Alex Madva and Michael Brownstein
Draft 2.6.15
Comments Welcome

# The Blurry Boundary between Stereotyping and Evaluation in Implicit Cognition

**Introduction**

Does the distinction between cognition and affect apply to implicit mental states? In the literature on intergroup relations, cognition and affect correspond to stereotypes— beliefs about social groups, or, more concretely, representations of traits associated with social groups (Stangor, 2009)—and prejudice—negative affective or evaluative responses toward others on the basis of social group members (McConahay & Hough, 1976; Dixon et al., 2012). While the distinction between explicit stereotypes and explicit prejudices is widely supported (Fiske & Linville, 1980; Martin & Halverson, 1981; Wol, 1982; Augoustinos & Innes, 1990; Duckitt, 1992; Dovidio et al., 1996; Fiske, 1998; Mackie & Smith, 1998; Fiske et al., 2002), its applicability to implicit mental states is controversial.

Understanding the relationship between stereotypes and prejudice is important as it can help illuminate the nature of the mental representations underlying implicit mental states. While some theories propose that stereotyping and prejudice are functionally integrated (Greenwald et al., 2002; Gawronski & Bodenhausen, 2006, 2011), others propose that the cognitive and affective components of implicit states represent separate constructs—namely implicit stereotypes (IS) and implicit evaluations (IE)—which reflect different mental processes and neural systems (Amodio & Devine, 2006, 2009; Amodio & Ratner, 2011). That stereotyping and prejudice are at least *conceptually* distinct in implicit cognition has been clear ever since Anthony Greenwald and Mahzarin Banaji (1995) coined these terms. But the question of their functional integration or independence leads to a number of significant conclusions. Most centrally, functional integration supports a "one-type" model of implicit states, according to which the core construct of the field is implicit attitudes as such (Fazio, 1990; Strack & Deutsche, 2004; Gawronski & Bodenhausen, 2006, 2011; Petty, 2006). Functional independence, on the other hand, supports a "two-type" model, according to which there are two fundamental kinds of implicit mental states: ISs and IEs (Amodio & Devine, 2006, 2009; Forbes & Schmader, 2010; Amodio & Ratner, 2011).

Considering the evidence for a one-type vs. two-type model bears upon a number of other important questions. First, understanding the relationship between stereotyping and prejudice in implicit mental states may help to improve the predictive validity of indirect measures of attitudes. The predictive validity of the most widely used indirect measure—the Implicit Association Test (IAT; Greenwald et al., 1998)—has been questioned,[1] and one potential limitation of the standard race-evaluation IAT is that it does not distinguish the relative contributions to test performance of liking/disliking social groups from beliefs about social groups. Second, given the centrality of the

[1] See, for instance, Oswald et al. (2013). We note, however, that even this very critical review found that the IAT is a small, but significant predictor of behavior across many domains. For reply, see Greenwald et al. (2014).

affect/cognition distinction in theories of the mind more generally, understanding the relationship between stereotyping and prejudice in implicit mental states can help to integrate theories of implicit social cognition with research in neuroscience (Phelps et al., 2000; Amodio & Ratner, 2011) and other branches of psychology. Finally, the development of effective interventions to combat implicit bias is likely to be aided by a clearer understanding of the relationship between feelings and beliefs about social groups (Amodio & Devine, 2006; Dasgupta, 2013).

The two-type model is widely influential and arguably comprises the most-well developed extant theory of the relationship between stereotypes and prejudice in implicit intergroup cognition. Despite holding significant appeal, however, we find that the overall evidence does not support the two-type view. We propose a conceptual framework for an alternate view, roughly according to which all ISs are irreducibly evaluative and affect-laden while all IEs are "semantic," in the sense that IEs stand in co-activating associative relations with concepts and beliefs. Implicit attitudes are best conceived, in other words, in terms of differences between particular "clusters" or "bundles" of semantic-affective associations, rather than between two broad types of mental state. These clusters differ in degree, rather than kind, of semantic and affective content. We discuss the relationship between our proposed framework and extant one-type theories, experiments that would be diagnostic for the different theories on offer, and future directions for improving the predictive validity of indirect measures and interventions to combat discrimination.

**The Two-Type Model of IS and IE**

In Amodio and Devine (2006), participants first took two distinctive IATs: the standard evaluative race IAT (Eval-IAT), which measures associations between black and white faces and generic pleasant 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. The authors interpret this dissociation in terms of the 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. Roughly, the Stereo-IAT measures "cold" implicit beliefs about racial groups while the Eval-IAT measures "hot" implicit likes, dislikes, and preferences.[2]

Further evidence for the 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 a distinctive range 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

[2] Earlier research measured more overtly evaluative stereotypes, like wealthy vs. poor or educated vs. ignorant (Judd et al. 2004, Rudman et al. 2001). The Stereo-IAT purports to avoid this confound.

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 IS and IE 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 lead Amodio and Ratner (2011, 143) to conclude that the IS/IE distinction paves the way for making "new and increasingly refined predictions."

This research spurred renewed interest in disentangling the roles of stereotyping and evaluation in other forms of discriminatory behavior. For example, researchers have explored the relative contributions of IS and IE 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 infer 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 and colleagues, 2004).

Data relating implicit social cognition to neuroanatomy has also been used to support the functional dissociation of IS and IE. The general claim is 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 (2008, 201) write, "and therefore it may be important to measure and conceive of affective and semantic associations independently." This claim is supported by studies using fMRI (Gilbert et al., 2012), EEG (Amodio 2009a; Amodio, Bartholow, and Ito 2014), measures of cortisol reactivity (Amodio, 2009b), and startle eyeblink responses (Amodio, Harmon-Jones, and Devine 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 motivation. 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).[3]

Relating the apparent dissociation of IS and IE to the development of interventions to combat discrimination, Forbes and Schmader (2010) studied the differential effects of *retraining* IEs

[3] We 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. A seating distance measure, for example, 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.

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 answering 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, 7). 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, 7).[4] Amodio and Lieberman (2009) stress similar conclusions in related research: "our findings suggest that different prejudice reduction techniques are needed to target these two types of implicit bias."

These studies lend support to related research programs in other areas of psychology and even outside psychology. For example, the socio-cognitive account of gender schemas—a matrix of related stereotypical associations—similarly focuses on the dissociation of coldly cognitive stereotypes from hot affective-motivational evaluations. Virginia Valian (2005, 198-200) writes:

> 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)

Outside of psychology, philosophers have stressed a similar distinction. For example, 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

[4] Forbes and Schmader also tested how IEs and ISs *interact* by pairing IE-retraining with IS-retraining. In one study, participants were trained either to like or dislike math, and either to reinforce or undermine math-gender stereotypes. Only women who had been trained *both* to like math and to reinforce math-gender stereotypes were susceptible to stereotype threat. Roughly, believing that they were not good at what they liked to do led them to try harder but not to perform better. By contrast, participants who underwent counterstereotype training were impervious to stereotype threat. They performed better without trying harder, and did so regardless whether they liked or disliked math. Such findings on the practical applications of implicit bias interventions are fascinating and extremely important (§7).

> 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.[5]

In these passages, Valian and Anderson apply the well-known distinction between explicit beliefs (stereotypes) and explicit feelings (prejudices or attitudes) to implicit bias. Doing so may make a bitter pill easier to swallow. Emphasizing that 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. The two-type model is also laudatory for illustrating 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 about the existence of implicit intergroup bias. The two-type model also points to the general importance of understanding the underlying nature of implicit bias in order to combat discrimination, and more specifically to the likelihood that successful interventions will have to target specific biases or kinds of biases.

However, in the next two sections we suggest that the empirical evidence is consistent with a one-type model of implicit mental states, or at least a far blurrier boundary between stereotyping and evaluation than the data above suggests.

**Reconsidering the Two-Type View**

In this section we raise four questions about the two-type model. First, we consider whether the prompts used in the Stereo-IAT are in fact evaluatively neutral. Second, we raise a conceptual question about the claim that ISs and IEs cause different types of behavior, rather than making different types of causal contribution to (one and the same type of) behavior. Third, we consider whether 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. Finally, we consider whether the neural data supports a two-type over a one-type model.

*Reconsidering the Stereo-IAT* The Stereo-IAT is designed to isolate the cold, cognitive core of ISs. The measure includes only positive and putatively "neutral" words regarding intelligence and athleticism. For most participants, white faces were more easily associated with positive words like "genius" and "smart" and 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 constitute "relatively neutral" ways of grouping the categories (655), rather than using more overtly evaluative groupings like "smart" and "athletic." To the extent that "mental" and "physical" *are* evaluative terms, they have a similar positive valence, unlike in the Eval-IAT, in which the two categories, "good" and "bad," clearly differ in evaluative

[5] For another philosophical account of the heterogeneity of implicit bias, see Holroyd & Sweetman (forthcoming).

standing. Thereby, the Stereo-IAT is claimed to be "only conceptual, but not evaluative" (Amodio and Hamilton, 2012, 1275).

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"). The terms "mental" and "physical" denote and connote traits people generally like to have, and often use to flatter others. 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.

Historical connotations notwithstanding, it is likely that all trait dimensions 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 may show 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.[6]

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. In general, a trait like intelligence or being "good at" some activity typically 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). Because intelligence and athleticism are typically desirable traits, we suspect that they enjoy a kind of default positive valence. Nevertheless, one might perceive either of them in a negative light in certain contexts. For example,

[6] A follow-up study could measure effects of training on reported feelings of confidence, "belonging", and stereotype endorsement. If participants reported 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 cognitive retraining procedure was affective after all.

an individual 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. By the same token, an individual who self-identifies as intellectual might come to disdain athleticism and "dumb jocks." We say more about these sorts of "compensation effects" later.

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 elements of visual processing are pervasively influenced by valence, then the same could very well be true of implicit stereotyping, but at a level too low to be captured by the Stereo-IAT.[7] Indeed, if the visual processing of non-social objects is valenced from the outset, the same is surely true of social perception, e.g., the visual processing of faces. Micro-valences in social perception should in turn influence the activation and operation of stereotypes, while stereotypes reciprocally influence social perception. Hugenberg and Bodenhausen (2003, 2004) find 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 cases, social perception seems to be intrinsically affective, and social affect intrinsically cognitive. Some of Amodio and colleagues' own research demonstrates how early perceptual processes are intertwined with higher-level cognition and affect (Amodio 2009a; Ofan, Rubin, and Amodio 2011; Ratner et al. in press). Amodio, Bartholow, and Ito (2014, 388) write that, "intergroup attitudes and goals can affect the way we see faces in the first place such that early perceptual biases may contribute to more elaborated forms of prejudice and stereotyping that have been traditionally found in social psychological research." A striking demonstration of the evaluative significance of stereotypes is Flannigan and colleagues' (2013) finding that men and women in counterstereotypical roles (men nurses, women pilots) are "implicitly bad"—the sheer fact that these stimuli are counterstereotypical leads individuals to implicitly dislike them.

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.[8] In research on self-reported emotions, Cottrell and colleagues (2010) found that specific intergroup emotions did, while general evaluative attitudes did not, predict policy views about gay rights and immigration. 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

[7] Whether this effect 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 2011).
[8] See also Dasgupta et al., 2009.

a threat). 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. 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 and 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. Which emotional response takes prominence, we propose, will vary with context.

*Conceptual Questions about Behavioral Prediction* The two-type model claims that stereotypes and evaluations predict different types of behavior. However, cognitive and non-cognitive states are not traditionally distinguished because they cause distinctive behaviors, but because they make distinctive causal contributions to behavior.[9] 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 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.[10] However, the contrast between, e.g., the *stereotype that one group is more athletic* with the *evaluation that one group is more likeable* does not control for the intentional content of what is being represented in this

[9] We believe there are limitations and problems in belief-desire psychology (Gendler 2008b; Brownstein and Madva 2012a,b), but this is the received view.

[10] For example, Rudy might believe it is the case that he is athletic, or he might *want* it to be the case that he is athletic. He might also hope, imagine, pretend, regret, or suppose it to be the case that he is athletic. Perhaps, if we hold fixed as many as possible of his other attitudes, Rudy will act in different ways depending on which attitude he takes toward the proposition that he is athletic. If he believes he is athletic, he might be less susceptible to stereotype threat and perform better in tests of "natural" athletic ability. If he merely wants to be athletic, he might spend more effort and time practicing, but be more susceptible to stereotype threat. Even here, we are speaking loosely. The effects of wanting to be athletic themselves depend on whether Rudy does or does not believe that he already is athletic, that practice is necessary to become and stay athletic, and so on. It may even be impossible to match the intentional content of cognitive and non-cognitive attitudes.

way. 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, Swencionis, and Amodio's (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 isolate the activation patterns of judgment about friendship, participants might have been asked, "Who is more likely to enjoy time with friends?" or "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?). 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 (thereby introducing concepts of enjoyment and pleasure into trait attribution) instead of a more straightforward stereotypic trait attributions, such as "Who is more athletic? Who is the better athlete? Who is stronger?"[11]

*Reconsidering the Double Dissociation* The double dissociations observed in studies that appear to support the two-type model may, therefore, not in fact reflect wholly separate cognitive and affective systems. These dissociations are perfectly consistent with the possibility 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*.

Evidence for dissociations between specific stereotypes is often overlooked. For example, Devine (1989) argues 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

[11] Amodio and colleagues suggest that athleticism is a relatively non-evaluative stereotype, but it might be 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?"

> cultural stereotypes… Individuals have unique, personal experiences of their cultural context and this is reflected in the fact that cultural perceptions vary across individuals…

Indeed, Amodio and Hamilton (2012) themselves found that 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 may reflect a particular (pernicious and negative) racial stereotype, which is perhaps dissociable from athleticism stereotypes, rather than a general disposition to associate racially typical faces with all culturally prevalent stereotypes.

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 might predict further behavioral dissociations, e.g., that some individuals 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 (see below). In other words, while the two-type model claims that IS and IE represent two broad classes of implicit bias, which are in turn made up of multiple "species" of specific ISs and IEs, their data is consistent with there being one class of implicit bias, some "species" of which are less affectively intense than others (§6).

*Neural data* It is unclear whether neuroscientific data clearly supports a two-type view of implicit mental states. Evidence for the traditional identification of the amygdala with affect and the prefrontal cortex with cognition is mixed (e.g., Salzman and 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. Duncan and Barrett focus primarily on non-social forms of cognition, rather than on implicit or social cognition. But if there is no such thing as non-affective non-social cognition, there is likely no such thing as non-affective social cognition either. We would expect that implicit social-cognitive processes are, if anything, even more pervasively shaped by affect (Mitchell 2009; Contreras and colleagues 2012).

Even proponents of the brain-basis of the cognitive/affective distinction have expressed similar concerns. David 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:

> 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 embedded in MSM and in the two-type model. Amygdala activation was suggested 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.[12,13]

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" (2010, 708).

We agree. But perhaps implicit attitudes are a high-level construct, on par with romantic love and the self, and so are unlikely to be localized in distinctive brain regions.[14] Perhaps relatively affective and semantic components of implicit attitudes are associated to greater or lesser degrees with specific regions or networks, but these could be viewed as components that subserve the (high-level) construct of interest (i.e., implicit attitudes). We return to this point below.

**Cognition and Emotion in Explicit Cognition**

The two-type model draws inspiration for the claim that IE and IS are independent constructs from research on explicit social cognition, yet 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 and Neuberg 2005) emphasize their

---

[12] Amodio himself has offered different responses to this issue. Amodio and Ratner (2011) and Amodio (2010) propose mapping IEs specifically to the amygdala's central nucleus, whereas Amodio and Lieberman (2009) propose that IEs may be more a matter of arousal and intensity than of negative valence.

[13] Amodio also rejects traditional views of how cognitive and affective processes *interact*: "early notions of emotion regulation hearken back to the Freudian and Cartesian ideas of inner conflict between *passion* and *reason*, and the belief that cognition (i.e., reason) must be invoked to directly down-regulate emotion (i.e., the passions). The idea that cognition directly down-regulates emotion is still pervasive today…" (2010, 710). By contrast, Amodio (2009a) argues that action control requires the harmonious coordination of perceptual, cognitive, affective, and motivational processes. Affective-motivational processes figure on this model not as rogue dispositions that need to be reined in by cold logic, but as "key mechanisms of self-regulation." We are sympathetic with this aspect of MSM, but we would invite Amodio and colleagues to go further in challenging received views—toward questioning why to rely on the cognitive/affective distinction at all in differentiating among these processes.

[14] Amodio notes that Gillihan and Farah (2005) raise similar concerns about neuroscientific research on the self.

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. 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 strongly 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: generic IEs correlated with ISs, about both warmth and intelligence. Similarly, Agerström, Carlsson, and Rooth (2007, 21) found that implicit dislike of Arabs on an Eval-IAT was strongly correlated with implicit associations of Arabs with *in*competence on a Stereo-IAT (r=.52). While SCM researchers have focused primarily on self-reports and other direct measures of attitudes, these studies exemplify how to investigate these phenomena with indirect measures.[15]

SCM researchers have been clear in acknowledging the irreducibly evaluative nature of ISs, but many have not appreciated the ways in which IEs are also "semantic."[16] In response, Cottrell and Neuberg argue 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 much better than did generic intergroup dislike and "negative feelings."

Like SCM, Cottrell and Neuberg's (2005) "threat-based" model (TBM) of intergroup emotions is based primarily on self-report. However, this "rich texturing of emotions" likely affects

[15] We are not concerned to defend the universal validity of SCM as a theory, but think that it has much to offer.
[16] For exceptions, see Tapias et al., 2007 and Dasgupta et al., 2009.

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.

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 too coarse-grained to capture, let alone differentiate among, the many affect-laden responses most relevant to social behavior. Below we make further suggestions for how the insights of SCM and TBM might be used to enhance the predictive validity of indirect measures.

**Toward a One-Type Model**

The two-type model can be characterized as "interactionist." That is, ISs and IEs are distinct but usually interact "in the wild." Hence, Amodio and Devine (2008) argue that future research should focus on "the interface of cognition and emotion." We agree that future research should explore the complex interactions among implicit biases, the various processes that influence the formation and change of implicit attitudes, and the ways in which manipulations and interventions influence specific biases in specific ways. However, we suggest that the case has not been made for a two-type model of implicit attitudes and that an "integrated" one-type account of ISs and IEs remains compelling. Implicit attitudes are indeed heterogeneous—some are more controllable and some less, some more conscious and some less, some more affectively intense and some less—but their variety is best conceived in terms of differences between particular "clusters" or "bundles" of semantic-affective associations, rather than between two broad types of association.

We lack the space to develop a full account of these clusters of semantic-affective associations, but a compelling framework is provided by the notion of "alief" in recent philosophical psychology (Gendler, 2008a,b). Alief represents a proposal for understanding what kind of mental state implicit attitudes are. The core idea is that implicit attitudes are in some respects like beliefs—they represent the world as being a certain way and play a role guiding behavior—but in other respects fail to display the properties of ordinary beliefs—they are insensitive to an agent's subjective perception of truth and evidence. What we find useful about alief in this context is the distinctive intentional content of these putative states. Aliefs are defined by their tightly bound representational, affective, and behavioral (or *R-A-B*) components. They involve "a cluster of dispositions to entertain simultaneously *R*-ish thoughts, experience *A*, and engage in *B*" (2008a, 645). In the case of an implicit race bias, the content of one's alief might be, "Black man! Scary! Avoid!" The *R-A-B* components of an occurrent alief are said to be automatically co-activating, such that the perception of a relevant cue simultaneously activates particular feelings and behaviors.

It is this automatic co-activation of thoughts, feelings, and behavior that we find promising as a conceptualization of the relation between prejudice and stereotypes in implicit cognition. Rather than interpret some implicit biases as purely cold, cognitive beliefs, and others as simple, generic dislikes of social groups, the alief model recognizes that implicit biases are generally constituted by a mélange of tightly intertwined cognitive, affective, and motivational factors. This approach incorporates the insights of the traditional tripartite model of explicit attitudes and prejudices into an account of implicit mental states. It 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).

On our view, alief-like "clusters" vary in degree, rather than kind, of semantic and affective content. For example, the co-activating semantic association of black men and weapons may *also* activate feelings of fear, but the intensity of the fear response will vary with context, the intensity of the stimulus, etc. Hence we predict that damping down the fear response should also reduce the semantic black-weapon association. The idea is a shift away from differences in kind and toward differences in degree. For example, despite championing the two-type model, Amodio and Lieberman (2009) discuss how changing views about the amygdala might call for reconceptualizing implicit prejudice. In light of evidence that amygdala activity is "associated with arousal or the emotional intensity of a stimulus, but not valence or fear per se," they propose "that implicit prejudice may be better conceived as reflecting the intensity of one's reaction to an outgroup (vs. ingroup) face." On this model, amygdala activation reflects the degree of intensity of one's response, rather than the activation of a distinctly affective-evaluative 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.[17]

In defending a one-type model of implicit attitudes, we of course 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, and motor preparations for fight-or-flight. This co-activating effect could be realized by a wide variety of neural mechanisms. It is consistent with there being dedicated neural regions for each "type" of response, or with the substrates and circuits for each response being distributed across different regions. Co-activation can occur "within" as well as "between" different regions and circuits. Put in other words, at different levels of explanation, the brain can rightly be described as

[17] See also Fazio's (2007) account of the "attitude/non-attitude continuum." Fazio also discusses Hermans, De Houwer, and Eelen's (2001) research on individual differences that can moderate the intensity of one's implicit attitudes, such as different individuals' "need to evaluate."

comprised 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.

Critics of the tripartite model of attitudes and of Gendler's account of alief ask why we should posit clusters of *R-A-B* content, 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?"[18] Ultimately, the principal advantages, as we see it, of the one-type interpretation are better, more ecologically valid, behavioral predictions, and better predictions of when implicit mental states do and do not change. These are vital questions for identifying and combating discrimination.

**Toward More Predictive Validity**

On our view, the principal virtue of the standard Eval-IAT is that it is a measure of generic likings and preferences, and as such it has the potential to predict a wide range of behaviors across many individuals and contexts. Its principal vice, however, is that the effect sizes are likely to be small. Like a jack of all trades and a master of none, the generic Eval-IAT should predict many behaviors, but at the cost of predicting few of them particularly well. This is precisely what Oswald and colleagues' (2013) review of the IAT found; it is a consistent but weak predictor of behavior.[19] By contrast, the principle value of the Stereo-IAT may be its high predictive success within a narrow range of contexts.[20] 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 "counter-balancing" stereotype that black people are athletically gifted. Thus 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

[18] Empirically, an enduring challenge for accounts that posit a single construct with multiple components is to explain why the components are not 100% correlated. (The enduring challenge for accounts that posit two or three types of wholly separate constructs is to explain why it is so difficult to generate conditions in which they are 0% correlated.) However, apparent evidence that these components are not always correlated (e.g., that the Stereo-IAT activates cognitive but not affective responses) might be better thought of as cases where some of the components are activated more intensely than others (e.g., perhaps affective responses to the Stereo-IAT are too subtle to be detected by that particular measure).

[19] We do not, however, endorse the normative gloss Oswald and colleagues cast on this finding (i.e., that the IAT is a poor measure for predicting behavior).

[20] Oswald and colleagues (2013) failed to find that Stereo-IATs predicted any behaviors better than the Eval-IAT, but meta-analyses of Stereo-IATs are premature. Whether, and how well, a given Stereo-IAT predicts any particular behavior depends to a great extent on the stimuli and behaviors at issue, in contrast to the Eval-IAT. For example, Rudman and Kilianski (2000) found that only one of three types of Stereo-IAT predicted prejudice against women leaders. We interpret this as an exploratory step toward pinpointing the specific biases that do and do not predict a specific response. Collapsing all three measures in a meta-analysis, however, would obscure this result and likely suggest that gender Stereo-IATs are simply unreliable.

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*.[21]

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. Recent research using (what we dub) Evaluative-Stereotype IATs (ES-IATs) is particularly valuable in this regard. 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, Rudman and Kilianski found that a highly specific evaluative stereotype predicted dislike of women leaders:

> Thus, 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)

Using a racial ES-IAT, Rudman and Lee (2002) found that listening to violent and misogynist hip hop increased racial 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, e.g., about the man's popularity. Stereotype application was better predicted by the ES-IAT than by explicit measures. And 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).[22]

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

[21] And positive stereotypes about ingroup members, although we are less confident than Brewer (1999) that "pure" ingroup favoritism is "more" of a culprit than "pure" outgroup derogation and comparative ingroup/outgroup preferences.

[22] In some recent work, Rudman and colleagues (2012) seem to have come around to Amodio and Devine's point of view, however: "When stereotype IATs are valenced (e.g., when warmth is contrasted with coldness), they assess evaluative rather than semantic associations (e.g., Rudman, Greenwald, & McGhee, 2001)." We think the earlier conceptualizations are more accurate—ES-IATS measure jointly evaluative *and* semantic associations.

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.

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. Rather than racial IEs and ISs being simply *unrelated*, SCM 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.[23]

Earlier we raised concerns about the claim that cognitive and non-cognitive states predict different types of behavior. 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?). In short, when it comes to predicting behavior, evaluative stereotypes are where the action is.

---

[23] Identifying the right stimuli for ES-IATs is likely a matter of trial and error. While Ebert (2009) found that implicit liking of women correlated with women-warmth associations, Rudman and Goodwin (2004) failed to detect a correlation between gender IEs and ISs. Rudman and Goodwin's Stereo-IAT had included terms related to both warmth and power, and it is possible that the influence of men-power associations limited the ability to measure women-warmth associations. Combining two stereotypes in a single implicit measure introduces the possibility that ingroup favoritism and halo effects could conceal implicit stereotypes, or that the activation of one stereotype overrides, enhances, or inhibits another stereotype. (Relatedly, Wade and Brewer (2006) measured warmth and competence associations toward businesswomen and homemakers in a single LDT. They found a general effect for valence (more positive associations toward homemakers) but no stereotype-specific effects. Distinctive LDTs for measuring warmth and competence separately may have led to very different results.) Ultimately, it is an open empirical question how best to refine the stimuli to target specific constructs and predict specific behaviors. Although it makes sense in hindsight, it would have been difficult to predict in advance that gender-authority ISs predict prejudice against women leaders while ISs related to careerism, domesticity, and agency do not.

**Toward Better Interventions**

We have raised several questions for the two-type model as well as considered it in light of influential accounts of explicit stereotypes and prejudices. We have made some tentative proposals for how to understand the co-activating nature of cognition and affect in implicit attitudes, and for future research to improve the predictive validity of indirect measures of attitudes. We conclude by considering the relevance of theoretical conceptions of implicit attitudes for creating effective interventions to combat bias.

Before advancing our own proposals, we first consider how the two-type model appeals to the IE/IS distinction to motivate strategies for practical intervention. For example, Gilbert, Swencionis, and Amodio (2012, 3609) endorse the "theoretical proposal 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 the independence of IEs and ISs, then, is said to be that we should recondition them separately—but perhaps we should infer just the opposite.

We have expressed doubts about the extent to which stereotypes and evaluations come apart in implicit social cognition, but suppose they are indeed dissociable. Even so, this fact might represent more of a cautionary tale about what *not* to do than an organizing principle for bias interventions. An intervention that seems to reduce IEs might leave ISs intact, 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.

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). 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 not that two interventions are more time-consuming than one. Our 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).[24] In general, learning is facilitated by combining information with affective and motivational allure. Just watch a TED talk to see this.

While we find little evidence that combating 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 trying to pinpoint 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 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). 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. 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 finding 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

[24] See Adcock et al. (2006) and Cohen et al. (2014) for examples of how motivation and value promote memory.

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.

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.

Amodio, D. M., Bartholow, B. D., & Ito, T. A. (in press). 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. (2008). 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). Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In B. Gawronski and B. K. Payne (Eds.) *Handbook of implicit social cognition* (pp. 353-374). New York: Guilford.

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

Aristotle (long time ago, galaxy far away). *Politics*. Translated and published by somebody.

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.

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

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.

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., & Dunantlaan, H. (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.

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

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

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.

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.

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

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.

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, 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.

Gillihan, S. J., & Farah, M. J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience.*Psychological bulletin*, *131*(1), 76.

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., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Holroyd, J. & Sweetman, J. (Forthcoming). "The Heterogeneity of Implicit Biases." Brownstein, M. and Saul, J. (Eds.) *Implicit Bias & Philosophy: Volume I, 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.

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.

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

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., & 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

Park, B., & Judd, C. M. (2005). Rethinking the link between categorization and prejudice within the social cognition perspective. *Personality and Social Psychology Review*, *9*(2), 108-130.

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.

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., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology, 87,* 494-509.

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. (in press). Reactions to gender egalitarian men: Feminization due to stigma-by-association. Group Processes and Intergroup Relations.

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

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of social issues*, *41*(3), 157-175.

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.

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.