dovidio's and Samuel Gaertner's work on "aversive racism" (Gaertner and Dovidio 1986; Dovidio and Gaertner 2004) has been very influential but does not account for the full scope of contemporary research. Aversive racism is characterized by unconscious negative feelings, but it is clear that implicit black-white racial bias is equally, or perhaps even primarily, driven by *preferences* for whites compared to aversions to blacks (Brewer 1999; Dixon et al. 2012; Greenwald and Pettigrew 2014).)

Fazio's work was influenced by the cognitive psychology of the 1970s, which distinguished between "controlled" and "automatic" information processing in memory (e.g., Shiffrin and Schneider 1977). What Fazio showed was that attitudes can also be understood as activated by controlled or automatic processes. The "sequential priming" technique (Fazio 1995) measures social attitudes by timing people's reactions (or "response latencies") to stereotypic words (e.g., "lazy" or "nurturing") after exposing them to social group labels ("black," "women," etc.). Most people are significantly faster to identify a word like "lazy" in a word-scramble after being exposed to the word "black" (compared with "white"). A faster reaction of this kind is thought to indicate a relatively automatic association between "lazy" and "black." According to Fazio's MODE model of attitudes ("Motivation and Opportunity as Determinants"; Fazio 1990; Fazio and Towles-Schwen 1999; Olson and Fazio 2009), these associations are activated in the presence of relevant cues and facilitate an automatic attitude-to-behavior process. In some cases, though, people have control over their automatic associations. The difference between direct and indirect measures (e.g., a questionnaire like the MRS and reaction time tests like sequential priming and the IAT), on this view, reflects a difference in the control subjects have over their responses. MODE understands control in terms of motivation and opportunity to exert effortful, deliberative control over one's behavior. When someone has low motivation or opportunity to rein in her automatic associations, those associations will guide her behavior and judgment. For example, MODE explains lower correlations between implicit and explicit racial attitudes, compared with higher correlations between implicit and explicit attitudes toward food and consumer preferences, in terms of race being a socially sensitive topic compared with the latter preferences. On socially sensitive topics like race, that is, people are more motivated to control their automatic reactions. Indirect measures like sequential priming and the IAT manufacture this situation (of low control due to low motivation and/or opportunity to deliberate).

The broader notion embedded in this research was that indirect measures offer a window onto people's attitudes themselves, independent of other factors that affect behavior, such as higher-order goals, self-presentation concerns, or cognitive depletion. Indeed, Fazio and colleagues (1995) characterized sequential priming as a "bona fide pipeline" to people's attitudes. In fact, MODE technically denies the distinction between implicit and explicit attitudes. Rather, it is a "one-process" model. Attitudes as such are captured by techniques like sequential priming; it is the degree of control people have over their attitudes, in conjunction with the strength of their attitudes, which determines their responses. This leads to the interpretation I call "True Attitudes," as it conceptualizes indirect measures as capturing our prejudices before they are "contaminated" by controlled processes, such as a desire to present oneself as unprejudiced. Such desires and motives occur *downstream* from racial attitudes, according to MODE.

In contrast to MODE, Greenwald, Banaji, and colleagues' research focused on unawareness of implicit attitudes. This stream of research interprets scores on direct measures of racial attitudes to represent the attitudes people know they have, while scores on indirect measures are thought to represent the introspectively unidentified "traces" of past experiences on one's feelings, thoughts, and behaviors. This research was influenced by theories of implicit memory, which was understood generally as the influence of past experience on later behavior without conscious memory of the past experience (e.g., Jacoby and Dallas 1981; Schacter 1987). One can see the role of theories of implicit memory in Greenwald and Banaji's seminal definition of implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (1995: 8). Here the emphasis is not on automaticity but on the introspective unavailability of implicit attitudes, or, alternately, the introspective unavailability of the past experiences that formed those attitudes. Borrowing a concept from Dovidio and Gaertner (1986), I call the interpretation this view leads to "Driven Underground." The guiding idea is that in the modern world people persist in holding both prejudiced and unprejudiced racial attitudes, but don't themselves know about the former. What precisely "knowing about" one's implicit attitudes means can be interpreted in different ways (see discussion below).

Awareness and Control Over Implicit Attitudes

In addition to emphasizing different features of implicitness—"True Attitudes" emphasizing automaticity and "Driven Underground" emphasizing unconsciousness—these two approaches also deny what the other claims. On the one hand, Fazio and colleagues deny that direct and indirect measures capture two different kinds of mental states, one conscious and the other unconscious. Olson and Fazio (2009: 49), for example, write, "[MODE] maintains that people tend to generally be aware of their attitudes and that it is motivational forces, not some consciousness-impervious shield, that prevents their verbal expression." On the other hand, Greenwald and Banaji emphasize that implicit attitudes are not reflections of automatic processing alone. Rather, much of their research focuses on the combined contributions of automatic and controlled processes to the formation and change of implicit attitudes themselves.

Both of these denials helpfully illuminate components of what is clearly a complex phenomenon. Implicit attitudes are neither outside of conscious awareness as such nor

are they completely automatic states. For example, despite uncorrelated scores on direct and indirect tests (i.e., what people report on a questionnaire vs. how they score on a test like the IAT), people are fairly good at predicting their own implicit attitudes (Hahn et al. 2013). And when people are told that the IAT is "as close to a lie detector test as is possible," most people's scores move significantly closer to their scores on direct measures (Nier 2005). The literature on awareness of implicit attitudes is growing too. At present, it appears that people are comparatively more aware of what their implicit attitudes are (their "content") than they are aware of the causal origins of their implicit attitudes or of the many ways that their implicit attitudes influence their behavior. This point is underscored by Nosek and Hansen (2008) who examined a wide range of implicit attitudes (toward race, gender, food, sports teams, brand names, political figures, etc.) in over 100,000 participants and found that the best way to predict a person's implicit attitudes toward ψ is to ask them explicitly how warmly they feel about ψ.

That people tend to lack source and impact awareness of their implicit attitudes may provide succor for "Driven Underground," depending on whether "Driven Underground" is a view about awareness of one's mental states as such or, rather, a view about awareness of the origins and effects of one's mental states. If "Driven Underground" defines implicit attitudes as introspectively unavailable mental *states*, then data demonstrating that people lack source and impact awareness of their implicit attitudes does not vindicate it. Wilson and colleagues (2000) and Banaji (2001) define implicit attitudes as mental states in this way. Greenwald and Banaji's definition of implicit attitudes as "traces of past experience" suggests a less mentalistic interpretation of "Driven Underground," according to which what people are unaware of is the origins of their implicit attitudes. This less mentalistic interpretation may be a better representation of the data, but it is unclear how much it can clarify what makes an attitude implicit, since we are often unaware of the origins and effects of our explicit attitudes (cf. the large literature on confabulation in consumer preferences, for example Nisbett and Wilson (1977); see Payne and Gawronski (2010) for discussion).

The literature on control over implicit attitudes is extensive and complex, but it makes clear that implicit biases are not automatic in the sense of being inescapable (akin to motor reflexes or automatisms like sleepwalking), as proponents of "True Attitudes" have suggested (e.g., Fazio et al. 1986; Bargh 1999). For example, when people adopt specific "if-then" plans (known as "implementation intentions") to behave in unbiased ways, the effects can be significant. In the context of a "shooter bias" simulation (see the first section of this chapter), in which one's goal is to shoot all and only those individuals shown holding guns (rather than other objects such as cell phones), biased responding decreases significantly if one says to oneself beforehand, "if I see a person with a gun, then I will shoot!" (Mendoza et al. 2010). Other effective strategies for shifting implicit biases involve increasing one's exposure to images, film clips, or even mental imagery depicting members of stigmatized groups acting in stereotype-discordant ways (Blair et al. 2001; Dasgupta and Greenwald 2001) and engaging in meaningful intergroup contact (Dasgupta and Rivera 2008). Of course, care must be taken in considering which data on control over implicit attitudes do and do not cut against "True Attitudes." MODE predicts that implicit attitudes can be controlled downstream from their activation, but it pointedly denies that implicit attitudes themselves involve controlled processes. To get a handle on this distinction, imagine the difference between Ulysses having his men bind him to the mast of his ship in order

to resist the allure of the Sirens—which is a form of self-control downstream from the activation of desire—and someone else who has managed to change or eliminate the desire itself.

What Is Implicit About Implicit Attitudes?

The discussion thus far may leave readers understandably confused. If implicit attitudes are neither unconscious nor automatic, then why consider them to be distinct from explicit attitudes? One way to move forward is to recognize that automaticity and (un) consciousness are kinds of operating *conditions*. That is, they are descriptions of when—the conditions under which—implicit attitudes are operating. Indirect tests like the IAT manufacture these conditions; they do not require people to know that their social attitudes are being measured, nor do they provide people sufficient time to control their responses deliberatively. But one must exercise care in making an inference from these characteristics of tests to the nature of the mental states illuminated by these tests. Moreover, explicit attitudes sometimes operate under these conditions as well. Personality inventories are explicit measures, but it is not clear that people know what is being measured when taking them (e.g., people might not know they are being asked about their degree of neuroticism when they are asked whether they get stressed out easily).

Instead of defining implicit attitudes in terms of the conditions under which they are manifested, an alternative is to define them in terms of the *principles* according to which they operate. Several groups of researchers have taken this route. Here I will focus on Gawronski and Bodenhausen's "Associative-Propositional" model of evaluation (APE; Gawronski and Bodenhausen 2006, 2011), which treats implicit and explicit attitudes as behavioral manifestations of two distinct kinds of mental process.

According to APE, all information is stored in the mind in the form of associations. For example, the statement, "black people are a disadvantaged group" represents the association between "black people" and "disadvantaged group" (Gawronski and Bodenhausen 2011). As we move through the world, the associations stored in our memories are constantly activated. Hearing the name "Malcolm X," for example, might activate the thought that black people are a disadvantaged group. Of course, a person will have many associations with the name Malcolm X, just as with virtually any cue. Nevertheless, APE offers a complex account of which associations will be activated in a given context. APE refers to this process of the activation of associations as *associative processing*. Sometimes, however, we are concerned to validate the information supplied by associative processing. That is, sometimes we are concerned with whether a given association is true or false. APE refers to this process of validation as *propositional processing*. The result of propositional processing might be the thought that "it is true (false) that black people are a disadvantaged group." The differences between associative and propositional processes comprise the heart of APE. And the fundamental difference between them are the "laws" according to which they operate. While associative processes are driven by the spatiotemporal contiguity of stimuli with associations stored in memory, propositional processes are driven by subjective assessments of truth.

APE is put to work to distinguish implicit and explicit attitudes in the following way. When a person reads through a pile of résumés (for example), she may notice (consciously or unconsciously) the names of the job candidates. These names will trigger associations with particular social groups (e.g., Jamal may trigger associations with black

men; Emily may trigger associations with white women). In addition, people often associate positive and negative stereotypes with particular social groups. For example, many white Americans associate negative stereotypes such as "lazy" and seemingly positive stereotypes such as "athletic" with black men. (I say "seemingly" because stereotypes such as "athletic" can be positive in some contexts but negative in others; see Madva and Brownstein (2016).) Upon registering the name "Jamal," these stereotypes may become activated. Because this is an associative process, the name Jamal may activate the concept lazy in independence of whether the person believes it to be true or false that people with the name Jamal tend to be lazy. Such activated associations may manifest in the consciousness of the résumé reader as a vague negative gut feeling, although this emergence into consciousness is not a defining feature. What *is* crucial, according to APE, are the ways in which an activated association gives rise to behavior. One possibility is that associative processing ~~alone~~ "guides" the résumé reader's response. This is the situation manufactured by the IAT and other indirect measures. A second possibility is that the reader transforms her association into a proposition (e.g., "black people are lazy"), which she then endorses or rejects. This is the situation manufactured by questionnaires and other direct measures of attitudes.

APE offers a complex account of the interaction of associative and propositional processes, as well as the relations between associative and propositional processing and consciousness and automaticity, the result of which are predictions about the conditions under which one or the other process will guide a person's behavior. APE is surely not accepted by all theorists, although there is widespread support for process-focused accounts of implicit attitudes. If APE represents a plausible model of the empirical data, what does it tell us about modern racism?

Two Faces of Prejudice

APE suggests that the implicit/explicit divide does not track automaticity/controlled processes, nor does it track non-conscious/conscious mental states per se, although control over and consciousness of one's mental states act as important moderators of the activation of implicit attitudes. Rather, the implicit/explicit divide tracks the difference between mental processes that are and are not "validity-apt." One face of prejudice—represented by explicit, "old-fashioned" racism—reflects what people take to be true. On some interpretations, a natural way to express this is that propositional processing issues in beliefs. This stems from the notion that beliefs "track truth" or have the "aim" of being true (Velleman 2000; Gendler 2008a, 2008b). Related interpretations of prejudiced propositional attitudes might instead stress concepts like "endorsement" or "identification" (e.g., Frankfurt 1971) or "judgment-sensitivity" (Scanlon 1988; Smith 2005, 2008, 2012). What holds all of these notions together is the idea that the person *takes a stand* toward the content of her attitudes, regarding them as true or false, mine or not mine, valid or invalid.

The other face of prejudice—represented by implicit bias—reflects the information one has encoded from one's social environment, media, and so on. But this reflection of information is complex. First, it is *not* a reflection of "material" reality alone. The black-violent implicit stereotype is not just a reflection of crime statistics, but also cultural messaging, such as depictions of black men in film, television, and other forms of culture. (See Jussim et al. (2009) and Hardin and Banaji (2013) for discussion.) Second,

associative processes are *not* merely a reflection of "cultural knowledge," that is, they are not a reflection of stereotypes "in the air" or what a person thinks other people think (Nosek and Hansen 2008). Rather, associative processes reflect one's own enculturation and social learning, in independence of whether one regards one's associations as one's own or as what others think. Third, many factors, including personality, age, socioeconomic status, and so on, mediate and moderate the ways in which one encodes information. For example, APE claims that the activation of associations depends largely on the "fit" between the stimuli a person encounters and her preexisting associations. Thus the activation of associations in response to one and the same stimulus may be quite different for two people who have different preexisting associations.

There is no obvious go-to category to represent what these associative processes issue in. Tamar Gendler (2008a, 2008b) has recommended the *sui generis* state that she calls an "alief," which, roughly can be understood as a mental state with relatively fixed and tightly bound representational, affective, and behavioral components. For Gendler, aliefs are automatic and (largely) unconscious, and are meant to represent the psychological underpinnings of implicit bias. Of course, the name we give to these states is far less important than the properties we ascribe to them. The key property, I have suggested, is that they reflect complex yet validity-inapt processes of enculturation. Implicit attitudes are normatively deviant because they are largely insensitive to the rules that we ourselves set down, this is, our beliefs, values, and ideals. Call this view "arationality": research on implicit bias shows that different kinds of racial attitudes operate according to different kinds of rules. Tests of implicit bias reveal neither our true attitudes nor our sublimated, underground attitudes. Rather, they reveal our arational attitudes.

Broadly, what this tells us is that contemporary racial attitudes reflect the ways in which we are imperfectly rational creatures. Human beings are, perhaps, distinct in the animal kingdom in our ability to set normative standards for ourselves, and to treat those standards not only as conventions but as sources of morality. Explicit racism plays this game. It takes a (grossly perverted) stand on what is right and wrong. Implicit bias, however, does not play this game. It shows us to be imperfectly rational in a different sense, namely, that much of what we think, feel, and do does not stem from the setting of normative standards, from a concern for morality, or for a concern for "who we are" or "who we want to be." In this sense, the empirical literature on implicit social cognition simply throws more fuel on the fire—the fire of bounded rationality—that cognitive and social psychologists, behavioral economists, and others have been burning for the past 40 or so years (Banaji 2003; Hardin and Banaji 2013). No doubt, automaticity and unconsciousness are central elements of what makes us boundedly rational creatures. But automaticity and unconsciousness are themselves not the processes that render an attitude implicit, and indeed, many of our beliefs, values, and other explicit attitudes can be automatic and unconscious (Railton 2009, 2014; Arpaly and Shroeder 2012). Rather, it is ~~an~~ a immunity to those features of our minds that enable our distinct rationality that marks implicitness.

In the broadest sense, research on implicit bias calls us to rethink the relationship between morality and rationality. It is increasingly and abundantly clear that having good intentions and respectable beliefs is insufficient for being the kind of moral creatures we wish to be, and, that having bad intentions and unrespectable beliefs is not

necessary for being an immoral creature. Research on implicit bias calls us to reject any conception of rational agency on which this is mysterious.

References

Amodio, D., and Devine, P. (2009) "On the Interpersonal Functions of Implicit Stereotyping and Evaluative Race Bias: Insights From Social Neuroscience," in R. Petty, R. Fazio, and P. Briñol (eds.), *Attitudes: Insights From the New Wave of Implicit Measures*, Hillsdale, NJ: Erlbaum, pp. 193–226.

Anderson, E. (2010) *The Imperative of Integration*, Princeton: Princeton University Press.

Arpaly, N., and Shroeder, T. (2012) "Deliberation and Acting for Reasons." *Philosophical Review* 121, no. 2: 209–239.

Ashburn-Nardo, L., Knowles, M.L., and Monteith, M.J. (2003) "Black Americans' Implicit Racial Associations and Their Implications for Intergroup Judgment." *Social Cognition* 21, no. 1: 61–87.

Banaji, M. (2001) "Implicit Attitudes Can Be Measured," in H.L. Roediger, J.S. Nairne, I. Neath, and A. Surprenant (eds.), *The Nature of Remembering: Essays in Remembering Robert G. Crowder*. Washington, DC: American Psychological Association, pp. 117–150.

Bargh, J. (1999) "The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects," in S. Chaiken and Y. Trope (eds.), *Dual-Process Theories in Social Psychology*, New York: Guilford Press, pp. 361–382.

Beeghly, E. (2014) *Seeing Difference: The Epistemology and Ethics of Stereotyping*, Ph.D. diss., University of California, Berkeley.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005) "Implicit Discrimination." *American Economic Review*: 94–98.

Brennan, S. (2013) "Rethinking the Moral Significance of Micro-Inequities: The Case of Women in Philosophy," in F. Jenkins and K. Hutchinson (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Brewer, M. (1999) "The Psychology of Prejudice: Ingroup Love and Outgroup Hate?" *Journal of Social Issues* 55, no. 3: 429–444.

Brownstein, M. (2015) "Attributionism and Moral Responsibility for Implicit Bias." *Review of Philosophy and Psychology* 7, no. 4: 765–786.

———. (2016) "Implicit Bias, Context, and Character," in M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*, Oxford: Oxford University Press.

Brownstein, M., and Saul, J. (eds.). (2016a) *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*, Oxford: Oxford University Press.

———. (eds.). (2016b) *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics*, Oxford: Oxford University Press.

Correll, J., Park, B., Judd, C., and Wittenbrink, B. (2002) "The Police Officer's Dilemma: Using Race to Disambiguate Potentially Threatening Individuals." *Journal of Personality and Social Psychology* 83: 1314–1329.

Cortina, L. (2008) "Unseen Injustice: Incivility as Modern Discrimination in Organizations." *Academy of Management Review* 33: 55–75.

Cortina, L., Kabat Farr, D., Leskinen, E., Huerta, M. and V. Magley. (2011) "Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact." *Journal of Management* 39, no. 6: 1579–1605.

Dasgupta, N. (2004). "Implicit Ingroup Favoritism, Outgroup Favoritism, and Their Behavioral Manifestations." *Social Justice Research* 17, no. 2: 143–168.

Dasgupta, N., and Greenwald, A. (2001) "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice With Images of Admired and Disliked Individuals." *Journal of Personality and Social Psychology* 81: 800–814.

Dasgupta, N. and L. Rivera (2008) "When Social Context Matters: The Influence of Long-Term Contact and Short-Term Exposure to Admired Group Members on Implicit Attitudes and Behavioral Intentions." *Social Cognition* 26: 112–123.

Devos, T., and Banaji, M.R. (2005) American = white? *Journal of Personality and Social Psychology* 88, no. 3: 447.

Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012) "Beyond Prejudice: Are Negative Evaluations the Problem and Is Getting Us to Like One Another More the Solution?" *Behavioral and Brain Sciences* 35(6): 411–425.

Dovidio, J., and Gaertner, S. (2004) "Aversive Racism." *Advances in Experimental Social Psychology* 36: 1–51.

Dovidio, J., Kawakami, K., and Gaertner, S. (2002) "Implicit and Explicit Prejudice and Interracial Interaction." *Journal of Personality and Social Psychology* 82: 62–68.

Dunham, Y., Baron, A.S., and Banaji, M.R. (2006) "From American City to Japanese Village: A Cross-Cultural Investigation of Implicit Race Attitudes." *Child Development* 77, no. 5: 1268–1281.

Fazio, R. (1990) "Multiple Processes by Which Attitudes Guide Behavior: The MODE Model as an Integrative Framework." *Advances in Experimental Social Psychology* 23: 75–109.

———. (1995) "Attitudes as Object-Evaluation Associations: Determinants, Consequences, and Correlates of Attitude Accessibility," in R. Petty and J. Krosnick (eds.), *Attitude Strength: Antecedents and Consequences. Ohio State University Series on Attitudes and Persuasion, Vol. 4*, Hillsdale, NJ: Lawrence Erlbaum, pp. 247–282.

Fazio, R.H., Jackson, J.R., Dunton, B.C., and Williams, C.J. (1995) "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Personality and Social Psychology* 69, no. 6: 1013.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C., and Kardes, F.R. (1986) "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology* 50: 229–238.

Fazio, R., and Towles-Schwen, T. (1999) "The MODE Model of Attitude-Behavior Processes," in S. Chaiken and Y. Trope (eds.), *Dual-Process Theories in Social Psychology*, New York: Guilford Press, pp. 97–116.

Frankfurt, H. (1971) "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1: 5–20.

Gaertner, S.L., and Dovidio, J.F. (1986) *The Aversive Form of Racism*, San Diego: Academic Press.

Gawronski, B., and Bodenhausen, G. (2006). "Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change." *Psychological Bulletin* 132, no. 5: 692–731.

———. (2011) "The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions." *Advances in Experimental Social Psychology* 44: 59–127.

Gendler, T. (2008a) "Alief and Belief." *Journal of Philosophy* 105, no. 10: 634–663.

———. (2008b) "Alief in Action (and Reaction)," *Mind and Language* 23, no. 5: 552–585.

Glaser, J. and Knowles, E. (2008) "Implicit Motivation to Control Prejudice." *Journal of Experimental Social Psychology* 44: 164–172.

Green, A., Carney, D., Pallin, D., Ngo, L., Raymond, K., Lezzoni, L., and Banaji, M. (2007) "Implicit Bias Among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients." *Journal of General Internal Medicine* 22: 1231–1238.

Greenwald, A., and Banaji, M. (1995) "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review* 102, no. 1: 4.

Greenwald, A., McGhee, D., and Schwartz, J. (1998) "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74: 1464–1480.

Greenwald, A., and Pettigrew, T. (2014) "With Malice Toward None and Charity for Some: Ingroup Favoritism Enables Discrimination." *American Psychologist* 69, no: 7: 669–684. http://dx.doi.org/10.1037/a0036056

Hahn, A., Judd, C., Hirsh, H., and Blair, I. (2013) "Awareness of Implicit Attitudes." *Journal of Experimental Psychology-General* 143, no. 3: 1369–1392.

Hardin, C., and Banaji, M. (2013) "The Nature of Implicit Prejudice: Implications for Personal and Public Policy." *Behavioral Foundations of Public Policy*: 13–30.

Holroyd, J. (2012) "Responsibility for Implicit Bias." *Journal of Social Philosophy* 43, no. 3: 274–306.

Jacoby, L., and Dallas, M. (1981) "On the Relationship Between Autobiographical Memory and Perceptual Learning." *Journal of Experimental Psychology: General* 110, no. 3: 306.

Judd, C. M., Park, B., Ryan, C. S., Brauer, M., and Kraus, S. (1995). "Stereotypes and Ethnocentrism: Diverging Interethnic Perceptions of African American and White American Youth." *Journal of Personality and Social Psychology* 69, no. 3: 460.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., and Cohen, F. (2009). "The Unbearable Accuracy of Stereotypes," in T. Nelson (ed.), *Handbook of Prejudice, Stereotyping, and Discrimination*, New York: Psychology Press, pp. 199–227.

Kelly, D., and Roedder, E. (2008) "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3, no. 3: 522–540.

Levy, N. (2012) "Consciousness, Implicit Attitudes, and Moral Responsibility." *Noûs* 48: 21–40.

———. (2015) "Neither Fish Nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs* 49, no. 4: 800–823.

Levy, N., and Mandelbaum, E. (2014) "The Powers That Bind: Doxastic Voluntarism and Epistemic Obligation," in J. Matheson and R. Vitz (eds.), *The Ethics of Belief*, Oxford: Oxford University Press, pp. 15–32.

Madva, A. (2012) *The Hidden Mechanisms of Prejudice: Implicit Bias and Interpersonal Fluency*, Ph.D. diss., Columbia University.

———. (2016) "Why Implicit Attitudes Are (Probably) Not Beliefs." *Synthese* 193: 2659–2684.

———. ~~(Forthcoming a) "Biased Against De-Biasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle Against Prejudice." *Ergo*,~~

———. (~~Forthcoming b~~) "Implicit Bias, Moods, and Moral Responsibility."

Madva, A., and Brownstein, M. (2016) "Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind." *Noûs*. doi:10.1111/nous.12182.

Mandelbaum, E. (2013) "Against Alief." *Philosophical Studies* 165: 197–211.

———. (2016) "Attitude, Association, and Inference: On the Propositional Structure of Implicit Bias." *Noûs* 50, no. 3: 629–658.

McConahay, J. (1982) "Self-Interest Versus Racial Attitudes as Correlates of Anti-Busing Attitudes in Louisville: Is It the Buses or the Blacks?" *Journal of Politics* 44, no. 3: 692–720.

———. (1986) "Modern Racism, Ambivalence, and the Modern Racism Scale," in J. Dovidio and S. Gaertner (eds.), *Prejudice, Discrimination, and Racism*, San Diego: Academic Press, pp. 91–125.

McConahay, J., Hardee, B., and Batts, V. (1981) "Has Racism Declined in America? It Depends on Who Is Asking and What Is Asked." *Journal of Conflict Resolution* 25, no. 4: 563–579.

Mendoza, S., Gollwitzer, P., and Amodio, D. (2010) "Reducing the Expression of Implicit Stereotypes: Reflexive Control Through Implementation Intentions," *Personality and Social Psychology Bulletin* 36, no. 4: 512–523.

Nier, J. (2005) "How Dissociated Are Implicit and Explicit Racial Attitudes? A Bogus Pipeline Approach." *Group Processes & Intergroup Relations* 8: 39–52.

Nisbett, R. E., and Wilson, T. D. (1977) "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84, no. 3: 231.

Nosek, B. A., and Banaji, M. R. (2009) "Implicit Attitude," in T. Bayne, A. Cleeremans, and P. Wilken (eds.), *The Oxford Companion to Consciousness*, Oxford: Oxford University Press, pp. 84–85.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002) "Harvesting Intergroup Implicit Attitudes and Beliefs From a Demonstration Website." *Group Dynamics* 6: 101–115.

Nosek, B., Greenwald, A., and Banaji, M. (2007) "The Implicit Association Test at Age 7: A Methodological and Conceptual Review," in J. A. Bargh (ed.), *Automatic Processes in Social Thinking and Behavior*, Philadelphia: Psychology Press.

Nosek, B. A., and Hansen, J. J. (2008) "The Associations in Our Heads Belong to Us: Searching for Attitudes and Knowledge in Implicit Evaluation." *Cognition and Emotion* 22, no. 4: 553–594.

Olson, M., and Fazio, R. (2009) "Implicit and Explicit Measures of Attitudes: The Perspective of the MODE Model." *Attitudes: Insights From the New Implicit Measures*: 19–63.

Payne, B., and Gawronski, B. (2010) "A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going?" in B. Gawronski, and B. Payne (eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, New York: Guilford Press, pp. 1–17.

Plant, E. A., and Peruche, B. M. (2005) "The Consequences of Race for Police Officers' Responses to Criminal Suspects." *Psychological Science* 16, no. 3: 180–183.

Railton, P. (2009) "Practical Competence and Fluent Agency," in D. Sobel and S. Wall (eds.), *Reasons for Action*, Cambridge: Cambridge University Press, pp. 81–115.

———. (2014) "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics* 124, no. 4: 813–859.

Saul, J. (2012) "Skepticism and Implicit Bias." *Disputatio Lecture* 5, no. 37: 243–263.

———. (2013) "Unconscious Influences and Women in Philosophy," in F. Jenkins and K. Hutchison (eds.), *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press.

Scanlon, T. (1998) *What We Owe Each Other*, Cambridge: Harvard University Press.

Schacter, D. (1987) "Implicit Memory: History and Current Status." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13: 501–518.

Schuman, H., Steeh, C., and Bobo, L. (1997) *Racial Attitudes in America: Trends and Interpretations*, 2nd edition, Cambridge, MA: Harvard University Press.

Shiffrin, R., and Schneider, W. (1977) "Controlled and Automatic Human Information Processing: Perceptual Learning, Automatic Attending, and a General Theory." *Psychological Review* 84: 127–190.

Smith, A. (2005) "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115, no. 2: 236–271.

———. (2008) "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138: 367–392.

———. (2012) "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics* 122, no. 3: 575–589.

Thurstone, L. L. (1928) "Attitudes Can Be Measured." *American Journal of Sociology* 33, no. 4: 529–554.

Valian, V. (1998) *Why So Slow? The Advancement of Women*, Cambridge, MA: MIT Press.

———. (2005) "Beyond Gender Schemas: Improving the Advancement of Women in Academia." *Hypatia* 20: 198–213.

Velleman, D. (2000) "On the Aim of Belief," in *The Possibility of Practical Reason*, Oxford: Oxford University Press.

Wilson, T. D., Lindsey, S., and Schooler, T. Y. (2000) "A Model of Dual Attitudes." *Psychological Review* 107: 101–126.