Forthcoming in *Noûs*
Alex Madva, California State Polytechnic University, Pomona
Michael Brownstein, John Jay College of Criminal Justice, CUNY

Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind[1]

Abstract

How do cognition and affect interact to produce action? Research in intergroup psychology illuminates this question by investigating the relationship between *stereotypes* and *prejudices* about social groups. Yet it is now clear that many social attitudes are implicit (roughly, nonconscious or involuntary). This raises the question: how does the distinction between cognition and affect apply to implicit mental states? An influential view—roughly analogous to a Humean theory of action—is that "implicit stereotypes" and "implicit prejudices" constitute two separate constructs, reflecting different mental processes and neural systems. On this basis, some have also argued that interventions to reduce discrimination should combat implicit stereotypes and prejudices separately. We propose an alternative (anti-Humean) framework. We argue that all putative implicit stereotypes are affect-laden and all putative implicit prejudices are "semantic," that is, they stand in co-activating associations with concepts and beliefs. Implicit biases, therefore, consist in "clusters" of semantic-affective associations, which differ in degree, rather than kind. This framework captures the psychological structure of implicit bias, promises to improve the power of indirect measures to predict behavior, and points toward the design of more effective interventions to combat discrimination.

1. Introduction

[1] For their invaluable feedback and inspiration, we are grateful to Mahzarin Banaji, Francisco Gallegos, Katie Gasdaglis, Bertram Gawronski, Matthew Goren, Jules Holroyd, Benedek Kurdi, Nabina Liebow, Jonathan Phillips, Victoria Plaut, Laurie Rudman, Christina Stevens Carbone, Joseph Sweetman, Robin Zheng, and anonymous reviewers at *Noûs*. Versions of this essay were presented at Mahzarin Banaji's Social Cognition Lab; Victoria Plaut's Culture, Diversity, and Intergroup Relations Lab; and the 2014 meetings of the European Philosophical Society for the Study of Emotions, the Emotions and Emotionality Conference, and the Southern Society for Philosophy and Psychology. We are grateful for the helpful comments and questions we received during these presentations. This essay was written and revised with institutional support (for Alex Madva) from the Mellon Postdoctoral Fellowship at the University of California, Berkeley; Vassar College; and California State Polytechnic University, Pomona; and (for Michael Brownstein) from the American Academy of Arts and Sciences; the American Council for Learned Societies; and John Jay College of Criminal Justice/CUNY. Finally, thanks to the Leverhulme Trust for funding the "Implicit Bias and Philosophy" workshops at the University of Sheffield from 2011-2013, out of which this work grew.

Research on implicit bias demonstrates that individuals can act in discriminatory ways even in the absence of explicitly prejudiced motivations. Stereotypes about leadership ability, for example, might cause an employer to discriminate against women and minority applicants for a management position, even though the employer harbors no ill will toward these groups. Some philosophers and psychologists interpret these findings by drawing a distinction between cold, cognitive stereotypes and hot, affective-motivational prejudices. Consider, for example, Virginia Valian's (1998, 2005) account of how implicit gender stereotypes, or "schemas," impede the professional advancement of women:

> The explanation I focus on is social-cognitive; it examines the moment-by-moment perceptions and judgments that disadvantage women… the gender schemas we all share result in our overrating men and underrating women in professional settings, only in small, barely visible ways: those small disparities accumulate over time to provide men with more advantages than women. As I present it, the social-cognitive account is "cold." It is purely cognitive rather than emotional or motivational. It is intended to explain what goes wrong in environments where nothing seems to be wrong, where people genuinely and sincerely espouse egalitarian beliefs and are well-intentioned, where few men or women overtly harass women… cognitions do not automatically carry a set of emotions and motivations with them. (2005, 198-200)

Similarly, Elizabeth Anderson writes (2010, 44-5):

> The content of stereotypes is not inherently derogatory, nor are stereotypes typically generated by preexisting group prejudice. They are more a matter of "cold" cognitive processing than "hot" emotion… They are crude, typically unconsciously held heuristics that enable people to economize on information processing and react quickly to situations involving the object. As such, they are not inherently morally objectionable.

In these passages, Valian and Anderson apply the well-known distinction between explicit beliefs (stereotypes) and explicit feelings (prejudices) to implicit bias. Doing so may make a bitter pill easier to swallow. Emphasizing that implicit stereotypes are ubiquitous, morally innocuous, and coldly cognitive seems less likely to elicit defensive backlash than does leveling accusations of prejudice against those who explicitly avow egalitarian ideals. In broad form, Valian and Anderson extend a Humean picture of action to the implicit mind, such that cognitive and affective-motivational states make distinctive contributions to the production of behavior. But are "cold" implicit stereotypes truly distinct from "hot" implicit prejudices?

In research on intergroup relations, the distinction between explicit stereotypes and prejudices is widely accepted, although the nature of their interaction remains contested. Generally speaking, explicit stereotypes are characterized as beliefs or generalizations about the traits of social groups (Allport, 1954; Ashmore & Del Boca, 1981, Stangor, 2009), while explicit

prejudices are characterized as negative affective or evaluative responses toward members of social groups (McConahay & Hough, 1976; Dixon et al., 2012).[2]

The distinction between *implicit* stereotypes and *implicit* prejudices has been controversial, however, ever since Anthony Greenwald and Mahzarin Banaji (1995) called for study into so-called "implicit" social cognition. Some propose that the cognitive and affective components of implicit states represent two separate constructs, which reflect different mental processes and neural systems (Glaser, 1999; Judd et al., 2004; Correll et al., 2007; Amodio & Devine, 2006, 2009; Amodio & Ratner, 2011; Forbes & Schmader, 2010). We will refer to this as the *two-type model* of implicit mental states. Following David Amodio, Patricia Devine, and colleagues, we will refer to these two ostensible types as implicit stereotypes (ISs) and implicit evaluations (IEs).

The two-type model is intuitive and widely influential. Its defenders make three related claims:

1. ISs and IEs constitute two separate constructs, which reflect different mental processes and neural systems.
2. ISs and IEs predict different behaviors.
3. Interventions should combat ISs and IEs differently.

These three claims rightly emphasize the heterogeneity of implicit bias, an umbrella term referring broadly to various discriminatory forms of implicit social cognition.[3] These claims also make important advancements on some of the most pressing theoretical, empirical, and practical questions about implicit bias, namely:

1. Metaphysics of mind: what is the underlying nature of implicit bias?

---

[2] Prejudices can also be understood as a subset of "attitudes," in particular negative attitudes. In psychology, attitudes are understood as likings or dislikings, or more formally, as associations between a concept and an evaluation (Nosek & Banaji, 2009). On this definition, prejudices are dislikings, or associations between concepts of social groups, negative evaluations, feelings, and autonomic responses. Muddying the terminological waters is the fact that implicit prejudices, or implicit evaluations, are also often referred to in psychology as "implicit attitudes." When we use the term "implicit attitudes," we refer to the general construct representing all implicit mental states, rather than to (putative) implicit prejudices specifically.

While psychologists use the term "prejudice" to refer to distinctively affect-laden attitudes, some philosophers use terms like "prejudice" (and "racism") in different or broader ways. For example, Appiah (1990) defines "racial prejudice" in cognitive terms, as consisting in false and evidence-recalcitrant beliefs. See Faucher and Machery (2009) for discussion of various cognitive, behavioral, and affective accounts of prejudice and racism.

[3] Holroyd and Sweetman (2016) argue *that* implicit biases are heterogeneous, and call on philosophers (and scientists) to be more attentive to this fact. However, they remain explicitly agnostic about the scope and causes of this heterogeneity. We respond to their call in what follows, offering an account of the nature and causes of (some of) this heterogeneity. We focus in particular on the relationship between affect and cognition in implicit biases, in part by drawing inspiration from the empirical literature on explicit emotion and belief (§4). Holroyd and Sweetman raise concerns similar to ours regarding Amodio and colleagues' characterization of the affect-cognition relationship in the implicit mind (§3). See §3.3, §7, and notes #9 and #32 for further discussion of Holroyd and Sweetman's claims. We are grateful to an anonymous reviewer for *Noûs* for pushing us to clarify these points.

2. Ecological validity: how well do lab-based measures of implicit bias predict "real-world" behavior?
3. Practical application: how can implicit bias research help to combat discrimination and inequality?

Moreover, because implicit mental states arguably affect nearly all of our moment-to-moment behavior, understanding their cognitive and affective properties promises to illuminate the complex relationship between "reason" and "passion" in a new way.

In what follows, we argue that the evidence supports a *one-type model* of implicit mental states. On leading one-type models (e.g., Greenwald et al., 2002; Gawronski & Bodenhausen, 2006, 2011), ISs and IEs might be *conceptually* distinct, by virtue of their definitions, but they remain *functionally* the same. For example, we might *call* an automatic association of "black" with "athletic" a stereotype, and an automatic association of "black" with "bad" (or of "black" with a negative feeling) a prejudice, but both are, functionally speaking, the same kind of mental construct. (The relata differ, but the relation is the same.)[4]

After describing the two-type model in more detail (§2), we raise some empirical and conceptual concerns for it (§3). We situate these concerns in relation to leading theories of explicit prejudice and stereotyping (§4). We then attempt to integrate the insights of the two-type model with those of other research programs to consider how to modify indirect measures in order to best predict real-world behavior (§5). Stepping back from measurement and prediction, we advance some speculative hypotheses about the nature of implicit mental states, and situate these in relation to broader theories of cognitive architecture (§6). We propose specific criteria a one-type model must meet, according to which, roughly, all ISs are irreducibly evaluative and affect-laden while all IEs are "semantic," in the sense that they stand in co-activating associative relations with concepts and beliefs.[5] We conclude by explaining how our framework can improve the design of effective interventions to combat discrimination (§7).

2. The Two-Type Model of ISs and IEs

The most influential and empirically-tested version of the two-type model of implicit mental states is advanced in a series of papers by Amodio, Devine, and colleagues. In Amodio and Devine (2006), participants first took two different IATs: the standard evaluative race IAT (Eval-IAT), which measures associations between black and white faces and generic pleasant

[4] Our objections to the two-type model should not be construed as objections to *dual-system* theories that divide the mind into implicit and explicit systems (e.g., automatic, associative System 1 and reflective, logical System 2). Our concern is specifically with the viability of the cognitive/affective distinction *within* the domain of the implicit social mind. Note also that, although we speak here of implicit *associations*, much of what we say may be adapted to "propositional" accounts of implicit attitudes. See §6.1 for discussion.

[5] Social psychologists often use the term "semantic" to refer solely to associations between concepts in memory, and do not necessarily believe that implicit attitudes are "semantic" in the sense of purporting to represent the world as being a certain way. The "propositional" theorists discussed in in §6 claim that all (or most) implicit attitudes are semantic in the full-blooded representational sense, but they may or may not agree that all implicit affective responses also stand in co-activating associations with other concepts and attitudes.

and unpleasant words (e.g., "love" versus "evil"); and a novel Stereotyping IAT (Stereo-IAT), which measures associations between black and white faces and words associated with racial stereotypes of athleticism and intelligence. Amodio and Devine found that majorities of participants exhibited implicit stereotypical and evaluative biases, but that these biases were uncorrelated with each other. For example, a given participant might exhibit a strong association of blacks with unpleasant words on the Eval-IAT but only a weak association of blacks with sports-related words on the Stereo-IAT, or vice versa. Amodio and Devine interpreted this dissociation in terms of the hallowed distinction between cognition and affect, arguing that the Stereo-IAT reflects semantic associations between concepts and attributes, whereas the Eval-IAT reflects evaluative associations between stimuli and positive or negative affective responses.[6] Roughly, the Stereo-IAT measures "cold" implicit beliefs about racial groups while the Eval-IAT measures "hot" implicit likes, dislikes, and preferences.

Further evidence for the apparent IS/IE dissociation is found in a variety of behavioral measures. Amodio and Devine (2006) found that the Eval-IAT and Stereo-IAT uniquely predicted different kinds of behavior. Participants with strongly negative IEs of blacks sat farther away from a black interlocutor, and rated a black student as less likeable based on a written essay. Participants with strong ISs, on the other hand, described the black essay-writer in more stereotypical terms, and predicted that another black student would perform worse on an SAT-based task than on a sports-trivia task. Relatedly, Amodio and Hamilton (2012) found that manipulations influence ISs and IEs differently. Participants who believed that they were about to interact with a black person demonstrated more negative IEs and reported feeling greater anxiety than participants who expected to interact with a white person. However, participants' ISs were unaffected by their expectations to interact with black versus white interlocutors. These and other findings led Amodio and Ratner (2011, 143) to conclude that the IS/IE distinction paves the way for making "new and increasingly refined predictions." For example, researchers

[6] Earlier research into the two-type model measured more overtly evaluative stereotypes, like wealthy vs. poor or educated vs. ignorant (Rudman et al., 2001; Judd et al., 2004). The Stereo-IAT purports to avoid this confound. A reviewer for *Noûs* points out, however, that while prejudice is typically conceived as a positive or negative *feeling* toward social groups (characterized by distinctive forms of autonomic arousal; see note #1 and the paragraphs that follow this note), the Eval-IAT only measures associations between social groups and evaluative *concepts* of good/pleasant versus bad/unpleasant. Thus it may be possible, for example, to activate these evaluative concepts *without* activating any affective-bodily responses. This potential dissociation might make Amodio and Devine's (2006) studies a less-than-ideal illustration of the two-type model. However, a diverse body of evidence has led social psychologists to conceptualize the Eval-IAT as an effective *proxy* for genuine affective responses. For example, Phelps et al. (2000) found that participants' amygdala activation while looking at white versus black faces correlated with Eval-IAT scores as well as with eyeblink-startle responses, and Terbeck et al. (2012) found that beta blockers reduced implicit racial bias (see also, e.g., Amodio et al. 2003; Gawronski & Lebel, 2008; Smith and Nosek, 2011). While Amodio and Mendoza (2010, 359) acknowledge that the Eval-IAT *might* reflect *either* "aroused affective reactions or cognitive associations pertaining to emotional appraisals," they believe that "Amodio and Devine's (2006) findings suggest that, barring abnormal brain function…, measures of implicit evaluative bias may reflect affective processes" (367). That said, as is clear from the many criticisms in this paper, we agree that Amodio and Devine (2006) doesn't show what it purports to show. However, Amodio and Devine (2006) remains far and away the most cited exposition and defense of the two-type model, and bringing its limitations into sharper relief is important. We discuss below how several papers that purport to build on the findings of Amodio and Devine (2006) have not addressed the theoretical limitations of the original studies.

have explored the relative contributions of ISs and IEs to "shooter" bias, which involves an automatic tendency to "shoot" more unarmed black men than unarmed white men in a computer simulation. Glaser and Knowles (2008) found that a race-weapons Stereo-IAT predicted shooter bias, but that the Eval-IAT did not. They inferred that shooter bias is primarily caused by ISs (semantic associations of blacks with guns and criminality), rather than by any emotionally charged racial animosity (see also Judd et al., 2004; Correll et al., 2007).

In addition to generating novel behavioral data, two-type theorists have investigated the neural substrates of ISs and IEs. They have argued that amygdala-based learning, which underlies IEs, is functionally, phylogenetically, and anatomically distinct from neocortical-based learning, which underlies ISs. Moreover, amygdala-based learning does not depend on semantic associations, and neocortical-based learning can proceed in the absence of affect. "Affective versus semantic associations may be learned, modulated, and unlearned through very different processes," Amodio and Devine (2009, 201) write, "and therefore it may be important to measure and conceive of affective and semantic associations independently." These claims were investigated in studies using fMRI (Gilbert et al., 2012), EEG (Amodio, 2009a; Amodio et al., 2014), measures of cortisol reactivity (Amodio, 2009b), and startle eyeblink responses (Amodio et al., 2003).

This stream of research culminated in the memory-systems model (MSM; Amodio & Ratner, 2011) of implicit social cognition. MSM actually identifies three fundamental kinds of implicit states: implicit semantic associations (i.e., implicit stereotypes), implicit affective evaluations, and implicit motivations. As its name implies, MSM draws from research on distinct memory systems, tying each kind of implicit process to distinct neural mechanisms (roughly, IS = prefrontal cortex and temporal cortex; IE = amygdala and the autonomic nervous system; and implicit motivation = basal ganglia). Here we restrict our focus to ISs and IEs, and leave aside implicit motivations, for two reasons. First, the brunt of the research supporting a "multiple-type" model of implicit cognition has focused on investigating the IS/IE distinction alone. Second, it is often hard to distinguish implicit evaluation from implicit motivation. For example, Amodio and Devine (2006) used a seating distance measure to test the distinctive behavioral effects of IEs, but this measure arguably says as much about approach/avoidance motivations as it does about likes and dislikes. *Ceteris paribus*, we approach what we like and we come to like what we (repeatedly) approach. Our concerns about the two-type IS/IE taxonomy apply *mutatis mutandis* to the three-type taxonomy.[7]

[7] Amodio and Ratner (2011) describe the putative third system—the implicit motivational system—as an "instrumental" system, grounded in habit- and reward-learning. This might provide reason to respect—rather than elide—MSM's distinction between implicit motivation and implicit evaluation. For example, perhaps in a seating distance measure approach-behaviors are subserved by instrumental learning/motivational processes and avoidance-behaviors by fear-based conditioning/avoidance processes. On our view, however, disentangling the contributions of habit- and reward-learning from the contributions of fear-based conditioning to a decision about where to sit (and the motor routine of sitting) is as difficult as disentangling motivation from affect. In other words, seating distance—which Amodio and Devine (2006) took to reflect IEs—appears to us just as much (if not more) a measure of habit- and reward-learning as of fear, arousal, or discomfort. On this point we actually follow Amodio and Ratner (2011, 145), who write that, "Findings connecting approach–avoidance body movements with attitudes also

Two-type theorists have further claimed that the distinction between ISs and IEs is vital for combating discrimination. Amodio and Lieberman (2009) write, "our findings suggest that different prejudice reduction techniques are needed to target these two types of implicit bias."[8] Taking up this proposal, Forbes and Schmader (2010) studied the differential effects of retraining IEs versus ISs. First, undergraduate women were trained to implicitly like math by repeatedly associating the phrase "I like" and idiosyncratic things they liked (television, coffee, jogging) with math-related terms. A day later, these participants invested greater effort on a math test by spending more time and attempting to answer more problems. The effect of this IE-retraining was especially pronounced under stereotype threat, e.g., when the test was described as a "diagnostic measure of their natural mathematical ability" (2010, 14). This increase in effort, however, did not translate into answering more problems correctly (see also Kawakami et al., 2008). By contrast, participants who retrained their math-gender stereotypes by associating the phrase "women are good at" with math terms performed significantly better on math and working-memory tests the next day. Citing the two-type model, Forbes and Schmader write, "improving working memory in situations of stereotype threat necessitates a change in stereotypes, not a change in [evaluative] attitudes" (2010, 11).

Research into the two-type model has advanced our understanding of implicit bias in several ways. First, it supports the view, articulated by Valian and Anderson (§1), that blanket negativity toward outgroups, or a general preference for whites over blacks and men over women, is not, all by itself, responsible for all forms of automatic or unconscious discrimination. Second, it illustrates the importance of understanding the specific implicit mental states that predict specific judgments and behaviors; improving the predictive power of indirect measures should allay doubts some critics have raised about the prevalence of implicit intergroup bias (§5). Finally, the two-type model points to the general importance of understanding the underlying psychological nature of implicit bias (§6) in order to combat discrimination, and more specifically to the likelihood that successful interventions will have to target specific biases or kinds of biases (§7).

However, we suggest that the empirical evidence better supports a one-type model of implicit mental states.

3. Reconsidering the Two-Type View

In this section we raise five concerns about the two-type model. First, we question whether the prompts used in the Stereo-IAT are in fact evaluatively neutral. Second, we raise a conceptual concern about the claim that ISs and IEs cause different types of behavior, rather than

---

likely rely on instrumental and reward-learning systems." See also Kawakami's (2007) counterconditioning procedure which reduces racial IEs by having participants repeatedly *approach* black faces as well as *avoid* white faces. Thanks to a reviewer for *Noûs* for pushing us to clarify this point.

[8] Similarly, Amodio and Ratner (2011, 146) claim that "the MSM suggests that interventions can be tailored to the specific characteristics of the underlying memory systems." Ultimately, Amodio and colleagues seem to think that distinctive interventions that target these separate systems should be combined into one more complex intervention. For example, Amodio and Lieberman (2009, 28) write, "it may be best to use both types of reduction techniques in conjunction in order to most effectively diminish bias." We return to this point in §7.

making different types of causal contribution to (one and the same type of) behavior. Third, we argue that the central dissociation between the Eval-IAT and Stereo-IAT, as well as the behaviors these measures predict, falls short of supporting a generalized two-type distinction, in contrast to a less theoretically weighty distinction between *particular* implicit mental states. Fourth, we sketch an experimental manipulation that could produce better evidence for the two-type model, but has not yet been tested. Finally, we argue that the current neural data does not support a two-type over a one-type model.

3.1 Reconsidering the Stereo-IAT

The Stereo-IAT is designed to isolate the cold, cognitive core of ISs. The measure includes only putatively positive and "neutral" words regarding intelligence and athleticism. For most participants, white faces were more easily associated with positive words like "genius" and "smart" and (putatively) neutral words like "math" and "read," while black faces were more easily associated with positive words like "agile" and "rhythmic" and neutral words like "run" and "basketball." Moreover, the categorization task instructs participants to sort the words according to whether they are "mental" or "physical," which the authors claim constitute "relatively neutral" ways of grouping the categories (Amodio & Devine, 2006, 654), rather than using more overtly evaluative groupings like "smart" and "athletic." To the extent that "mental" and "physical" *are* evaluative terms, Amodio and colleagues claim that they have a similar positive valence, unlike in the Eval-IAT, in which the two categories, "good" and "bad," clearly differ in evaluative standing.[9] Thereby, the Stereo-IAT is claimed to be "only conceptual, but not evaluative" (Amodio & Hamilton, 2012, 1275).[10]

[9] A reviewer for *Noûs* suggests that the phrase "relatively neutral" means "neutral relative to each other, not that the attributes are independently 'close enough' to neutral." We believe that Amodio and colleagues vacillate between both claims. Sometimes they characterize the terms as neutral in the sense of lacking evaluative content altogether (which is especially apparent in light of their attempt to avoid evaluative stimuli, as we discuss in §3.1.1), and other times they characterize the terms as neutral relative to each other, in the sense of sharing a similar (slightly) positive valence. We address both claims in what follows.

[10] Holroyd and Sweetman argue that Amodio and colleagues' experiments do not adequately operationalize the cognitive/affective distinction. First, they point out that, "those terms that are supposed to indicate semantic associations in Amodio and Devine's studies (intelligence, athleticism, smart) have both evaluative and semantic content (the characteristics are positive, good, features)" (2016, 94). We are sympathetic with Holroyd and Sweetman's general concern here and expand on similar concerns in the following sections. However, Amodio and colleagues assert that the sheer presence of slightly valenced terms in the Stereo-IAT need not be problematic—as long as the direction (positive vs. negative) and strength of the valences are approximately matched. Holroyd and Sweetman also point out that the stimuli in the Eval-IAT "incorporate both semantic content and affect," and recommend finding ways to experimentally induce affective responses without relying on semantically meaningful stimuli. Amodio and colleagues would argue that the semantic content of terms on the Eval-IAT are highly diverse (positive terms were "honor, lucky, diamond, loyal, freedom, rainbow, love, honest, peace and heaven"). That is, these stimuli are matched not for semantic content, as on the Stereo-IAT, but for valence. Although we think Holroyd and Sweetman's arguments do not conclusively establish that Amodio and colleagues have not properly operationalized the distinction they're aiming at, we are sympathetic with the point that the semantic nature of prejudice is being overlooked. We say more about this in §4.

3.1.1. Our first set of questions has to do with the specific stereotypes in question. Amodio and Devine (2006) explain that they also pre-tested other stereotypical associations, including

> sets of target words related to poor (vs. wealthy), hostile (vs. friendly), and lazy (vs. motivated). In each case, however, the stereotype was strongly related to evaluation (e.g., poor is negative and wealthy is positive), and therefore these were not suitable for examining the independence of implicit evaluation and implicit stereotyping. (654n2)

As we see it, the sheer difficulty of finding stereotypes that were "relatively neutral" is significant. If most prevalent stereotypes tend to be strongly evaluative, this suggests that any genuinely neutral stereotypes are outliers. It also strikes us that intelligence and athleticism *are* clearly evaluative, as are many of the particular terms used in the Stereo-IAT, like "educated" and "genius."[11] Warriner and colleagues (2013) compiled affective ratings for thousands of words by asking participants to rank their valence on a scale of 1 to 9 (negative to positive), and found that many of these terms have, on average, a strong positive valence (7.65 for "intelligence," 7.48 for "educate," 6.73 for "read", etc.).[12] Indeed, all trait dimensions likely have an evaluative component (Rosenberg and Sedlak, 1972). Consider the IS-retraining in Forbes and Schmader (2010). Participants were not merely trained to associate math-related terms with women, but to associate math-related terms with the overtly evaluative phrase, "women are good at" (and to associate language-related terms with the phrase "men are good at"). Rather than providing evidence for the general claim that "hot" affective-motivational dispositions are less relevant to test performance than "cold" cognitive dispositions, Forbes and Schmader's research shows how one evaluative disposition (feeling good in a given domain, and perhaps an attendant sense of confidence and "belonging") is more important than another evaluative disposition (liking and investing effort in that domain) when it comes to countering stereotype threat. We return to this point in §3.4.

Forbes and Schmader's studies also help to illuminate how the precise evaluative significance of ISs will vary across individuals and contexts. For women taking a math test, the stereotype that men are "good" at math may have a negative valence, as may the stereotype that women are "good" in domains unrelated to math. Often, a trait like intelligence or being "good at" some activity has a positive valence when it is attributed to oneself or one's ingroup, but a negative valence when attributed to an outgroup (Degner and Wentura, 2011). Moreover, traits like "intelligent" and "athletic" likely have negative valence in certain contexts. An individual

[11] Historically, the stereotypical division of groups into the mental and physical formed a cornerstone of defenses of social hierarchy and slavery. In the *Politics*, Aristotle claimed that the putative physical talents of ethnic outgroups ("barbarians"), in conjunction with their intellectual inferiority, made it natural and just for them to be ruled by the intellectually superior Greek men.

[12] For ease of exposition, we focus on participants' ratings of valence ("the pleasantness of a stimulus"), although they also rated arousal ("the intensity of emotion provoked by a stimulus"; Warriner et al. 2013, 1191). The pattern of findings is the same, e.g., "intelligence" is, on average, positively valenced (7.65) and emotionally arousing (6.32). These ratings of arousal are all based on self-report, and it is possible that neural or behavioral measures of arousal (e.g., of skin conductance or startle reflex) might tell a different story. See §3.1.2.

who self-identifies as a "jock" and believes that being a "brainiac" is inconsonant with this athletic self-concept might evaluate certain sorts of intelligence negatively in certain contexts (see, e.g., Greenwald et al., 2002). By the same token, an individual who self-identifies as intellectual might come to disdain athleticism and "dumb jocks." In this vein, while Warriner and colleagues (2013) found positive *average* valences for terms like "athlete" (6.21) and "geek" (5.56), they also found large standard deviations (above 2), suggesting that *some* individuals ranked these as strongly positive while others ranked them as strongly negative.[13] It seems, then, not only that these terms *do* tend to be significantly valenced, but that their valence may vary significantly across particular participants in particular contexts.

3.1.2. Our second set of questions regards the apparent absence of affect on the Stereo-IAT, which could reflect: (a) that the intensity of a causally significant affective response is too low or subtle for the measurement tool; (b) that the relevant type of affective response is not being measured.

Regarding (a), consider recent work on "micro-valences." Lebrecht and colleagues (2012) propose that valence is an intrinsic component of all object perception. On their view, the perception of "everyday objects such as chairs and clocks possess a micro-valence and so are either slightly preferred or anti-preferred" (2). If the most primitive non-social elements of visual processing are pervasively influenced by valence, then the same is likely to be true of implicit stereotyping, albeit perhaps at a level too low to be captured by the Stereo-IAT.[14] For example, Hugenberg and Bodenhausen (2003, 2004) found that angry faces are more likely to be seen as black and that dark-skinned faces are more likely to be seen as angry. In these and other cases (Amodio, 2009a; Ofan et al., 2011; Ratner et al., in press; Correll et al., 2015), social perception seems to be intrinsically affective, and social affect intrinsically cognitive. A striking demonstration of the evaluative significance of stereotypes is Flannigan and colleagues' (2013) finding that men and women in counterstereotypical roles (e.g., men nurses, women pilots) are "implicitly bad"—the sheer fact that these stimuli are counterstereotypical leads individuals to implicitly dislike them.

[13] *All* of the terms used in the Stereo-IAT have a significantly positive or negative valence according to Warriner et al. (2013). Some related terms, e.g., "mathematics," have a neutral valence on average (4.98), yet even the valence of these terms exhibits large standard deviations. Moreover, neutral words may have microvalences (§3.1.2). Also noteworthy is that the standard deviations for ratings of arousal tend to be significantly larger than for valence, suggesting that the emotional intensity, as well as the valence, of these terms is highly variable across participants and contexts.

[14] Research on micro-valences is relatively new and further investigation is required to vindicate it. In particular, it will be important to determine the extent to which micro-valences are related to actual autonomic responses, and not merely to cognitive appraisals of preference. We introduce this research only as an example to make salient the live possibility of causally efficacious affective responses that are too small or subtle to be detected by the Stereo-IAT. Our claim is that Amodio and colleagues have to *rule this possibility out* in order to claim that the Stereo-IAT delivers evidence of absence of affect, instead of just an absence of evidence of affect. Moreover, whether the effect of micro-valences is due to the penetration of early visual processing, or, e.g., to a shift in attention (Roskos-Ewoldsen and Fazio, 1992) is an open question (Siegel, 2012).

Regarding (b), consider research that suggests that specific types of affective response are triggered by specific stimuli. For example, Tapias and colleagues (2007) found that priming participants to think of African-American men triggers anger, while priming thoughts of gay men triggers disgust.[15] It could be, then, that the affective responses specific to the Stereo-IAT have not been measured. While Amodio and Hamilton (2012) found that inducing social anxiety strengthened participants' racial IEs but not ISs, perhaps inducing fear of personal safety might do so (because a member of a physically imposing social group is perceived as more of a threat).[16] In this vein, Rudman and Ashmore (2007) found that non-black participants who reported having been excluded, given "the finger," physically threatened, or assaulted by blacks subsequently exhibited stronger ISs but not stronger IEs.[17] This runs counter to the general claim that the emotions "elicited in real-life intergroup interactions… [have] more direct implications for affective and evaluative forms of implicit bias than for implicit stereotyping" (Amodio & Hamilton, 2012, 1273). It also poses a concern for the proposal that "the most prominent emotional response" in interracial interactions is anxiety, rather than hostility, threat, or guilt (Amodio & Hamilton, 2012, 1273; for further discussion, see Amodio and Mendoza, 2010, 360-1). Which emotional response takes prominence, we propose, will vary with context (Cottrell & Neuberg, 2005; Cesario & Jonas, 2014).

3.2 Conceptual Questions about Behavioral Prediction

The two-type model claims that ISs and IEs predict different types of behavior. However, on a traditional Humean view, cognitive and non-cognitive states are not distinguished because they cause different behaviors, but because they make distinctive causal contributions to behavior.[18] Beliefs are thought to cause behaviors only in conjunction with desires (and other beliefs). The very same belief might lead to radically different behaviors depending on the individual and the context. Suppose Lou and Nancy simultaneously form the belief that Bonnie is a drug dealer. Lou wants to buy drugs, and so approaches Bonnie, but Nancy wants drug dealers to go to prison, and so avoids Bonnie and calls the police. *A priori*, one would not predict that distinctive spheres of behavior would be uniquely predicted by cognitive versus non-cognitive attitudes about drug dealers.

In a typical experimental context, one might contrast the effects of one belief with the effects of *another belief*, while holding fixed as many as possible of participants' other attitudes. In Amodio and Hamilton (2012), for example, participants were led to believe that they were about to interact either with a white person or with a black person. Then the effects of these

[15] See also Dasgupta and colleagues (2009).

[16] On situation-specific racial bias related to fear, see, e.g., Cesario et al. (2010), who found that priming participants with photos of black men activated flight-related, avoidance responses when done in an open field (a context which affords flight), but aggressive, fight-related responses when done in an enclosed booth (a context which does not afford flight).

[17] Rudman and Ashmore used an *evaluative* measure of ISs, rather than the putatively non-evaluative Stereo-IAT used by Amodio and colleagues. We return to this point in§5.

[18] See §6, Gendler (2008a,b), and Brownstein and Madva (2012a,b) for discussion of alternatives to Humean (belief-desire) psychology.

different beliefs were contrasted. In Forbes and Schmader (2010), liking math was contrasted with disliking math, and reinforcing math-gender stereotypes was contrasted with undermining math-gender stereotypes—and thereafter both manipulations were simultaneously investigated in a 2x2 analysis of variance.

We find it more difficult to interpret studies that purport to differentiate the behavioral effects of cognitive versus non-cognitive attitudes about a social group. To do so, one would have to hold fixed the intentional content being represented across the two types of mental state (e.g., do participants *believe* that whites are more intelligent than blacks, or do they merely *want* this to be the case?). However, the contrast between, for example, the *stereotype that one group is more athletic* with the *evaluation that one group is more likeable* does not control for the intentional content in this way. It is akin to claiming that Lou's belief that Bonnie is a drug dealer predicts different behavior compared with Nancy's desire to see drug dealers go to prison. This potential confound between attitude and content makes it difficult to infer underlying differences between types of implicit attitude, as opposed to differences between specific attitude-content combinations.

This concern is especially salient for Gilbert and colleagues' (2012) investigation into the neural substrates of ISs and IEs. Participants repeatedly saw either a pair of white faces or a pair of black faces, and were asked one of two questions about the faces. ISs were measured with the question, "who is more likely to enjoy athletic activities?" while IEs were measured with the question, "who would you be more likely to befriend?" These highly specific questions differ in a number of ways besides stereotypical-trait attribution versus social liking. To ensure that differential brain activation does not simply reflect the activation of two distinct concepts (athleticism versus friendship), other questions could be asked, such as, "who is more likely to enjoy math?" to measure ISs and "who is more outgoing/likeable/pleasant?" to measure IEs. To better isolate the activation patterns of judgment about friendship, participants might have been asked, "who has more friends?" Moreover, the first question in the study merely requires deciding which of two people enjoys an activity more while the second invokes the self-concept (who would "you" befriend?). This presents an additional confound. Answering this friendship question likely requires assessing one's own traits, comparing oneself with another, activating a memory search of one's friends, imagining social interactions, and so on. We are also interested to know why the IS question asks about "enjoying" athletic activities, instead of directly asking, for example, "who is more athletic?" Gilbert and colleagues suggest that athleticism is a relatively non-evaluative stereotype, but we think it is more accurate to say that they are asking a relatively non-evaluative question about athleticism (and a relatively evaluative question about friendship). We suspect that the affective-motivational significance of physicality stereotypes could be better revealed by other questions, such as, "who would you pick to be on your sports team? Who would you rather compete against? Who would win in a fight? Who would you rather fight?"

3.3 Reconsidering the Double Dissociation

The double dissociations observed in studies that appear to support the two-type model may not, therefore, reflect wholly separate cognitive and affective systems. These dissociations may reflect only that *particular* ISs are dissociable from *particular* IEs, in the same way that particular ISs (about, for example, athleticism and intelligence) are dissociable from *each other*. Holroyd and Sweetman suggest that it would be more parsimonious to explain these dissociations solely in terms of differences in "the content of the associations, without recourse to distinct underlying mechanisms" (2016, 93). We agree. However, appealing to parsimony here suggests that the content-based and mechanism-based explanations are equally consistent with the data. Yet there is, in fact, substantive evidence that specific stereotypes (which, according to the two-type model, differ only in content, and not in underlying mechanisms) are frequently dissociated.

Such evidence is often overlooked. For example, Devine (1989) argued that although a majority of Americans have come to personally disavow racial stereotypes, a consensus remains regarding which stereotypes Americans *perceive* to be prevalent and "culturally shared," and therefore which stereotypes are harbored at the implicit level. Nosek and Hansen (2008) note, however, that:

> In retrospect, data from Devine (1989) also showed variability in perceptions of stereotypes. In the first study, participants reported the cultural stereotype about African Americans. Far from consensus, not a single characteristic was generated by all participants. In fact, most qualities (e.g., low intelligence, uneducated, sexually perverse) were mentioned by between just 20% and 50% of the respondents indicating substantial variability in the perception of cultural stereotypes…[19]

Indeed, Amodio and Hamilton (2012) themselves found that white participants who implicitly stereotype black people as unintelligent do not necessarily also stereotype them as athletic. Evidently, racial bias on the Stereo-IAT "was primarily driven by the activation of the 'Black-unintelligent' stereotype" rather than by the black-physical, white-unphysical, or white-intelligent stereotypes (1276). The Stereo-IAT seems to reflect a particular (pernicious and negative) racial stereotype, which is dissociable from athleticism stereotypes, rather than a general disposition to associate racially typical faces with all culturally prevalent stereotypes.[20]

Given that some individuals' Stereo-IAT scores primarily reflect a difficulty in associating blacks with intelligence, rather than a corresponding ease in associating blacks with athleticism, we predict further behavioral dissociations, for example, that some individuals

[19] We are, of course, not predicting that there will *never* be pervasively shared stereotypes. The "headline" about stereotypes should read *both* that stereotypes are often shared among diverse individuals *and* that there are nevertheless many individual- and group-based differences in stereotyping.

[20] Evidence of double dissociations among particular ISs would be acutely troublesome for the two-type theory, given that, on that theory, double dissociations of putative ISs and IEs indicate a difference in underlying mechanisms. If particular ISs double-dissociated, would we then be forced to conclude that these ISs were supported by separate underlying mechanisms too?

would use stereotypical terms to describe a black writer but not a black athlete, and vice versa. We also expect that particular implicit evaluations are dissociable (§3.1.2; §§4-5). In other words, while two-type theorists claims that ISs and IEs represent two broad classes of implicit bias, which are in turn made up of multiple "species" of specific ISs and IEs, their data are consistent with there being one class of implicit bias, some "species" of which are less affectively intense than others (§6).

3.4 Reconsidering Manipulations that Affect IEs but not ISs

Over and above IE/IS double dissociations, two-type theorists also argue that one kind of implicit mental state can be manipulated without influencing the other. In principle, such manipulations could produce powerful evidence for the two-type model. However, the experimental manipulation most diagnostic for the two-type model remains untested.[21] While Amodio and Hamilton (2012) manipulated IEs without finding evidence of effects on specific ISs, no experiment we know of has manipulated ISs without influencing IEs. Only the latter finding would pose a problem for leading one-type models.

For example, on Gawronski and Bodenhausen's (2006) one-type theory, evaluations are modeled not as a separate type of mental state, but as the net valence of activated ISs, which are specific, semantic associations between groups and traits (e.g., black-criminal, black-unintelligent). If the temporarily salient ISs toward a social group are predominantly negative in valence, then an individual will have a negative evaluative response to members of that group (e.g., as measured by an Eval-IAT). On this model, it is possible to manipulate net evaluations without changing any particular IS. A manipulation that induces social anxiety and strengthens racial IEs may have no measurable effect on any *specific* IS (e.g., neither strengthening nor weakening black-athletic or black-unintelligent associations).

The central question, then, is whether ISs can be manipulated without influencing IEs.[22] We find no evidence for this possibility, but some evidence to the contrary. Glaser (1999) and Gawronski and colleagues (2008) both found that retraining ISs led to changes in IEs (see §7).

Consider again Forbes and Schmader's (2010) claim that changing ISs is more important for countering stereotype threat than changing IEs. None of their studies actually measured

---

[21] Thanks to Bertram Gawronski for pushing us to elaborate on this point.

[22] In fact, a reviewer for *Noûs* points out that even this type of manipulation might fail to be decisive, under certain conditions. For example, a manipulation might de-associate a social group with one stereotype but replace it with another stereotype that shares the same valence. Because the net valence of associated stereotypes would remain the same, a one-type theorist would not predict that this sort of change in ISs would bring about a change in IEs. But the cases we have in mind are simpler: the two-type theory predicts that it should be relatively easy to change only one IS, or one "cluster" of ISs (e.g., associations of blacks with athletic terms) without changing any IEs. A one-type theorist would predict that this intervention would change IEs, although the effect could certainly be canceled out in numerous ways, such as in more complex cases when a *further* IS is also changed simultaneously (or when participants' affective states are manipulated in some further away, such as by taking beta blockers). Consider, by analogy, the question whether a certain chemical is poisonous. A key test is what happens when the chemical is imbibed in isolation, although it might be true that if the chemical is imbibed simultaneously with an antidote, there may be no poisonous effect at all. Changing two ISs simultaneously is analogous to imbibing the poison and the antidote simultaneously.

whether changes in ISs occurred *without* corresponding changes in IEs. Would training women to associate the phrase "women are good at" with math-related words influence their performance on a Math Eval-IAT? We expect so. A follow-up study could also measure the effects of IS-retraining on reported feelings of confidence, "belonging," and stereotype endorsement. If IS-retraining leads to increased confidence and sense of belonging, but not changes in explicit endorsement of math-gender stereotypes, this would suggest that the primary effect of this putatively cold-cognitive retraining procedure was affective after all. We say more about strategies for changing IEs and ISs in §7.

3.5 Reconsidering the Neural Data

Amodio and colleagues have been pioneers in the burgeoning integration of social psychology and neuroscience. The neuroscientific data, however, does not clearly favor the two-type model. Evidence for the traditional identification of the amygdala with affect and the prefrontal cortex with cognition is mixed (e.g., Salzman & Fusi, 2010). Reviewing the literature demonstrating the role of affect in the processing of conscious experience, language fluency, and memory, Duncan and Barrett (2007) argue that "there is no such thing as a 'nonaffective thought.' Affect plays a role in perception and cognition, even when people cannot feel its influence…" and conclude that "the affect-cognition divide is grounded in phenomenology." On this view, the cognitive/affective distinction is, ultimately, an empirically unsupported posit of folk psychology, which persists primarily because it derives intuitive support from qualitative experiences of emotion. We typically experience affect only when it is especially *intense*, but low-level affect exerts a pervasive influence on ostensibly cognitive processes (§3.1). Duncan and Barrett focus primarily on non-social forms of cognition, but if there is no such thing as non-affective non-social cognition, there is surely no such thing as non-affective social cognition either. Implicit social-cognitive processes are pervasively shaped by affect (Mitchell, 2009; Contreras et al., 2012).[23]

Some two-type theorists have expressed similar concerns. Amodio suggests that received opinion about the amygdala "as the fear center, and often as the locus of emotion broadly" (2010, 710) has not been confirmed. Instead, the amygdala seems to reflect:

[23] While these considerations speak against the existence of "nonaffective thought," a further question is whether they speak against "noncognitive affect." Is it possible to activate affective responses without, say, activating any concepts and beliefs? A reviewer for *Noûs* suggests that one reason to think so is that human beings likely share IEs (i.e., automatic affective responses to stimuli) with non-human animals who are incapable of processing semantic information, and so will not form associations of IEs with concepts. We are sympathetic to the idea that non-human animals may share these mental states (for a one-type defense of this claim, see, e.g., Gendler, 2008), but we are less certain than the reviewer that animals who share certain kinds of IEs with human beings lack the relevant semantic or conceptual capacities (or nonlinguistic analogues of these capacities). Animals are certainly capable of associative learning, and may very well associate affective responses with concepts or "proto-concepts." We suspect that phenomena structurally similar, if not identical, to associative priming occur in many animals. How to conceive of these analogues to the semantic components of IEs in non-human animals is a fascinating question that clearly requires further thought, however. More generally, how to extend research on implicit social cognition to non-human minds, and vice versa, is an important area for future research (see, e.g., Gawronski and Cesario, 2013). See also Duncan and Barrett (2007) for further arguments that affect is in fact a *type* or *aspect* of cognition.

> a diverse set of processes involved in attention, vigilance, memory, and the coordination of both autonomic and instrumental responses… Furthermore, the amygdala comprises multiple nuclei associated with different functions, connected within an inhibitory network …. These subnuclei cannot be differentiated with current neuroimaging methods, and thus it is very difficult to infer the specific meaning of an amygdala activation using fMRI… (2010, 710-1)

We are very sympathetic with these notes of caution about interpreting amygdala activation, but we find the caution expressed here somewhat inconsonant with claims Amodio and other two-type theorists make elsewhere. Amygdala activation was posited as the hallmark of IEs, whereas it now seems to serve a variety of ostensibly cognitive, evaluative, and motivational functions, including processes of attention and memory (as noted by Duncan and Barrett) that are surely relevant to learning and unlearning long-term semantic associations.[24]

A final note of caution is also well-articulated in Amodio (2010). The aptly titled, "Can Neuroscience Advance Social Psychological Theory?" explains that the exploratory enterprise of mapping psychological constructs onto brain regions is much more tractable for "low-level" than "high-level" processes. Low-level processes, such as edge-detection in vision, map much more directly onto specific physiological processes than high-level processes such as self-concepts, trait impressions, political attitudes, and social emotions like romantic love (698). The article concludes, "it is often advisable to interpret brain activity in terms of lower-level psychological processes that then contribute to the higher-level processes that are typically of interest to social psychologists" (708). We agree. But we suspect that implicit attitudes are a high-level construct, more akin to romantic love and the self, and so are unlikely to be localized in distinctive brain regions. Amodio does not articulate what ultimately distinguishes low- from high-level constructs; however, if visual edge-detection exemplifies a low-level construct and romantic love exemplifies a high one, then some likely factors underlying this distinction are the construct's functional and operational complexity, neural localization, malleability, and sensitivity to context, goals, social pressure, and cultural influence. Unlike edge-detection, but like romantic love, implicit associations are functionally complex, neutrally distributed, malleable, socially learned, cross-contextually unstable, and amenable to various forms of self-regulation and change. Perhaps relatively affective and semantic components of implicit attitudes are subserved in greater or lesser degrees by specific neural regions or networks, but these should be viewed as components that underlie the (high-level) construct of interest (i.e., implicit attitudes). We return to this point in §6, where we elaborate on the internal structure and complexity of implicit attitudes.

4. Explicit Cognition and Emotion

[24] Amodio himself has entertained several alternative hypotheses about the IE/amygdala relationship (Amodio & Lieberman, 2009; Amodio, 2010; Amodio & Ratner, 2011).

The two-type model draws inspiration for the claim that IEs and ISs are independent constructs from research on explicit social cognition, but some leading theories about explicit stereotypes and prejudices, such as the Stereotype Content Model (SCM; Fiske et al., 2002) and the "threat-based" model of intergroup prejudice (TBM; Cottrell & Neuberg, 2005), emphasize their interrelations. SCM argues that prevalent stereotypes about social groups tend to form around two central dimensions: warm versus cold, and competent versus incompetent. Cognitive judgments about both of these dimensions are significantly influenced by affective and motivational processes. For example, the motivations to protect one's self-esteem and maintain the status quo play a large role in leading individuals to judge that some groups are warm but incompetent (e.g., the elderly, housewives), while others are competent but cold (e.g., Asians, Jews, businesswomen). Fiske and colleagues (2002, 879) write:

> different combinations of stereotypic warmth and competence result in unique intergroup emotions— prejudices—directed toward various kinds of groups in society. Pity targets the warm but not competent subordinates; envy targets the competent but not warm competitors; contempt is reserved for out-groups deemed neither warm nor competent.

This model shows how stereotypes take on specific sorts of evaluative significance for specific individuals in specific contexts. It explains phenomena like "benevolent sexism," which is the tendency to compensate for negative gender stereotypes with "warm" feelings (Dardenne et al., 2007). The cognitive stereotype that a group is warm is likely to be related to the judgments of likeability and approach behaviors (e.g., seating distance) that the two-type model identifies as uniquely predicted by evaluative attitudes. While SCM researchers focus on self-reported stereotypes, studies using indirect measures show the applicability of SCM to implicit cognition.[25] Ebert (2009; see also Ebert et al., 2014) found that implicit associations of women with warmth were strongly correlated with implicit liking (evaluation) of women. Strikingly, Ebert also found that implicit liking of women significantly correlated with implicit associations of women with competence (r = .59). In other words, Ebert found in the case of gender exactly what Amodio and Devine (2006) *didn't* find in the case of race: putatively generic IEs correlated with putatively non-evaluative ISs about both warmth and intelligence.[26] Similarly, Agerström and colleagues (2007) found that implicit dislike of Arabs on an Eval-IAT significantly correlated with implicit associations of Arabs with *in*competence on a Stereo-IAT (r = .52).

SCM researchers have been clear in acknowledging the irreducibly evaluative nature of stereotypes, but many have not appreciated the ways in which prejudices are also "semantic." In

[25] We are not concerned to defend the universal validity of SCM as a theory, but think it has much to offer.

[26] A note about terminology: in what follows, we sometimes refer to "IEs" and "ISs" as if they were distinct entities. In these instances, we are merely following the linguistic conventions of the researchers being cited. Publications that report similarities and differences between putative IEs and putative ISs would be, on our terms, better characterized as reports on two or more particular clusters of semantic-affective associations, or as separate aspects of one cluster, rather than as studies of distinctive types of mental kind. Thanks to an anonymous reviewer for *Noûs* for pushing us to clarify this.

response, Cottrell and Neuberg developed TBM, arguing that "the traditional view of prejudice—conceptualized as a general attitude and operationalized via simple evaluation items—is often too gross a tool for understanding the often highly textured nature of intergroup affect"(2005, 787). Social affect comes in all shapes and sizes: fear, disgust, pity, and envy, to say nothing of moral emotions like resentment, admiration, praise, and blame. Thus, Cottrell and colleagues (2010) found that self-reported intergroup emotions such as resentment, pity, disgust, and fear predicted policy attitudes about gay rights and immigration, while generic intergroup dislike and "negative feelings" did not.

Like SCM, TBM is based on self-report. However, the "rich texturing of emotions" TBM describes (Cottrell & Neuberg, 2005, 771) likely affects implicit intergroup biases as well. For example, Stewart and Payne (2008) found that racial weapon bias could be reduced by rehearsing the plan to think the word "safe" upon seeing a black face. This is, on its face, both a semantically relevant and a highly affect-laden word to think in this context (in contrast to the more cognitive terms, "quick" and "accurate," which failed to reduce weapon bias). This suggests that the potential for affect to influence weapon bias is not via a generic dislike, as if people will be more likely to "shoot" anything they dislike. A more relevant emotion is clearly *fear*. Thinking the word "safe" likely activates both thoughts and feelings that interfere with the association of black men with weapons.[27]

We propose, therefore, that the aim of refining indirect measures to make increasingly precise behavioral predictions may be well served by incorporating Cottrell and colleagues' insight that intergroup affect is not simply a matter of generic likes or dislikes (i.e., the mere net valence of multiple associations) of social groups. The Eval-IAT may be too coarse-grained to capture, let alone differentiate among, the many affect-laden responses most relevant to social behavior. In §5, we make further suggestions for how the insights of SCM and TBM might be used to enhance the predictive validity of indirect measures.[28]

5. Toward More Predictive Validity

On our view, the principal virtue of the Eval-IAT is that it measures generic likings and preferences, and as such can predict a wide range of behaviors. Its principal vice, however, is that effect sizes are likely to be small. Like a jack of all trades and a master of none, the Eval-IAT should predict many behaviors at the cost of predicting few of them particularly well. This

[27] See also Correll et al. (2007).

[28] A reviewer for *Noûs* suggests that perhaps existing measures, like the race-weapons IAT, do in fact already capture the nuanced affective component of negative stereotypes, even if they are not thought of as doing this by most researchers. We think this is indeed the case, insofar as the stereotypes being measured are, on our view, ineluctably affective. However, researchers who study "shooter bias" and "weapon bias" tend to endorse the two-type view, and explicitly argue that these measures primarily tap (putative) ISs rather than IEs (Correll et al. 2007; Correll et al. 2015; Glaser 1999; Glaser and Knowles 2008; Stewart and Payne 2008). As we discuss in §5, we think the predictive power of measures like the IAT could be improved by designing tests that home in on the specific affective content of specific ISs. For example, to test our suggestion that fear is more relevant than generic dislike to weapons associations , one might compare the predictive power of a standard race-weapons IAT to a race-weapons IAT using strong fear-inducing stimuli (e.g., threatening images of assault rifles). For research in this vein, see Gschwendner et al. (2008).

is precisely what Oswald and colleagues' (2013) meta-analysis of the IAT found; it is a consistent but weak predictor of behavior.[29] By contrast, the principal value of the Stereo-IAT may be its high predictive success within a narrow range of contexts. Recall Amodio and Hamilton's (2012) finding that the Stereo-IAT primarily reflected the difficulty participants had in associating blacks with intelligence. This suggests that, rather than tracking coldly cognitive stereotypes alone, the Stereo-IAT tracks the insidious and plainly negative stereotype that black people are unintelligent. This negative *evaluative stereotype* may be primarily responsible for the effect, rather than a "compensatory" stereotype that black people are athletically gifted. While we agree with defenders of the IS/IE distinction that neither blanket negativity toward outgroups nor some affectless set of beliefs are solely responsible for all forms of unconscious discrimination, in many cases we think the effects are driven by a conjunction of the two: *negative stereotypes about disadvantaged groups*.[30]

Rather than assessing generic likes and dislikes, or affectless semantic associations, future research should identify those pernicious stereotypes that "stick" precisely because of their affective and motivational significance. Laurie Rudman and colleagues' research on implicit evaluative stereotypes is exemplary in this respect. Rudman has experimented with a wide range of (what we dub) Evaluative-Stereotype IATs (ES-IATs) with a number of fascinating results. For example, Rudman and Kilianski (2000) found that gender-authority ISs (associating men's names with high-status occupational roles (boss, expert, authority) and women's names with low-status roles (assistant, subordinate, helper)) predicted implicit and explicit prejudice toward women authority figures. However, prejudice toward women authority figures was *not* predicted by gender-career ISs (associations with career, job, domestic, and family) or gender-agency ISs (associations with self-sufficient, competitive, communal, and supportive). In other words, they found that a highly specific evaluative stereotype predicted dislike of women leaders:

> prejudice against female authority may be due more to associations linking men to power and influence than to role or trait expectancies. In other words, women may be viewed as legitimate careerists, possessed of the agency necessary for flying 747s and performing surgery. However, if they violate expectancies that men (not women) occupy powerful roles, their authority in the cockpit or the operating room may not be welcomed… (2000, 1326)

[29] We do not, however, endorse the normative gloss Oswald and colleagues cast on this finding, that the IAT is a "poor" measure for predicting behavior. For a convincing reply, see Greenwald et al. (2015), who show how small effects can aggregate to have significant social impacts. We also note that the low predictive validity of the IAT is a central factor in Machery (2016), Huebner (2016), and Lee's (ms) arguments that implicit attitudes are dispositional "traits." If the predictive validity of the IAT improved significantly, perhaps because researchers took the suggestions we make below, that might spell trouble for the trait view. See §6.1 for brief discussion of the trait view.

[30] And positive stereotypes about advantaged groups (although we are less confident than Brewer (1999) and Greenwald and Pettigrew (2014) that ingroup favoritism is "more" of a culprit than outgroup derogation). See Gilbert et al. (2015) regarding the distinctive effects of men's and women's implicit *ingroup* versus *outgroup* stereotypes regarding math and English ability.

Using a racial ES-IAT,[31] Rudman and Lee (2002) found that listening to violent and misogynist hip hop increased ISs and led participants to interpret a black (but not a white) man's ambiguous behavior as hostile and sexist. The manipulation also affected stereotypical judgments about the man's intelligence, but not stereotype-irrelevant judgments, for example, about the man's popularity. Stereotype application was better predicted by the ES-IAT than by explicit measures. Rudman and Ashmore (2007) found that ES-IATs predicted discriminatory behavior against blacks, Jews, and Asians significantly better than a generic Eval-IAT. The ES-IAT predicted economic discrimination (how much money participants would distribute to student groups at their university) as well as autobiographical reports of slur use, social avoidance, property violations, and physical assault. Because the ES-IAT "combines beliefs with evaluation," write Rudman and Ashmore, it "may be a superior measure of implicit bias" (2007, 363).[32]

Some of the most celebrated studies demonstrating the predictive power of IATs have used ES-IATs. Rooth and colleagues found that implicit *work-performance* stereotypes predicted real-world hiring discrimination against both Arab-Muslims (Rooth, 2010) and obese individuals (Agerström and Rooth, 2011) in Sweden. Employers who associated these social groups with laziness and incompetence were less likely to contact job applicants from these groups for an interview. These landmark studies directly tied evidence of persistent hiring discrimination to implicit bias research, and specifically to implicit stereotypes related to competence (a core dimension of SCM). In both cases, the ES-IAT significantly predicted hiring discrimination over and above explicit measures of attitudes and stereotypes, which were uncorrelated or very weakly correlated with the ES-IAT. The predictive power of the obesity ES-IAT was particularly striking, because explicit measures of anti-obesity bias did not predict hiring discrimination at all—even though a full 58% of participants openly admitted a preference for hiring normal-weight over obese individuals. Given Rooth and colleagues' success in

[31] Rather than using allegedly neutral terms like "athletic," Rudman and Lee's racial ES-IAT used 7 overtly negative terms for black stereotypes ("hostile, violent, sexist, criminal, dangerous, crude, loud") and 7 positive terms for white stereotypes ("calm, lawful, ethical, trustworthy, polite, respectful").

[32] A reviewer for *Noûs* suggests that, if our view is right, "and all the implicit bias states being measured by the IATs are necessarily affect- and semantic-laden, then it seems all IATs are de facto ES-IATs," in which case it is not clear why ES-IATs should be superior to Eval-IATs. We do believe that the stimuli in Eval-IATs activate an array of ISs. What undermines their predictive power, relative to the ES-IAT, is the lack of conceptual coherence in the ISs being activated. The positive stimuli in Amodio and Devine's (2006, 654) Eval-IAT are as follows: "*honor*, *lucky*, *diamond*, *loyal*, *freedom*, *rainbow*, *love*, *honest*, *peace*, and *heaven*." The predictive power of this measure is lessened, we propose, because it activates too many different ISs, some of which may be entirely irrelevant to the experimental context. It adds noise to the data. ES-IATs are better by virtue of homing in on a specific cluster of situation-relevant, affect-arousing, strongly valenced ISs.

Also, we note that Rudman and colleagues have not (in print) explicitly defended a one-type view. Indeed, in one recent paper, Rudman and colleagues (2012, 18) seem to endorse the two-type model: "When stereotype IATs are valenced (e.g., when warmth is contrasted with coldness), they assess evaluative rather than semantic associations." We think the earlier conceptualizations are more accurate—ES-IATS measure jointly evaluative *and* semantic associations. In fact, Rudman finds our overall arguments against the two-type view convincing (p.c.). We are therefore hopeful that our arguments might reinvigorate research on implicit evaluative stereotypes.

predicting real-world discrimination, one might expect that ES-IATs would have become more widely used. However, subsequent field studies of a similar nature, such as Derous and colleagues (2014), instead used generic Eval-IATs—and (unsurprisingly, by our lights) failed to find that the latter significantly predicted hiring discrimination. Null findings like this are grist for the mill of IAT critics like Oswald and colleagues (2015, 567), who cite Derous and colleagues' findings in order to cast doubt on the legitimacy and replicability of Rooth's. On our view, the upshot is not that the predictive power of IATs is irredeemably hit-or-miss (or that Rooth's studies fail to replicate), but that researchers too often use the *wrong* measures for a given task.

IAT research should not only target more specific biases, but also explore how specific biases interact in specific contexts. SCM and TBM offer a variety of promising avenues to explore. SCM, for example, models perceptions of warmth and competence as deeply intertwined. Warmth and competence are inversely related in some contexts (e.g., a compensation effect toward outgroups), but positively related in others (a halo effect or favoritism effect for ingroups). Using separate IATs to measure warmth and competence, Carlsson and Björklund (2010) found evidence for implicit compensation effects toward outgroups, but not ingroups. Psychology students implicitly stereotyped lawyers as competent and cold, and preschool teachers as incompetent and warm. Preschool teachers, by contrast, implicitly stereotyped their own group as both warm and competent. Greenwald and colleagues (2002) explain how a one-type model can account for these sorts of effects (roughly, an individual who implicitly associates "self" with "preschool teacher," "self" with positive valence, and "competence" with positive valence, will also associate both "self" and "preschool teacher" with "competence").

Earlier we raised concerns about the claim that cognitive and non-cognitive states predict different types of behavior, in contrast to the claim that they make different contributions to one-and-the-same behavior (§3.2). If in fact the cognitive and the non-cognitive are dissociable, then stereotypes and evaluations should only predict behaviors in conjunction with other beliefs, feelings, and motivations. This problem of predictive underdetermination is, however, decidedly less acute when evaluative stereotypes are measured, precisely by virtue of measuring a cognitive/non-cognitive bundle. Stereotyping an outgroup as lazy and incompetent is apt to predict hiring discrimination better than stereotyping an outgroup as less "mental" than "physical." Implicitly associating blacks with positive physical traits of athleticism and rhythmicity likely predicts one set of interracial dispositions (who is picked first for the basketball team?), while implicitly associating blacks with negative physical traits of violence and danger predicts a different set of interracial dispositions (who is picked first in a suspect lineup?).[33] In short, when it comes to predicting behavior, evaluative stereotypes are where the action is.

[33] Although Holroyd and Sweetman (2016) share some of our concerns with the IE/IS distinction, they do not consistently steer clear of making it. For example, they suggest, citing Amodio and Devine's (2006) original studies, that it would likely be misguided to retrain mental/physical ISs if one's aim was to improve the quality of "interracial interactions" (e.g., to induce more pro-social, approach-oriented behaviors; see their §4.1). In such

6. Toward a One-Type Model

The Humean two-type model can be characterized as "interactionist." That is, ISs and IEs are distinct but usually interact "in the wild." For example, Glaser (1999, 4-5) writes:

> prejudice and stereotypes are, in fact, distinct constructs. They do interact, and are therefore highly correlated, but can develop and operate separately and even in contradiction. The dissociation of stereotyping and prejudice is important because if these constructs are conflated, which they often are, efforts to examine and redress them will be misguided and potentially wasted.

We agree that future research should explore the complex interactions among implicit biases, the various processes that influence their formation and change, and the ways in which interventions influence specific biases in specific ways. A one-type model, however, is more compelling. In §6.1, we briefly situate one-type theories with respect to other recent philosophical and psychological accounts of implicit attitudes. Then, in §6.2, we describe eight features of implicit mental states, understood as a singular kind of semantic-affective-behavioral cluster.

6.1 Associative, Propositional, and Dispositional Theories of Implicit Attitudes

One-type models tend to interpret implicit biases in terms of associations in a single semantic network. Some model IEs as the net valence of an activated set of semantic associations (e.g., Gawronski & Bodenhausen, 2006, 2011), while others model IEs as a specific (theoretically nondescript) subtype of semantic association, for example, between "black" and "unpleasant" (Greenwald et al., 2002). Both models offer plausible accounts of the co-activating properties of semantic associations. But both leave it somewhat of a mystery (a) whether merely cold, semantic ISs exist (which would point toward a two-type model) and (b) *why* the activation of semantic associations has any bearing on how individuals feel and act. More needs to be said about why the activation of the concept "unpleasant," for example, *feels* unpleasant, and initiates avoidance behavior.

Alternative theories that model implicit attitudes as propositionally structured states with language-like compositional structure, may also be consistent with a one-type view, although it is not always clear (e.g., Mitchell et al., 2009; de Houwer, 2014; Mandelbaum, 2013, 2015; Levy, 2014a,b, 2015, forthcoming). For example, Mandelbaum (2015) argues that implicit attitudes are "Structured Beliefs." Because Mandelbaum (2013, 200-201) also argues that occurrent beliefs are frequently, and perhaps always, associated with affect and motor routines, and because he suggests that all quotidian mental states stand in representational, affective, and

---

passages, Holroyd and Sweetman—like Amodio and colleagues—underestimate the potential affective-motivational significance of mental/physical ISs. The operative notion of "interracial interactions" is too coarse-grained. Retraining physicality ISs could very well affect interracial interactions in such contexts as who is approached and actively recruited to play on the basketball team, etc. We nevertheless agree with Holroyd and Sweetman's general point that interventions should be tailored to target precisely those social dispositions we deem most unjust. See §7.

behavioral associative relations, we suspect that the Structured Belief view of implicit attitudes is a one-type view. It conceptualizes implicit mental states as beliefs associatively clustered with various affective and behavioral responses. However, elsewhere Mandelbaum (2015) acknowledges an in-principle, conceptual distinction between "purely cognitive" associations between concepts and "hybrid" associations that yoke concepts with valence. He ultimately seems agnostic about whether this conceptual distinction, which would be suggestive of a two-type view, is borne out in psychological reality. Similarly, Levy's (2014a,b; 2015; forthcoming) account of implicit attitudes as "patchy endorsements"—propositionally structured states that dispose agents to respond to cues in particular ways—seems in-principle consistent with a one-type view (see esp. 2014a, 30; forthcoming, section on "Reasons Responsiveness and Implicit Attitudes"). For we take it that patchy endorsements dispose agents to particular feelings and behaviors. However, in response to Brownstein and Madva (2012a), Levy (2014b) claims that only a subset of implicit attitudes are associated with affective-behavioral dispositions. Some implicit attitudes, he writes, are "merely representational" states that "are acquired and stored as mere associations between contents" (2014b, 99n5). Levy's claims here are difficult to evaluate, however, because he does not provide evidence for or references regarding non-affective merely representational implicit attitudes.

Our purpose is not to grind axes in the debate between associative and propositional interpretations of implicit states. Rather, just as the associative accounts described above leave unexplained why the activation of semantic associations bears on how individuals feel and act, so too it seems that propositional accounts presuppose, rather than rival, existing accounts of the causal relations between concepts (or beliefs), valence, and affect in implicit states. Propositional accounts face the challenge of explaining how, as Walther and colleagues (2011, 193) put it, "propositional knowledge is translated into liking." For example, these views seem to simply stipulate, without explaining, that the belief *that Ss are unpleasant* causes aversive feelings and behaviors toward Ss (and vice versa). Mandelbaum and Levy clearly argue that at least some propositionally structured implicit states stand in associative relations with feelings and actions, and they might very well accept our claims in §6.2 about the properties and structure of these associative relations.

A final set of views understands implicit attitudes not as occurrent mental states at all, but rather as dispositions, akin to traits like courage, honesty, and so on (Schwitzgebel, 2013; Machery, 2016; Huebner, 2016; Lee, ms). Some dispositional views seem to be one-type. Machery characterizes a trait as a "disposition to perceive, attend, cognize, and behave in a particular way in a range of social and non-social situations" (2016, 111). One does not possess the trait of courage if one merely has courageous thoughts; cognitive, affective, and behavioral components are necessary. (Traits are "multi-track.") Schwitzgebel (2010) also argues that the "dispositional stereotype" for beliefs includes affective and behavioral dispositions. However, dispositional theorists might argue that cognitive kinds like implicit belief and conative kinds like implicit emotion and motivation have fundamentally different dispositional profiles, perhaps construing beliefs as dispositions to act as if *P* were true and desires as dispositions to make it

the case that *P* is true.[34] As we mentioned earlier, much of the appeal of the trait view rests on how well measures of implicit attitudes can predict agents' feelings and behavior. The following sketch of a one-type model is meant as a prompt for future research that might eventually improve these predictions.

6.2 Constraints for a One-Type Theory

We think implicit states are best conceived in terms of mutually co-activating semantic-affective-behavioral "clusters" or "bundles.". This is clearly similar to Gendler's (2008a,b) notion of "alief," a *sui generis* mental state with representational, affective, and behavioral (or *R*-*A*-*B*) components. Aliefs are a co-activating "cluster of dispositions to entertain simultaneously *R*-ish thoughts, experience *A*, and engage in *B*" (2008a, 645). Gendler's account adapts the traditional tripartite model of explicit attitudes to implicit attitudes. The tripartite model posits three related but distinct components of one type of mental construct:

> Prejudice is typically conceptualized as an attitude that, like other attitudes, has a cognitive component (e.g., beliefs about a target group), an affective component (e.g., dislike), and a conative component (e.g., a behavioral predisposition to behave negatively toward the target group). (Dovidio et al., 2010).

Critics of the tripartite model and of alief ask why we should posit unified *R*-*A*-*B* clusters, rather than merely co-occurring beliefs, feelings, and behaviors. What is the value, as Nagel (2012) asks, of explaining judgment and action in terms of "alief-shaped lumps?" A person who feels a craving to smoke may also experience an impulse to reach in her pocket for cigarettes, and unreflectively reaching in her pocket might activate a craving to smoke. But the co-occurring craving and behavioral impulse should still be understood as psychologically distinct types, the objection goes, because the connection can be weakened or severed, for example, by practicing another behavioral response when the craving is experienced (e.g., "when I feel a craving for cigarettes, then I will chew gum").

Ultimately, we think that the following list of features of implicit mental states ought to constrain theorizing about the nature of these states. Moreover, we think that this list indicates the distinctive advantages of understanding implicit mental states as semantic-affective-behavioral clusters. That is, it is because the following features tend to be shared among a set of states that we think these states make up a genuine mental kind.[35]

---

[34] See Schroeder (2015) for discussion.

[35] Although we think the following features are most at home in a one-type theory, we do not offer any knockdown argument against the possibility of their being incorporated into, or even better explained by, a two-type theory. We say more about why to think of implicit attitudes as a distinctive mental kind in Brownstein and Madva (2012a,b). On the value of setting out a list of features to characterize a mental kind, see Fodor's (1983) discussion of the nine features which he claims characterize modularity. There, Fodor articulates a cluster of traits that define a type of mental *system*, rather than a type of mental state. A system, for Fodor, counts as modular if it shares these features

6.2.1 Co-Activation A fundamental feature of implicit mental states is co-activation. Thoughts of birthdays activate thoughts of cake and thoughts of cake activate thoughts of birthdays. *Which* particular associations are activated in any given context varies in ways we begin to explain below, but *that* a "spread" of activation occurs does not vary. In many cases, mental items (like the concepts of "birthday" and "cake") may stand in a simple co-activating relation to each other, but there may be other, more cognitively complex relations among implicit associations as well (see Smith, 1996). For example, there may also be an inhibition relation (e.g., thoughts of safety may inhibit thoughts of fear).

6.2.2 Cross Modal Co-Activation Co-activation is not encapsulated within any particular mental system or type of mental state. It is not the case that semantic associations only co-activate other semantic associations. Thoughts of birthdays activate thoughts of cake, as well as feelings of hunger or joy (or sadness, as the case may be), and behavioral impulses to eat or sing or blow out candles. A response to a stimulus does not just induce thoughts, or induce feelings, or induce behaviors. It always induces all three, in part because they are co-activating. To the extent that a smoker can weaken the co-activating association between nicotine cravings and smoking-initiation behaviors, this change will be wrought not by *eliminating* all behavioral responses to nicotine cravings, but by *replacing* one behavioral response with another. There is always some behavioral inclination or other, but its specific content and intensity can change (see §6.2.7).

6.2.3 Degree, Not Kind These co-activating clusters vary in degree, rather than kind, of semantic, affective, and behavioral content. For example, the co-activating semantic association of black men and weapons *also* activates feelings of fear, but the intensity of the fear response will vary with context, the intensity of the stimulus, etc. (e.g., Correll et al., 2007). Hence we predict that damping down the fear response should also reduce the semantic black-weapon association. A better appreciation of the mediators, moderators, and downstream effects of such differences in degree should, we submit, be just as central to theoretical models of implicit attitudes, and to practical strategies for combating discrimination, as are differences in kind.[36]

6.2.4 Omnipresent Affect Intensity In lieu of using the "presence" or "absence" of affect to distinguish between mental kinds, we propose that implicit mental states vary in affective intensity along a continuum.[37] Affect continues to exert causal effects even at very low

---

"to some interesting extent." Although we focus on what constitutes implicit mental states as a class, an important further question is: which features characterize the mental *systems* that operate over these states?

[36] A striking example of how to investigate perceptual-semantic contents as continua rather than discrete categories (as well as how to investigate the cognitive intersections between race and gender) is Johnson et al.'s (2012) "technique of varying apparent race [of faces] on a continuum from Black to Caucasian to Asian, as opposed to the use of distinct Black/White categories in isolation," or the use of a continuum between only two racial categories.

[37] See also Fazio's (2007) account of the "attitude/non-attitude continuum," and Amodio and Lieberman's (2009) proposal (in response to emerging evidence that amygdala activity is "associated with arousal or the emotional

intensities, although low-intensity effects will typically escape an individual's focal attention (and may require more fine-grained measures than the IAT).

6.2.5 Moderators of Affect Intensity We predict that the affective intensity of implicit attitudes is a function of several a) content-specific, b) person-specific, and c) context-specific variables. Regarding a), implicit attitudes with certain contents (e.g., automatic evaluative responses to perceptions of spiders or vomit) will tend to be more affectively intense than others (e.g., perceptions of lamps or doorknobs). Regarding b), some individuals will have stronger affective dispositions toward certain stimuli than others. For example, participants who were obese as children or had beloved, obese mothers exhibit less negative automatic evaluations of obesity (Rudman et al., 2007); individuals with a strong dispositional "need to evaluate" exhibit more intense automatic evaluations in general (Jarvis & Petty, 1996; Hermans et al., 2001); and low-prejudice people tend to be less susceptible than high-prejudice people to negative affective conditioning in general (Livingston & Drwecki, 2007). Regarding c), certain contexts, such as being in a dark room, will amplify the strength of automatic evaluations of threatening stimuli (Schaller et al., 2003; Cesario & Jonas, 2014; see Brownstein, 2016 and Madva, 2016 for discussion).

6.2.6 Multipolar Valence In lieu of modeling affect intensity as a bipolar continuum between positive and negative valence, we differentiate among distinctive types of affective-motivational responses. For example, anxiety and anger are both "negative" affective responses, but research suggests that while anxiety often activates an *avoidance* orientation, anger typically activates an *approach* orientation (Carver & Harmon-Jones, 2009). Thus, affective responses can differ along *at least* two dimensions—positive vs. negative, and approach vs. avoid. We further predict differences among types of approach motivation. Anger, joy, hunger, and sexual arousal may each activate distinctive sorts of approach orientation.

6.2.7 Omnipresent Behavioral Impulse Intensity "Behavior" is, in a certain sense, part of the content of implicit mental states (see Gendler, 2008a, 635n4; Brownstein and Madva, 2012a,b). This point is not intended as a defense of behaviorism. The proposal is that a psychological *motor routine* is activated, which is functionally connected to the agent's activated thoughts and feelings. The activation of this motor routine may not issue in a full-fledged action, although there will typically be some behavioral expression, such as changes in heart rate, eyeblink response, or muscle tone. The mediators and moderators of affect discussed in §§6.2.5-6 also likely influence behavioral impulses, such that, for example, certain impulses in certain contexts are more controllable than others.

---

intensity of a stimulus, but not valence or fear per se") that "implicit prejudice may be better conceived as reflecting the intensity of one's reaction to an outgroup (vs. ingroup) face."

6.2.7 Neural Distribution Our one-type model opposes quasi-phrenological efforts to isolate the specific brain regions that subserve specific types of mental state. Of course, we do not deny that there are meaningful differentiations between brain networks, nor that these networks can in some respects be thought of as self-standing systems. Suppose that, during the shooter bias task, perceiving a black man activates semantic associations with criminality and guns, affective responses of danger, perceptual and attentional biases to detect signs of threat, and motor preparations for fight-or-flight (Correll et al., 2015). This co-activating effect can occur "within" as well as "between" different networks. In other words, at different levels of explanation, the brain can rightly be described as composed of several subsystems, as a unified system unto itself, and as one component of a larger bodily-environmental system. A single semantic network model at a higher-psychological level is consistent with a multi-system model at a lower-neural level.

7. Toward Better Interventions

We raised several questions for the two-type model and considered it in light of influential accounts of explicit stereotypes and prejudices. We made proposals about the structure of co-activation in implicit cognition, and the predictive validity of indirect measures of bias. We conclude by considering the relevance of theoretical conceptions of implicit bias for combating discrimination. We agree with Holroyd and Sweetman (2016) that research on prejudice reduction should appreciate the heterogeneity of implicit biases, and should take into account the distinctive "contents" of different biases. In what follows we make recommendations about concrete directions future research should take.

The two-type model appeals to the IE/IS distinction to motivate strategies for practical intervention. Gilbert and colleagues (2012, 3609) argue that "evaluative and stereotypic information may be learned, stored, and unlearned via different networks of information" and that "a consideration of these distinctions is critical when designing interventions to change social attitudes or stereotypes." The practical upshot of IE/IS independence, then, is said to be that we should recondition them separately. But perhaps we should infer just the opposite.

Suppose, contrary to what we have argued, that IEs and ISs are indeed dissociable. Even so, dissociation might represent a cautionary tale about what *not* to do in devising bias interventions. An intervention that seems to reduce IEs might leave ISs intact, or vice versa, in which case individuals might continue to act in discriminatory ways in many contexts. For example, Glaser (1999) found that stereotype-retraining reduced implicit prejudice but *not* implicit stereotyping. This finding ostensibly supports distinguishing between these two constructs, but it has exactly the opposite practical implication from the one proponents of the two-type model draw. Moreover, if stereotypes and prejudices are to any extent mutually supporting, then removing one but leaving the other intact might render the effects of the intervention especially short-lived. If an intervention reduces negative evaluations of blacks, but individuals continue to implicitly stereotype blacks as violent and unintelligent, then it may only be a matter of time before those stereotypes lead to the renewal of negative evaluations.

Likewise, if an intervention leads individuals to stop stereotyping, but individuals continue to have negative gut reactions toward blacks, then they will likely relearn the stereotypical beliefs that rationalize those gut reactions. Indeed, Crandall and colleagues' (2011, 1496) research on the "affective basis of stereotypes" found that inducing content-free "mere" prejudice toward novel social groups generated stereotypical responses toward those groups.

Aiming for a comprehensive debiasing intervention no doubt motivates the assertion that the separate interventions should be paired together as "complementary" (Gilbert et al., 2012). That is, they believe that interventions should combine "both types of reduction technique" (Amodio & Lieberman, 2009, 28). As a sheer matter of time and resources, however, it seems preferable to design fewer interventions that simultaneously combat as many biases or kinds of bias as possible. Two separate interventions are presumably more time-consuming and resource-intensive than one. Thus, even on the terms of the two-type model, it is not obvious that a two-type model of implicit biases warrants two-pronged intervention strategies.

Although getting the most debiasing bang for our interventional buck is no trivial matter, our primary concern is that two separate interventions will be less effective than one. Interventions may be least likely to work in stable and context-general ways when they target evaluation and stereotyping separately. In general, it is much harder to form enduring associations between meaningless semantic items (e.g., memorizing how to translate words between two foreign languages without knowing how any of those words translate into one's native tongue) than between meaningful items with affective-motivational significance (e.g., remembering to avoid foods to which one has a violent allergic reaction, not to mention remembering the name, smell, and sight of those noxious stimuli). In general, learning is facilitated by combining information with affective and motivational allure (Adcock et al., 2006; Cohen et al., 2014). Just watch a TED talk to see this.

While we find little evidence that combating (putative) ISs and IEs separately is effective (though time will tell), some extant data does suggest that retraining ISs and IEs together is effective. For example, Gawronski and colleagues (2008) found that training participants to associate, specifically, negative-black stereotypes with whites, and positive-white stereotypes with blacks, led to reductions in negative IEs. Similarly, Forbes and Schmader (2010) did not simply retrain non-evaluative semantic associations between women and math terms, but between the phrase "women are good at" and math. Rather than pinpointing evaluatively neutral semantic associations or semantically meaningless evaluative associations, these studies suggest that we should retrain heavily affect-laden stereotypes.

Interventions must also consider the concrete meanings that evaluative stereotypes take on for specific individuals in specific contexts—and how these contexts give rise to and maintain these biases. Retraining math biases in Forbes and Schmader (2010) had no effect on men's test performance, and it affected women's performance only in the context of stereotype threat (during a purported test of natural ability; see also Gilbert et al., 2015). Moreover, research on benevolent sexism shows that ostensibly positive attitudes can, in certain contexts, be causally related to insidious stereotypes. For example, saluting women or minorities as "hard-working"

can be a way of implicitly questioning their intelligence.[38] The limitations of enhanced intergroup liking are also evident in Bergsieker and colleagues' (2010) finding that, during interracial interactions, whites seek to be perceived as warm and likeable, while blacks and Latinos seek to be seen as competent and worthy of respect. This result is especially striking in light of Rudman and Ebert's findings that men implicitly like women, but do not implicitly associate them with leadership or respect. If certain types of positive affect are integrally related to pernicious stereotypes, then merely increasing warm feelings toward disadvantaged groups may be ineffective or even counterproductive for combating discrimination. If blanket negativity is not the problem, then blanket positivity is not the solution. We ought to target precisely those affect-laden stereotypes that perpetuate discrimination and inequality, whichever they may be.

Finally, if stereotypes are intrinsically affective, and if evaluations are intrinsically cognitive, then the rhetorical emphasis often put on the cold, cognitive core of implicit bias seems misleading. Theorists overgeneralize from the true claim that "negative" outgroup attitudes are not solely responsible for discrimination to the sweeping pronouncement that affective-motivational processes play no fundamental role at all (or at least play a secondary role to cognitive processes). More modestly, we should say that putatively innocuous, "positive," and "normal" intergroup feelings and desires can contribute to discrimination. Rather than proving that stereotyping and prejudice are fundamentally independent, that may just go to show how deeply and complexly they are intertwined.

Works Cited

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron*, *50*(3), 507-517.

Agerström, J., Carlsson, R., & Rooth, D. O. (2007). *Ethnicity and obesity: Evidence of implicit work performance stereo-types in Sweden* (No. 2007: 20). Working Paper, IFAU-Institute for Labour Market Policy Evaluation.

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790.

Allport, G. (1954). *The Nature of Prejudice*, Reading: Addison-Wesley.Amodio, D. M. (2009a). Coordinated roles of motivation and perception in the regulation of intergroup responses: Frontal cortical asymmetry effects on the P2 event-related potential and behavior. *Journal of Cognitive Neuroscience*, *22*, 2609-2617.

Amodio, D. M. (2009b). Intergroup anxiety effects on the control of racial stereotypes: A psychoneuroendocrine analysis. *Journal of Experimental Social Psychology*, *45*(1), 60-67.

Amodio, D. M. (2010). Can neuroscience advance social psychological theory? Social neuroscience for the behavioral social psychologist. *Social Cognition*, *28*, 695-716.

[38] See Holroyd and Sweetman (2016, 99).

Amodio, D. M., Bartholow, B. D., & Ito, T. A. (2014). Tracking the dynamics of the social brain: ERP approaches for Social Cognitive & Affective Neuroscience. *Social Cognitive & Affective Neuroscience*.

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., & 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*(4), 652.

Amodio, D. M. & Devine, P. G. (2009). On the interpersonal functions of implicit stereotyping and evaluative race bias: Insights from social neuroscience. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new wave of implicit measures*(pp. 193-226). Hillsdale, NJ: Erlbaum.

Amodio, D. M., & Hamilton, H. K. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion*, *12*(6), 1273.

Amodio, D. M., & Lieberman, M. D. (2009). Pictures in our heads: Contributions of fMRI to the study of prejudice and stereotyping. *Handbook of Prejudice, Stereotyping, and Discrimination.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Amodio, D. M., & Mendoza, S. A. (2010). 19. Implicit intergroup bias: cognitive, affective, and motivational underpinnings. In *Handbook of implicit social cognition: Measurement, theory, and applications*, 353-374).

Amodio, D. M., and Ratner, K. (2011). "A memory systems model of implicit social cognition," *Current Directions in Psychological Science, 20(3),* 143-148.

Anderson, E. (2010). *The imperative of integration*. Princeton University Press.

Appiah, K. A. 1990. Racisms. In *Anatomy of racism,* ed. D. T. Goldberg. Minneapolis: University of Minnesota Press.

Bergsieker, H. B., Shelton, J. N., & Richeson, J. A. (2010). To be liked versus respected: Divergent goals in interracial interactions. *Journal of personality and social psychology*, *99*(2), 248.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues*, *55*(3), 429-444.

Brownstein, M. (2016). Context and the ethics of implicit bias. Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 2: Responsibility, Structural Injustice, and Ethics*. Oxford: Oxford University Press.

Brownstein, M., & Madva, A. (2012a). The normativity of automaticity. *Mind & Language*, *27*(4), 410-434.

Brownstein, M., & Madva, A. (2012). Ethical automaticity. *Philosophy of the Social Sciences*, *42*(1), 68-98.

Carlsson, R., & Björklund, F. (2010). Implicit stereotype content: Mixed stereotypes can be measured with the implicit association test. *Social psychology*, *41*(4), 213.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychological bulletin*, *135*(2), 183.

Cesario, J., & Jonas, K. J. (2014). Replicability and models of priming: What a resource computation framework can tell us about expectations of replicability. *Social Cognition*, 32, 124-136.

Cesario, J., Plaks, J. E., Hagiwara, N., Navarrete, C. D., & Higgins, E. T. (2010). The ecology of automaticity: how situational contingencies shape action semantics and social behavior. *Psychological Science*, 21(9) 1311–1317.

Cohen, M. S., Rissman, J., Suthana, N. A., Castel, A. D., & Knowlton, B. J. (2014). Value-based modulation of memory encoding involves strategic engagement of fronto-temporal semantic processing regions. *Cognitive, Affective, & Behavioral Neuroscience*, 1-15.

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social cognitive and affective neuroscience*, *7*(7), 764-770.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology*,*37*(6), 1102-1117.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of personality and social psychology*, *108*(2), 219.

Cottrell, C. A., Richards, D. A., & Nichols, A. L. (2010). Predicting policy attitudes from general prejudice versus specific intergroup emotions. *Journal of Experimental Social Psychology*, *46*(2), 247-254.

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: a sociofunctional threat-based approach to "prejudice." *Journal of Personality and Social Psychology*, *88*(5), 770.

Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of Personality and Social Psychology*, *93*(5), 764.

Dasgupta, N., DeSteno, D., Williams, L. A., & Hunsinger, M. (2009). Fanning the flames of prejudice: the influence of specific incidental emotions on implicit prejudice. *Emotion*, *9*(4), 585

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social Psychology and Personality Compass*.

Degner, J., & Wentura, D. (2011). Types of automatically activated prejudice: Assessing possessor-versus other-relevant valence in the evaluative priming task. *Social Cognition*, *29*(2), 182-209.

Derous, E., Ryan, A. M., & Serlie, A. W. (2014). Double jeopardy upon resume screening: when Achmed is less employable than Aisha. *Personnel Psychology*.

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. *Journal of personality and social psychology*, *56*(1), 5.

Dixon, J., Levine, M., Reicher, S., & 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*(06), 411-425.

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: theoretical and empirical overview. *The Sage handbook of prejudice, stereotyping and discrimination*, 1.

Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition and emotion*, *21*(6), 1184-1211.

Ebert, I.D. (2009). Don't Be Afraid! Competent Women Are Great. Implicit Gender Attitudes and Stereotypes of Today. Doctoral dissertation.

Ebert, I.D., Steffens, M.C., Kroth, A. (2014). Warm, but Maybe Not So Competent?—Contemporary Implicit Stereotypes of Women and Men in Germany. *Sex Roles* 70:359–375

Faucher, L., & Machery, E. (2009). Racism: Against Jorge Garcia's moral and psychological monism. *Philosophy of the Social Sciences*.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878.

Flannigan, N., Miles, L. K., Quadflieg, S., & Macrae, C. N. (2013). Seeing the Unexpected: Counterstereotypes are Implicitly Bad. *Social Cognition*, *31*(6), 712-720.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of Personality and Social Psychology*, *99*(5), 740.

Gawronski, B. & Bodenhausen, G. (2006). "Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change," *Psychological bulletin, 132(5)*, 692-731.

Gawronski, B. & Bodenhausen, G. (2011). "The associative-propositional evaluation model: Theory, evidence, and open questions," *Advances in Experimental Social Psychology*, *44*, 59-127.

Gawronski, B., & Cesario, J. (2013). Of Mice and Men What Animal Research Can Tell Us About Context Effects on Automatic Responses in Humans. *Personality and Social Psychology Review*, *17*(2), 187-215.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, *44*(2), 370-377.

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, *44*(5), 1355-1361.

Gendler, T. S. (2008a). Alief and belief. *Journal of Philosophy*, *105*(10), 634.

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, *23*(5), 552-585.
Gilbert, P. N., O'Brien, L. T., Garcia, D. M., & Marx, D. M. (2015). Not the Sum of Its Parts: Decomposing Implicit Academic Stereotypes to Understand Sense of Fit in Math and English. *Sex Roles*, *72*(1-2), 25-39.
Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia* 50, 3600-3611.
Glaser, J. C. (1999). *The relation between stereotyping and prejudice: Measures of newly formed automatic associations*. Doctoral dissertation, Harvard University.
Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*(1), 164-172.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.
Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2014). Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects.
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*(1), 3.
Greenwald, A. G. & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69*(7), 669-684.
Gschwendner, T., Hoffman, W., and Schmitt, M. (2008). Differential Stability: The Effects of Acute and Chronic Construct Accessibility on the Temporal Stability of the Implicit Association Test. *Journal of Individual Differences, 29*(2), 70-79.
Hermans, D., De Houwer, J., & Eelen, P. (2001). A time course analysis of the affective priming effect. *Cognition & Emotion*, *15*(2), 143-165.
Holroyd, J. & Sweetman, J. (2016). "The Heterogeneity of Implicit Biases." Brownstein, M. and Saul, J. (Eds.) *Implicit Bias and Philosophy: Volume 1: Metaphysics and Epistemology*. Oxford: Oxford University Press.
Huebner, B. (2016). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. Brownstein, M. and Saul, J. (Eds.) *Implicit Bias and Philosophy: Volume 1: Metaphysics and Epistemology*. Oxford: Oxford University Press.
Hugenberg, K., & Bodenhausen, G. V. (2003). Facing Prejudice Implicit Prejudice and the Perception of Facial Threat. *Psychological Science*, *14*(6), 640-643.
Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization the role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342-345.
Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*(1), 172.
Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: how covarying phenotypes and stereotypes bias sex categorization. *Journal of personality and social psychology*, *102*(1), 116.

Judd, C. M., Blair, I. V., & Chapleau, K. M. (2004). Automatic stereotypes vs. automatic prejudice: Sorting out the possibilities in the weapon paradigm. *Journal of Experimental Social Psychology*, *40*(1), 75-81.

Kawakami, K., Phills, C. E., 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*(6), 957-971.

Kawakami, K., Steele, J., Cifa, C., Phills, C., and Dovidio, J. (2008). "Approaching math increases math = me, math = pleasant," *Journal of Experimental Social Psychology*, *44,* 818-825.

Lebrecht, S., Bar, M., Barrett, L. F., & Tarr, M. J. (2012). Micro-valences: perceiving affective valence in everyday objects. *Frontiers in Psychology, 3*

Lee, C. (Manuscript.) A Dispositional Account of Aversive Racism.

Levy, N. (2014a). Consciousness, implicit attitudes and moral responsibility. *Noûs, 48*(1), 21-40.

Levy, N. (2014b). *Consciousness and moral responsibility*. Oxford University Press.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*, 49 (4), 800-823.

Levy, N. (forthcoming). Implicit Bias and Moral Responsibility: Probing the Data. *Philosophy and Phenomenological Research*.

Machery, E. (2016). De-Freuding Implicit Attitudes. Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 1: Metaphysics and Epistemology*. Oxford: Oxford University Press.

Madva, A. (2016). Virtue, social knowledge, and implicit bias. Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume 1: Metaphysics and Epistemology*. Oxford: Oxford University Press.

Mandelbaum, E. (2013). Against alief. *Philosophical studies*, *165*(1), 197-211.

Mandelbaum, E. (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*. DOI: 10.1111/nous.12089

McConahay, J. B., & Hough, J. C. (1976). Symbolic racism. *Journal of social issues*, *32*(2), 23-45.

Mitchell, J. P. (2009). Social psychology as a natural kind. *Trends in cognitive sciences, 13*(6), 246-251.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183-198.

Nagel, J. (2012). Gendler on alief. *Analysis*, *72*(4), 774-788.

Nosek, B. A. and Banaji, M. R. (2009). Implicit attitudes. *Oxford Companion to Consciousness*. Oxford: Oxford University Press, 84-85.

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion*, *22*(4), 553-594.

Ofan, R. H., Rubin, N., Amodio, D. M. (2011). Seeing race: N170 responses to race and their relation to automatic racial attitudes and controlled processing. *Journal of Cognitive Neuroscience, 23*, 3152-3161.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*. doi: 10.1037/a0032734

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance.

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*(5), 729-738.

Ratner, K. G., Dotsch, R., Wigboldus, D., van Knippenberg, A., & Amodio, D. M. (in press). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523-534.

Rosenberg, S., & Sedlak, A. (1972). Structural representations of implicit personality theory. *Advances in Experimental Social Psychology, 6*, 235-297.

Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of personality and social psychology, 63*(2), 198.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations, 10*, 359-372.

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

Rudman, L. A. & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin, 26,* 1315-1328.

Rudman, L. A. & Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes & Intergroup Relations, 5,* 133-150.

Rudman, L. A., & Mescher, K., & Moss-Racusin, C. A. (2012). Reactions to gender egalitarian men: Feminization due to stigma-by-association. Group Processes and Intergroup Relations.

Rudman, L. A., Phelan, J. E., & Heppen, J. B. (2007). Developmental sources of implicit attitudes. *Personality and Social Psychology Bulletin, 33*(12), 1700-1713.

Salzman, C. D., & Fusi, S. (2010). Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annual review of neuroscience, 33*, 173.

Schaller, M., Park, J.J., and Mueller, A. (2003). Fear of the dark: Interactive effects of beliefs about danger and ambient darkness on ethnic stereotypes. *Personality and Social Psychology Bulletin 29*, 637-649.

Schroeder, T. (Summer 2015 Edition). Desire. *The Stanford Encyclopedia of Philosophy* , Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/desire/>.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*(4), 531-553.

Schwitzgebel, E. (2013). A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. *New Essays on Belief: Constitution, Content and Structure*, 75.

Siegel, S. (2012). "Cognitive Penetrability and Perceptual Justification." *Nous, 46(2),* 201-222.

Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology*, 42(4), 300–313.

Stangor, C. (2009). The study of stereotyping, prejudice, and discrimination within social psychology. *Handbook of prejudice, stereotyping, and discrimination*, 1-22.

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*(10), 1332-1345.

Tapias, M. P., Glaser, J., Keltner, D., Vasquez, K., & Wickens, T. (2007). Emotion and prejudice: Specific emotions toward outgroups. *Group Processes & Intergroup Relations*, *10*(1), 27-39.

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Cowen, P. J., & Hewstone, M. (2012). Propranolol reduces implicit negative racial bias. *Psychopharmacology*, *222*(3), 419-424.

Valian, V. (1998). *Why so slow?: The advancement of women*. MIT press.

Valian, V. (2005). Beyond gender schemas: Improving the advancement of women in academia. *Hypatia*, *20*(3), 198-213.

Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science*, *20*(3), 192-196.

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45, 1191-1207.