1.2 Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition

Bryce Huebner

Recent data from the cognitive and behavioral sciences suggest that irrelevant features of our environment can often play a role in shaping our morally significant decisions. This isn't always a bad thing. But our inability to suppress or moderate our reflexive reactions can lead us to behave in ways that diverge from our reflectively held ideals, and to pursue worse options while knowing there are better ones available (Spinoza 2002, 320). Nowhere is this clearer than it is where racial biases persist in those who have adopted egalitarian ideals. My primary aim in this paper is to sketch a computational framework of implicit biases, which can explain both their emergence and their stability in the face of egalitarian ideals; I then use this framework to explain why some strategies for intervening on implicit biases are likely to be more successful than others. I argue that there are plausible ways to adjust our goals and manipulate our local environments to moderate the expression of implicit bias in the short-run, but I also maintain that the dynamic nature of learning and valuation, as well as the impact of stress on cognitive processing, necessitate more comprehensive interventions designed to re-shape the cognitive niche we inhabit. Put bluntly, the freedom to act on what we know to be best can only arise in a world where our reflexive and reflective attitudes are, by their very nature, already aligned. So if we wish to gain control over our implicit biases, we must intervene on the world to which our attitudes are attuned, and do so in a way that instills anti-racist attitudes.

1. The dual-system hypothesis

The phenomenon of implicit bias has become quite familiar.[1] White people often display *reflexive* signs of distrust and aversion, including increased blinking rates and decreased eye contact, when they interact with black people (Dovidio et al 1997). People are more likely to misidentify someone walking along a dimly lit street as dangerous if he is black; and many white people routinely feel the urge to cross the street to avoid black people. Such behavior undoubtedly helps to sustain racialized hierarchies, perpetuates racist assumptions about social status, and affects the viability of multiracial interactions. More troublingly, the assumption that black men are likely to be dangerous can, and often does, have a significant impact on prosecution rates and sentencing decisions, and it can play a deeply tragic role when quick decisions are made about whether a person is carrying a gun

[1] The numerous forms of implicit bias are likely to diverge in many ways from those I consider in this paper. I focus primary on the American context, and use the terms 'white' and 'black' as shorthand for categories to which contemporary North Americans are likely to be attuned. The mechanisms I discuss below are likely to track things like 'typically-African-American' features, and differences in social transactions may modulate the strength and scope of these associations in ways that shape the typical characteristics of the people who are racialized as white and black.

(Correll et al 2002). In any particular case, there is room to debate about the relative contribution of explicit prejudice and implicit bias, but it is hard to deny that racial stereotypes can—and often do—have a serious and deleterious impact on morally salient behavior.

Philosophers have long acknowledged that we sometimes act on habits or impulses that seem misguided on further reflection, and that we sometimes make judgments that we are unwilling to reflectively endorse. But implicit biases prove fiendishly difficult to extinguish, even when we acknowledge that race shouldn't be treated as a predictor of social status or threat (Dunham, Chen & Banaji In press; Gregg, Seibt, & Banaji 2006; Huebner 2009). This suggests a deep distinction between our reflexive and rational motivations, which most existing accounts of implicit bias see as evidence of two types of cognitive systems: a slow, controlled, inferential *system* that produces reflectively endorsed beliefs; and a heterogeneous network of systems that rapidly and reflexively generate expectations based on prior experience, yielding "attitudes and beliefs that are acquired passively without individuals' awareness and that influence subsequent judgments, decisions, and actions without intention or volition" (Dasgupta 2013, 235). There are debates about the precise commitments of such dual-systems hypotheses, and there are many ways of spelling out the difference between these kinds of systems (Evans & Stanovich 2013). But for now, what matters is that some version of this hypothesis has played an integral role in structuring the investigation of implicit bias—and it's easy to see why.

The films we watch, the news we read, and the narratives that dominate the public imagination typically present racial out-groups in ways that highlight their stereotypical features. This makes an associationist hypothesis about the origin and stability of implicit bias seem plausible (Devine 1989): If socially salient attitudes are encoded automatically and stored associatively, *exposure* to these stimuli should generate expectations about the attributes likely to be associated with particular groups; where such associations are triggered by encounters with out-group members, this will yield biased behavior and biased judgments. Put schematically:

> When people encounter a person, a group, or an issue they are familiar with, the attitude or belief associated with it pops into mind quickly and automatically in a split second. People may be unaware of attitude activation or only semiaware of it. But once an implicit attitude or belief is activated, it is difficult to inhibit or suppress right away and the activated attitude or belief is more likely to drive subsequent behavior, judgments, and decisions (Dasgupta 2013, 236).

It has often been assumed that such attitudes are internalized gradually through repeated exposures to culturally dominant presentations of groups. And on this approach, it becomes unsurprising that such attitudes often conflict with the beliefs we later consciously endorse.

The hypothesis that associative mechanisms play an important role in the encoding and retrieval of implicit attitudes also guides a great deal of research in social psychology. Indeed, many tools for examining implicit bias are designed to track such associations. For example, participants in the *implicit association* test (IAT) are asked to rapidly categorize two groups of people (black vs. white) and two attributes ('good' vs. 'bad'). Differences in response latency (and sometimes differences in error-rates) are then treated as a measure of the association between the target group and the target attribute (Greenwald, McGhee & Schwartz 1998). Similarly, the Go/No-Go association task (GNAT) presents a target group (black people) and a target attribute (good), and uses accuracy and response latency as measures of associative strength (Nosek & Banaji 2001). In both cases, faster responses and lower error rates are seen as evidence about the strength of non-inferential and associative computations. Such experiments reveal attitudes that are sensitive to associative considerations, and as the dual-system hypothesis predicts, these patterns persist in people who endorse egalitarian attitudes. But this isn't enough to demonstrate that our biases are implemented by non-inferential and associative computations.

As Eric Mandelbaum (in prep) notes, many prominent models of associative processing assume that associative representations must be encoded gradually and that they are modified only by changes in contextual or situational cues. But statistically robust effects in an IAT can be induced by asking participants to suppose that one imaginary group is bad and another is good—and this novel association can be transferred to a new group using similarly abstract suppositions (Gregg, Seibt, & Banaji 2006). This suggests that the emergence of implicit associations doesn't *always* depend on the repeated pairing of group members and evaluative attributes, which would be necessary for gradual associative encoding (though these data do not rule out the possibility that *other* types of associations are formed gradually).[2] More intriguingly, abstractly induced associations cannot always be extinguished with similarly abstract suppositions. Once they are encoded, they rapidly solidify in a way that suggests one-shot learning. Further, although changes in environmental contingencies *can* have a significant impact performance on performance in an IAT (Dasgupta 2013; Gawronski & Bodenhausen 2011), associative strength can also be modulated using logical and inferential strategies (Mandelbaum in prep). For example, reading and considering persuasive arguments about a group can significantly affect subsequent performance in an IAT (Briñol, Petty, & McCaslin 2009); and learning that your peers have prejudiced beliefs similar to your own can enhance the accessibility of biases, modulating response times in a lexical decision task and affecting a white person's decisions about how close to sit to a black person (Sechrist & Stengor 2001). A purely associative model would not predict either effect, though—as I argue below—there may be ways of explaining such effects using a computational model that makes a great deal of room for associative processing. So, where does this leave us?

Unlike the beliefs we reflectively endorse, implicit attitudes arise rapidly and automatically, often outside of conscious awareness. They track some types of associative relations, and they do so in ways that are difficult to modify except by changing environmental contingencies. Such facts are commonly taken to support a dual-system hypothesis, according to which implicit attitudes are implemented by associative and affective systems, while reflectively endorsed beliefs are implemented by a controlled and inferential system. But recent data have revealed that inferential reasoning and one-shot learning can sometimes affect implicit attitudes. This would be surprising if such attitudes were implemented exclusively by associative systems. But while the existing data strongly suggest that implicit biases are not implemented *exclusively* by associative mechanisms, there

[2] This experiment targets imaginary groups using an IAT. But as Michael Brownstein (p.c.) notes, the IAT is only one measure of implicit attitudes, and there are many ways of changing the scores on an IAT without affecting implicit bias *per se*. If taking an IAT is a partially controlled behavior, for example, this effect is unsurprising. I return to a similar issue in Section 3.

is no obvious reason to suppose that only one type of mechanism is implicated in the production of implicit attitudes. Most dual-system hypotheses explicitly allow for the operation of numerous fast and automatic systems, often in parallel, but this claim is often underdeveloped. My aim in the next two sections is to show how associative and inferential systems can *collectively* guide our behavior. To this end, I sketch a computational account of implicit biases, which is grounded in a "neurobiologically plausible and mechanistic description of how values are assigned to mental representations of actions and outcomes, and how those values are integrated to produce" morally significant decisions and morally significant behavior (cf., Crockett 2013).

The computational theory I advance is consistent with the existing behavioral and neuroscientific data on implicit bias, and my argument is made probable by the fact that the systems I discuss play a critical role in many types of learning and decision-making. Furthermore, recent theoretical and empirical perspectives have suggested that these systems are likely to be operative in the production of reflexive and automatic moral judgments (Crockett 2013; Cushman 2013; Huebner 2013; Railton 2014). But I cannot establish *conclusively* that this is the correct story about how our implicit biases are produced and maintained; decisive support for this computational theory therefore awaits further empirical investigation.

2. Three kinds of systems

Across the phylogenetic tree, associative learning is the primary means by which organisms represent the structure of their world, and there is reason to believe that associative learning systems play an important role in human social cognition (Crockett 2013; Cushman 2013; Heyes 2012). But human decision making is not exhausted by associative processing, and there is growing evidence that many of our decisions and actions depend on contributions from three distinct types of systems, which frequently operate in parallel: associative Pavlovian systems that reflexively trigger approach and withdrawal in response to biologically salient stimuli; associative model-free systems that reflexively assign values to actions on the basis of previously experienced outcomes; and computationally expensive model-based systems that generate forward-looking decision trees to represent the contingent relationships between potential actions and outcomes, and then assign values to these action-outcome pairs (Crockett 2013). In this section, I explain how each of these three systems operates in the context of human cognition more generally; and in the next section, I return to the role that these systems are likely to play in the production and maintenance of implicit bias.

2.1. Pavlovian learning and decision-making

Pavlov famously observed that associations could be induced by ringing a bell each time he gave food to a dog; over time, the dog gradually began to salivate whenever the bell rang, even when there was no accompanying food. This widely replicated result suggests that associations linking an innate response to biologically salient rewards or punishments can be formed *passively*. Capitalizing on this fact, contemporary approaches to associative learning emphasize the role of experienced discrepancies "between the actual state of the world and the organism's representation of that state. They see learning as a process by which the two are brought into line" (Rescorla 1988, 153). The algorithms that facilitate associative learning are error-driven: they generate predictions about the associative links between events, and update their predictions when surprising outcomes are observed; and the strength of the resulting associations is a function of updating the prior associative strength in light of the difference between a weighted average of past rewards and the most recently experienced reward (Glimcher 2011, 15650).

Pavlovian systems can learn at different rates, depending on the kind of information they track: some systems learn gradually, and depend on repeated exposure to an associative relation; others learn rapidly, treating recent experience as authoritative. But once these systems are calibrated against a pattern of rewards and punishments, they often become inflexible (Dayan et al., 2006). To see why, it will help to consider the type of food aversion that is learned rapidly as a result of stomach discomfort (Garcia & Koelling 1966). Such associations are difficult to extinguish because the offending food will be avoided at almost any cost. But even if it is eaten again, the system that guides this aversion did not evolve to encode associative relations with things that do-not-make-you-sick. The unsurprising upshot is that gustatory learning evolved to rapidly encode predictions about poisonous substances, and to generate long-lasting aversions to things that are potentially poisonous.

A network of mechanisms in the ventral striatum, basolateral amygdala, and orbitofrontal cortex represent associative relations between stimuli and biologically significant events (Niv 2009, 22), and there appear to be many different Pavlovian systems that evolved to track things as diverse as food, water, sexual reward, and social exchange. Each of these systems is narcissistic, in the sense that it only 'cares' about the associative relations that are salient to its own purposes (Akins 1996). But since biologically salient situations often involve multiple sources of reward and punishment, these systems operate in parallel, leading to competitions for the guidance of biologically and socially significant actions (Rescorla 1988). Such competitions must be settled by systems that evaluate the extent to which some value ought to be acted on, and this is why biological cognition also requires model-free systems that can reflexively assign value to actions on the basis of previously experienced outcomes.

2.2. Model-free learning and decision-making

Not all rewards grow on trees, and even those that do must often be cultivated. Pavlovian learning on its own can attune us to the *presence* of existing rewards and punishments, but something more is necessary when we must decide which rewards to pursue. Put differently, Pavlovian systems encode passively associations between an innate response and a biologically salient reward, but model free systems are necessary where we encode learned associations between actions and outcomes. Like savvy investors, we should only adopt a plan of action if doing so is likely to make us better off in the long run; this requires forming expectations about valuable outcomes, as well as estimating the likelihood of those outcomes given our chosen plan of action. Of course, it would be nice if we could carefully map out all of the alternative possibilities and evaluate them before making such decisions, but the world rarely cooperates. We are often forced to make choices on the basis of our current experience, either because we lack information that would be necessary to build such a plan, or because we do not have time to construct and evaluate all of the relevant models of how things *might* turn out. Nonetheless even rapid decisions must be sensitive to the *cost of acting* and the *value of expected payoffs*. And fortunately, evolution has found a clever solution to the problem of rapid evaluative decision-making, which is commonly known as the Law of Effect: actions that tend to produce positive outcomes are repeated, while actions that tend to produce negative outcomes are avoided. In many environments, this strategy is quite successful, and it has been preserved as the core of model-free decision-making.

Beginning in the mid-1990s, research in computer science and cognitive neuroscience converged on an account of the predictive and associative algorithms that make this type of instrumental decision-making possible. Model-free algorithms were discovered, which generated predictions about future outcomes on the basis of current sensory cues. In their most general form, these algorithms were designed to compute a prediction-error signal whenever outcomes were *better* or *worse* than expected*,* given the current value of a reward and the expected future value of that reward—which could be computed by way of a temporal derivative (Sutton & Barto 1998). These algorithms monitor the discrepancies between predictions and outcomes, and they adjust future predictions in light of experienced discrepancies; over time, this yields a form of error-driven learning that allows a learner to become attuned to the stable patterns of rewards and punishments in its environment. These algorithms thus lead an entity to pursue positive outcomes while avoiding negative outcomes. Simultaneously, neuroscientists discovered phasic spiking activity of midbrain dopamine neurons that was consistent with such algorithms: this activity increased when outcomes were *better-than-expected*, decreased when outcomes were *worse-than-expected*, and was unaffected by outcomes whose time and value are accurately predicted (Montague et al., 1996; Schultz et al 1993). Bringing these two insights together, it was suggested that these midbrain neurons function as "detectors of the 'goodness' of environmental events relative to learned predictions about those events" (Schultz, Dyan, & Montague 1997). In the intervening years, numerous tasks, using numerous sources of reward, and species ranging from honeybees to nonhuman primates, have revealed a midbrain network that computes a multimodal and polysensory teaching signal that directs attention, learning, and action-selection in response to valuable outcomes (Schultz, 1998, 2010).

While there are still debates over the precise computational algorithms that are implemented by model-free systems, there is broad consensus that risk- and reward-based decision-making recruit a network of evolutionarily old, computationally simple systems, which allow us to produce, maintain, and revise behavioral policies by tracking patterns of reward and their predictors (Glimcher 2011). Indeed, recent neuroscientific studies have revealed model-free systems in the midbrain and orbitofrontal cortex, which collectively implement the learning signals and motivational "umph" required to get habitual learning off the ground (Rangel et al., 2008; Liljeholm and O'Doherty, 2012). Mechanisms in the ventral striatum compute expectations when the distribution and likelihood of a reward is uncertain, while distinct circuits in the ventral striatum and anterior insula evaluate risk and compute risk-prediction-error signals (Preuschoff et al., 2006, 2008; Quartz, 2009). Another network including the orbitofrontal cortex and basal ganglia represents reward values in ways that are sensitive to both the probability of positive outcomes given recent gains and losses (Frank & Claus 2006; Shenhav & Greene 2009), and relative satiation (Critchley & Rolls 1996). Finally, and most intriguingly, the failure to conform to social norms evokes activity in the ventral striatum and posterior medial prefrontal cortex (mPFC), which is consistent with a prediction-error signal whose amplitude covaries with the likelihood of adjusting behavior to restore conformity (Klucharev et al 2009); norm conformity can even be decreased by interrupting the activity of this circuit using transcranial magnetic stimulation (Klucharev et al 2011).

Considerations of computational economy and efficiency militate against coding for multiple sources of reward when all necessary information can be encoded more simply; as a result model-free systems monitor for prediction errors at many points in temporally extended chains of events—this is what allows them to treat current experience as evidence for the success or failure of a long-term behavioral policy (Dayan 2012).[3] We can ignore the mathematical and neurological complexities of these algorithms for now, as the import of this fact is fairly intuitive. Model-free systems initially respond to rewarding outcomes, but where there are clear predictors of those outcomes, they become more sensitive to the presence of those reward-predicting stimuli. For example, such a system may initially respond to the delicious taste of a fine chocolate bar. But when this taste is repeatedly preceded by seeing that chocolate bar's label, the experience of seeing that label will be treated as rewarding in itself—so long as the label remains a clear signal that there is delicious chocolate is on the way. Similarly, if every trip the chocolate shop leads to the purchase of that delicious chocolate bar, entering the shop may come to predict the purchasing of the chocolate bar, with the label that indicates the presence of delicious chocolate; in which case entering the shop will come to be treated as rewarding. And if every paycheck leads to a trip to the chocolate shop...

The upshot is that model-free systems can track predictors of rewards, predictors of predictors of rewards, and more. This allows the reward signal triggered by the predictor of a reward to *stand-in* for the value of a behavioral policy; where things go better or worse than expected, given that policy, prediction-error signals can be used to revise these behavioral policies to make valuable outcomes more likely (Balleine, Daw, & O'Doherty 2008). But these are error-driven algorithms, and they can *only* adjust behavior where things go better or worse than expected. This type of learning is sometimes quite successful, as it is when the receipt of a reward is wholly contingent upon the actions taken in its pursuit. Indeed, as systems that treat the earliest successful predictor of rewards as *stand-ins* for the success of a long-term behavioral policy, they will be relatively successful in any environment where action-outcome contingencies are relatively stable.

Where ecological and social pressures produce stable environmental contingencies, the Pavlovian systems discussed above will yield motivations to approach beneficial things and to avoid harmful things. Similarly, when ecological and social pressures produce stable environments where an agent can expect the future to be relevantly like the past, model-free systems will facilitate the formation of behavioral policies that bring about good results in the long run. But like Pavlovian

[3] Changes in the phasic spiking rates of dopamine neurons are consistent with an ability to track exponentially weighted sums of previous rewards, and computational models reveal that this activity yields a prediction-error signal that can be adjusted to fit a temporally extended policy in light of the experienced value of recent rewards (Glimcher 2011, 15653). This is why the repeated experience of the predictor of a reward leads to a shift in the phasic activity of these neurons: rather than firing immediately after a reward is received, they fire when the earliest predictor of that reward is observed.

learning, model-free systems rely on error-driven associative processes that are difficult to modify once they are fixed, and they tend to yield suboptimal outcomes "where traditionally rewarding actions lead to undesirable outcomes, or vice versa" (Crockett 2013, 363).

More troublingly, the mere adoption of a new behavioral policy, which is grounded on new predictions about action-outcome contingencies, can leave prior predictions in place—although these associations may fade into the background. For example, failing to find one's favorite chocolate bar may lead to the purchase of a different one, yielding a new preference, and a new behavioral policy. But this shift in behavioral policies may leave the associative link between the original label and chocolate bar intact, since no errors were ever detected in *that* predicted link. Moreover, since statistical regularities do not always align with our reflectively held ideals, such systems can get us into trouble. Pavlovian and model-free systems are designed to get the patterns of environmental contingencies and the patterns of action-outcome contingencies right. So where an environment is organized in ways that leads us to pursue bad ends, we must adopt other decisions making strategies, and preferably strategies that allow us to act with forethought, choosing between different potential courses of action by anticipating future outcomes (cf., Prinz 2002, 4).

2.3. Looking forward, and model-based decisions

One way of acting with forethought is by re-purposing model-free systems to evaluate *merely possible* situations, and this strategy appears to be operative in human cognition. In a recent experiment where stock values were changed *after* decisions about whether to buy them had been made, activity in the ventral caudate was sensitive to differences between how much *could have been* won and how much actually was won (Lohrenz et al 2007). This activity is consistent with a fictive-error signal, that is, a signal that reflexively compares actual outcomes against *things that might have been.* Such fictive-error signals provide an initial way for us to test our actions before carrying them out; and the systems that produce these signals are integrated with the action-guiding systems discussed above. However—and this point will be key below—these fictive-error signals are computed by model-free systems, and they compete with the signals computed by other model-free systems, contributing to the guidance of action by strengthening or weakening the force of existing associations. In a striking demonstration of this fact, it was shown that the brains of smokers produce and ignore fictive-error signals in ongoing behavioral choice (Chiu et al 2008). The competitions between associative processes reveal important computational limitation on human decisions making, and they will be significant to my discussion of the architecture of implicit bias in Section 3. But before addressing this issue, I must address a final class of decisions-making systems that are present in human cognition.

We are not simply stimulus-response machines. While it is often difficult to do so, we can sometimes construct and evaluate counterfactual models of the world that allow us to predict the consequences of our actions before we act. Indeed, there is growing evidence that systems dedicated to model-based reasoning allow us to produce decision trees that represent the value of various potential actions-outcome pairs, and then search these trees to determine which actions are likely to produce the best outcomes overall (Crockett 2013, 363).[4] This process relies on mechanisms in the prefrontal cortex that represent stored goals and values in working memory. And while model-based decisions can be more accurate and flexible than those produced by predictive learning systems, "their computation is costly in terms of neural resources and time" (Niv 2009, 21). Furthermore, situational or endogenous variables can modulate the salience of our goals, values, and ideals, and they often do where they conflict with representations produced by simpler associative systems (Braver & Cohen 2000). This is important, as there are many cases where the outputs of model-based systems will not converge on the same decision that would be made on the basis of Pavlovian or model-free values—and conflicts between these systems must be settled by way of comparative evaluations that determine which representation is most salient in the current situation (Dennett 1991; Huebner 2014; Selfridge 1959). Increased activity in PFC and insula is often observed when a decision is made under risk or uncertainty; and increased activity of the ventromedial PFC is often observed when distinct reward and avoidance values must be converted into a 'common currency' to allow for a comparative evaluation (Levy & Glimcher 2011; Montague & Berns 2002; Montague, Hyman, & Cohen 2004). By contrast, activity is increased in the dorsolateral PFC when ongoing behavior is organized in light of existing goals or values (D'ardenne et al 2012; Miller & Cohen 2001; Knoch et al. 2008); and the there is evidence that the insula serves as the interface between model-based, model-free, and Pavlovian systems, which allows for the silencing of these systems when computational resources are abundant (Bechara, 2001; Moll et al 2006).

There is a great deal of evidence that our decisions, as well as our behavior, are typically produced by aggregating the 'votes' cast by Pavlovian, model-free, and model-based systems in support of their preferred actions (Crockett 2013; Daw et al 2011; Huys et al 2012). According to this view, the attitudes we express at a particular time causally depend on the outputs of multiple systems; this yields a more dynamic account of attitudes, as their contents can shift as we triangulate stable model-based representations against the dynamic evaluative representations that are encoded by Pavlovian and model-free learning systems. In the remainder of this chapter, I argue that our implicit and explicit attitudes are also likely to reflect the combined influence of these three types of computationally and psychologically distinct evaluative systems. This is often hard to see, because the 'attitudes' examined by social psychologists are often stabilized as components of large-scale behavioral dispositions (cf., Machery this volume). In part, I will argue, this is because there are robust regularities in the world to which Pavlovian and model-free systems can readily become attuned. By pulling apart the relative contributions of these systems to our biased decisions and behavior, I hope to make it clear how both inferential and environmental interventions can contribute to changes in our behavior; and I hope to provide some insight into the reasons why various short-term and long-term interventions on bias are as successful as they are. But first, I must explain how implicit biases could be produced and sustained by Pavlovian, model-free, and model-based mechanisms.

[4] I cannot defend the claim that such systems operate by constructing and evaluating decision trees in this paper, as doing so would take us far afield from my central argument. At present, tree-based searches offer the most promising account of model-based cognition (cf., Crockett 2013; Daw et al 2011; Huys et al; 2012).

3. A plausible architecture for implicit bias

We reflexively see human behavior as indicative of psychological dispositions (Gilbert & Malone 1995), and the information we encode about group membership allows us to rapidly draw inferences about the features of group members on the basis of incredibly sparse information. This capacity is part of our evolutionary inheritance (Mahajan et al, 2011), and it emerges early in development. By the age of 4, children already make inferences about psychological properties on the basis of group membership (Diesendruck & Eldror 2011), and they will even use linguistic labels designating group membership to guide inferences about the psychological and behavioral traits of new members (Baron et al 2013). Our tendency to classify others is triggered by the presentation of even minimal information (e.g., "you belong to the red group"), and this feature of our psychology would be difficult, if not impossible to eliminate. But there is a great deal we can do to moderate its impact, for this capacity is subserved by a content-poor mechanism that must be calibrated against environmental feedback to yield biases; put differently, while we may have evolved to distinguish between in-groups and out-groups, it is unlikely that we evolved to treat race, gender, and ability as grouping criteria—and doing so requires setting the parameters on general purpose mechanisms for social cognition. This is where the three systems I have discussed become critically important.

Let's start with Pavlovian systems. Intriguingly, images of group members with shared physical characteristics are insufficient to license category-based inference in 4-year-olds (Baron et al 2013). But the fact that they do not yet show this tendency shouldn't be too surprising if learning about groups on the basis of visual information recruits Pavlovian algorithms that can be attuned to the co-variations between observed physical features and the experience of biologically significant reactions (e.g., fear, threat, dislike, or reward). Since there is no intrinsic relationship between race and danger, learning these reactions has to occur in a rather round about way. But there are evolutionarily old mechanisms that allow us to track the emotional reactions of our friends and family, and we can treat these signals as indications of which things are threatening or dangerous; of course, we also reflexively monitor the dangers and rewards presented in the media where the members of racialized out-groups are often represented as sources of danger, threat, or sexual reward. Pavlovian mechanisms could slowly attune to these racialized representations, yielding attitudes grounded on fear, disgust, and sexual lust. Were racialized out-groups *only* experienced as sources of threats and dangers, the predictive algorithms employed by Pavlovian systems would yield fully calcified associations that would always guide behavior in light of (often mistaken) predictions about the dangers, threats, and rewards that interactions with out-groups will afford. But even less totalizing representations could drive strong avoidance reactions, which could lead us to suppress behavior and avoid situations where we predict (often mistakenly) that an aversive outcome is likely.[5]

The hypothesis that some implicit biases are implemented by Pavlovian algorithms gains support from the fact that faces of racialized out-group members trigger increased activity in the

[5] While such associations are likely to be pervasive, they are unlikely to be totalizing. While young black men may be tracked as threats, they may also be tracked as successful athletes; and young black women are likely to be primarily treated as sexual objects. In short, the even the Pavlovian associations encoded by a person are likely to be a heterogeneous lot, and this will cause all sorts of problems in attempts to intervene on these representations, I return to this point in Section 4.

amygdala (see Stanley, Phelps, & Banaji 2008 for a review). This effect is even observed when white participants are exposed subconsciously to photos of black people (Cunningham et al. 2004). Although the amygdala has long been thought to play an important role in fear conditioning, more recent data suggest that it also evaluates the biological and social *salience* of stimuli in ways that track more abstract patterns of risk and reward (Adolphs 2010). Amygdala lesions reduce loss aversion, increase risk-taking and social curiosity, and make people who appear unapproachable and untrustworthy *seem* more trustworthy and approachable (Adolphs, Tranel & Damasio 1998; Demartino, Camerer, & Adolphs 2010). And connections to the striatum, allow the amygdala to generate an avoidance signal based on the risks associated with socially and biologically significant stimuli. I contend that when the faces of racialized out-group members trigger increased activity in the amygdala, this is likely to be because participants are reflexively evaluating potential dangers, risks, and indicators of trustworthiness when they observe the face of a member of a racialized out-group.[6]

In socially relevant interactions, multiple Pavlovian associations may be relevant to rapid decision-making and the guidance of behavior. A variety of partially overlapping associations could be encoded for the members of a particular racial group, and numerous associations will often be applicable to a single person (since every person belongs to multiple groups). In this case risk-categorization would become a dynamic process in which numerous factors are monitored in parallel and evaluated for salience relative to a particular task (Quadflieg & Macrae 2011, 221; Rescorla 1988). Where multiple Pavlovian associations are relevant to ongoing behavior, competitive algorithms must be employed to determine which associative representations are most salient in light of current task-demands. So approach and avoidance responses depend on the value of particular associations, as well as the extent to which distinct representations converge or diverge from one another. When multiple associations converge on a single response, that response will be facilitated, but where there are conflicts between associations, responses will be slower to come on-line (Huang & Bargh in press); as a result, the excitatory and inhibitory relations between multiple associations becomes a critical variable in explaining many kinds of human behavior. Faces that display typically-African-American features will elicit stronger stereotype activations than faces with less-prototypical features (Blair et al 2002); images of people wearing clothing that are atypical for an out-group member are less likely to evoke racialized judgments (Barden et al 2004); and seeing a person in an unexpected setting tends to evoke less biased responses (Wittenbrink, Judd, & Park, 2001). Put simply whether "looking young, black and male will elicit the activation of stereotypic beliefs along any (or all) of those dimensions depends on his unique appearance and the situation in question" (Quadflieg & Macrae 2011, 223). I maintain that is partly because of the complex relations between the underlying Pavlovian associations.

If this is right, environments that contain multiple overlapping sources of racialized reward and punishment will allow Pavlovian systems to attune to environmental contingencies in ways that

[6] As Brownstien (p.c.) notes, this hypothesis predicts that people who have not been exposed to racialized environments would not show this type of response. This is right, but I'm not sure how it could be tested in our current world. That said, there are data suggesting that *familiar* faces of well-respected Blacks do not trigger this same response in white participants; and there is reason to believe that this may reveal an effect of top-down control, such that the evaluations carried out by Pavlovain systems are inhibited as a result of familiarity (cf., Stanley, Phelps, & Banaji 2008).

sustain robust and stable behavioral dispositions. At the same time, we should expect the operation of these systems to be sensitive to individual differences in learning history, and the strength of encoded associations will be modulated by individual differences in impulsivity as well as risk-aversion. In line with the hypothesis that tasks like the IAT and the GNAT provide evidence about the strength of associations, we find general trends that are modulated by differences between populations, as well as contextual and situational cues (Greenwald et al 2009). We also find that different populations and different tasks evoke different patterns of activity in the amygdala (Kubota, Banaji, & Phelps 2012). These data are, however, also consistent with downstream effects on model-free and model-based evaluations. Indeed, existing models of social cognition suggest that signals from Pavlovian systems can suppress behavior that would otherwise be produced by model-free systems, and they can 'prune' the decision trees produced by model-based systems by eliminating options that are 'too aversive' (Crockett 2013). [7] So before assuming that the effects revealed by these tasks are a simple function of Pavlovian processing, we must consider the possibility that non-Pavlovian systems also play an important role in the production and stability of implicit bias.

Model-free systems are also associative, they monitor experience for prediction-errors and continually adjust their predictions in light of experienced rewards and punishments. But as I noted above, these systems are also sensitive to fictive-error signals, which allow for the adjustment of behavior in light of simulated outcomes, and to other abstract forms of social instruction (Crockett 2013; Dayan 2012). For example, when we use linguistic labels and other minimal types of information to guide *inferences* about the unobserved features of group members, this process must initially depend on the use of abstract representations, and perhaps even model-based computations. This is likely the reason why such tasks typically recruit activity in regions of the medial PFC that are commonly associated with *abstract social reasoning* (Cooper et al 2012; Mitchel, McCrae, & Banaji 2006). But connections between the amygdala and mPFC allow these more abstract impressions to rapidly be off-loaded onto model-free systems. Put simply, signals from the mPFC can modulate approach and avoidance behavior to drive the rapid construction of behavioral policies (Kim et al 2011). For example, in a novel categorization tasks, the solidification of biases that can be tracked using an IAT occurs quite rapidly (Dunham, Chen, & Banaji in press). As people carry out this task, they must mentally rehearse the new action-outcome associations, and in so doing they are likely to reflexively generate course-grained mental simulations that can train model-free systems. After all, model-free systems respond to real as well as imagined feedback, and these systems don't know the difference between the two. Finally, since there are no signals that errors are being made in the novel

[7] What would it take to show that these effects were the result of integrating multiple competing representations, as opposed to revealing a simple effect of Pavlovian processing? As I argue in the remainder of this section, there are both conceptual as well as empirical arguments that speak in favor of the competitive processing hypothesis. Since model-based, model-free, and Pavlovian processes *can,* and *often* do operate in parallel, it would be surprising if they did not do so in this case. Of course, alternative computational models may someday be developed, which would speak decisively in favor of more localized Pavlovian effects. But at present, the most promising "computational approaches to decision-making account for choices by adding up model-based, model-free, and Pavlovian action values, and then converting those values into action probabilities using a softmax function, essentially treating the three systems as separate experts, each of which 'votes' for its preferred action (Crockett 2013, 364). While this fact is not decisive, I hope that arguments I advance in the remainder of this section lend credence to this hypothesis. (Thanks to Jennifer Saul and Michael Brownstein for pushing me to clarify this issue).

categorization task—indeed, the participants are categorizing just as they are supposed to—this process of offloading could easily yield a self-perpetuating system, where the resonance between the initial forms of abstract reasoning and the new model-free policies become evidence that everything is going as planned.

In real-world environments, this situation is complicated by the fact that robust patterns of social feedback also line up with these simulations, causing habitual responses to solidify, and making it difficult to get rid of these associations using only model-based interventions. Model-free systems can track more abstract relations between actions and outcomes, and they can generate behavioral policies that are sensitive to social norms rather than perceived or imagined threats (Klucharev et al 2009, 2011). In typical situations, the result of this process is likely to be quite troubling. Problematic biases are often encoded through a process of verbal and behavioral instruction, which is used—carelessly—to teach children about how *we* respond to the members of racialized out-groups; warnings are made about the threats and dangers posed by an out-group, and about the situations in which these dangers are likely to arise. Model-free systems learn from this type of social instruction, and this could lead to the production of *habits* that attach negative (and more rarely, positive) values to engagements with the members of racialized out-groups. Where this type of information resonates with Pavlovian processing, or with consciously held beliefs that are also formed as a result of this instruction, this process could rapidly transform social feedback into patterns of habitual response. Where behavioral policies track social norms that systematically disadvantage racialized out-groups, the prediction-errors that would be relevant to the adjustment of these policies must take the form of evidence that the norms are different than expected (i.e., they must reveal that the predicted statistically normal practices were not in fact statistically normal!); for these backward-looking mechanisms, the fact that a particular groups is exploited or dehumanized will always be irrelevant—for error-driven learning is only sensitive to statistical norms. But in spite of the dangers of this process of training model-free systems by way of simulations or direct instruction, the fact that this is possible also suggests that it may be possible to *retrain* model-free systems using other types of simulations—and I return to this point in the final section of this paper. But first, I address the potential role of model-based systems, which could be used to guide behavior in accordance with our reflectively endorsed goals and values.

As I noted above, model-based systems can produce and search through decision trees, allowing us to determine whether a particular action will align with, for example, our commitment to egalitarian values. These systems can operate consciously or subconsciously, mapping out various potential actions and evaluating the extent to which they satisfy our stored representations of goals and values; as a result they could generate value-driven aversions to racist behavior and exploitation, as well as preferences for the promotion of anti-kierarchical ideals. I maintain that this is one of the primary reasons why studies of implicit bias often reveal increased activity in the dorsolateral PFC and anterior cingulate cortex (ACC) when people attempt to suppress their biases (Stanley, Phelps, & Banaji 2008). The ACC is commonly activated in situations that trigger conflicts between multiple evaluative systems; and in this case, there is likely to be a conflict between biased Pavlovian and model-free representations, and more egalitarian ideals. The activity of the dorsolateral PFC, by contrast indicates the deployment of working memory representations, which are employed in modeling and evaluating potential actions in light of our goals and values. There is evidence that these working memory systems in the PFC allow us to exercise control over ongoing behavior, using bias signals to modulate the activity of other evaluative systems in light of our representations of goals and values (Miller & Cohen 2001). This provides a way for implicit associations to be regulated using model-based evaluations. But importantly model-based systems operate in parallel to model-free and Pavlovian systems, and most biologically and socially salient choices will represent the combined influence of multiple computationally distinct systems (Crockett 2013).

So it is significant that these systems could produce conflicting pulls toward everything from the positive value of norm conformity (understood as attunement to locally common patterns of behavior), to the aversive fear associated with an out-group, and the desire to produce and sustain egalitarian values, among many other situation relevant values. Where the outputs of these systems diverge, each will cast a vote for its preferred course of action; where some of these votes are negative, this will yield a bias toward behavioral inhibition (though the size of this effect will always depend on the strength of these votes, as well as the excitatory and inhibatory relations between multiple representational systems); where there is conflict between the outputs of multiple systems, we will see the familiar increase in response latency revealed by measures like the IAT and the GNAT. The key point is that model-based, model-free, and Pavlovian systems can exert inhibitory and excitatory effects on one another, and that this fact is likely to have a significant effect on behavior (Huebner & Rupert 2014). In the IAT and GNAT, the task-demands may generate a working-memory signal that conflicts with the representations produced by Pavlovian and model-free systems. This would then yield a difference in latency for counter-stereotypical associations, because there is a conflict between distinct decision-making systems. But the point is not merely that there is a conflict between these systems, which is where the familiar analysis stops. By approaching the data from a perspective that highlights the existence of multiple competing systems, which are carrying out different types of computations, we can begin to understand a puzzling piece of data that is rarely discussed in the literature on implicit bias.

Patients with focal lesions in vmPFC show less bias in implicit association tasks, but they make explicit judgments that are indistinguishable from controls (Milne & Graffman 2001). The most promising computational models of mPFC suggest that these circuits play a critical role in translating the reward and avoidance values computed by model-free and Pavlovian systems into a common currency that can be triangulated against currently active goals (Levy & Glimcher 2011; Montague & Berns 2002). Damage to these circuits yields a pronounced tendency to make impulsive decisions on the basis of currently active task-demands. For example, people with lesions to vmPFC make high-risk decisions in the face of ongoing economic losses, and they accept risky bets even where they know that their odds of winning are vanishingly small (Clark et al. 2008; Saver & Damasio 1991). They also tend to accept more unfair offers in ultimatum games, suggesting an inability to adjust the value of accepting such offers against the reflexively computed value of punishing someone who behaves unfairly. (Koenigs & Tranel 2007). I maintain that the data reported by Milne & Graffman (2001) are also plausibly explained by the failure to integrate multiple values. Specifically, while stereotypical values encoded by Pavlovian and model-free systems are still available to guide some behavior, damage to mPFC prevents these values from being integrated with the currently salient goal of matching faces to evaluative categories.[8] In this experiment, though perhaps not in every other case, the impact of Pavlovian and Model-free values is swamped by the salience of the goal-directed representations that are operative in the task.

Building on this suggestion, we can begin to see why inferential processing sometimes has an effect on implicit bias (Mandelbaum in prep), and why inferential processes are less useful when working memory resources are depleted as a result of increased stress or because of increases in cognitive or affective load. People who are *strongly committed* to egalitarian goals and values show an increased capacity to inhibit or suppress the influence of stereotypes on their judgments. When people with chronic egalitarian goals attempt to compensate for their previously expressed biases, those with the strongest commitments to egalitarian values are less likely to exhibit implicit biases (Moskowitz et al. 1999). A similar effect is also present in tasks measuring shooter-bias (Glaser & Knowles 2008), which is even more striking since these goals impact incredibly rapid responses that must be made rapidly. Together, such data suggest "that a chronic motivation is capable of dominating even the strongest and fastest-acting conflicting responses" (Bargh 2006). But even without such chronic motivations, people can use consciously held goals to temporarily modulate reflexive responses. For example, a stimulus that typically evokes negative attitudes (such as a rat) can be *treated* as having a positive value in the context of a currently active goal; but as soon as this goal is inactive, the valence of that stimulus will revert to its default state (Ferguson & Bargh 2004).

There is evidence, however, that even minor stress can 'flip a switch' that can lead us to abandon computationally taxing model-based processing, and to rely on computationally cheaper forms of model-free or Pavlovian processing (Crockett 2013; Schwabe & Wolf 2013). While it is commonly noted that associative processing dominates judgment and decision-making under stress or cognitive load, this reveals a deep fact about the architecture of human cognition, and more specifically about the importance of more chronic goals in bias-relevant decision-making. The increased load imposed by making rapid judgments can increase reliance on Pavlovian associations and model-free processing, and this is likely to be the reason why many people are slower to offer egalitarian responses when asked to rapidly classify stereotype-consistent claims as true or false (Wegner 1994). It is also likely to be the reason why stress is likely to affect everything likelihood of arresting or prosecuting someone to decisions about the severity of a sentence, as well as rapid judgments about whether a black person is carrying a gun or behaving in a threatening manner (each of which appears to be a case where implicit biases have a significant impact). While chronic goals can help to moderate these responses, it is unwise to assume that they always will.

Beyond this fact, we must remember that the computational architecture of implicit bias is likely to be incredibly complex. As I noted above, most socially significant decisions require sifting through numerous sources of information to distill multiple parallel possibilities into a single value that can guide action (Bargh 2006; Levy & Glimcher 2011; Montague & Berns 2002). Even in the best of cases, where we are not under stress, and where we are not facing an increase in cognitive or affective load, the extent to which we rely on a particular association, or the extent to which our

[8] Crockett (2013) notes serotonin functioning modulates the salience of Pavlovian computations. If the model I am sketching is roughly correct, manipulations of serotonin should also modulate implicit bias, but the relevant hypotheses have yet to be tested.

decisions are guided by a particular value, will be sensitive to a wide range of situational factors that guide the online construction of action-guiding representations. As the brain attempts to find a way to sift through all of the representations that are relevant to our ongoing behavior, it must rely on a variety of wide variety of pressures and contextual cues that arise in a particular situation. Thus, cigarette smokers who show a negative implicit attitude toward smoking are less likely to display this attitude when it has been a long time since their last cigarette (Sherman et al., 2003). While the pursuit of nicotine rewards can often be inhibited by the model-based value assigned to not smoking, the value of that reward varies with changes in the neurochemical and social context in which a judgment must be made. In the same way, the negative Pavlovian values associated with walking through a particular neighborhood, or seeing someone who is wearing a particular type of clothing, may lead to decisions that privilege associative processing over consciously endorsed goals—yielding a biased aversion, or much worse. Precisely how such computations are carried out will depend on the context in a decision must be made; if so, this would have serious implications in moral-psychology, since both the strength and accessibility of implicit biases will vary across contexts.

I maintain that the strength of various associations, the impact of particular model-free policies and model-based goals, and the inhibitory and excitatory relations between multiple systems are crucial variables that ought to be taken into account in developing a mechanistic account of how implicit biases are produced (cf., Rupert 2011). In light of this fact, we should expect low between-measure correlations for different ways of examining implicit bias. Different tasks will not always be measuring a single, unified thing, which is properly called an attitude (Machery this volume). Instead, these tasks are likely to track the influence of multiple causal factors, all of which are sensitive to individual differences in reinforcement history, experimental context, and corresponding differences in the temporary processing goals evoked by a particular task. I contend that Machery (this volume) and Holroyd & Sweetman (this volume) are right to be skeptical of the claim that implicit bias is a unitary cognitive phenomenon. And I hold that Machery is likely to be right that what emerges in an individual are stable dispositional traits—my core claim is that these traits are likely to be produced and sustained by the complex and dynamic interaction between model-based, model-free, and Pavlovian systems, as well as the world to which the learning systems we rely on are attuned. By focusing on this fact, I suggest that we can begin to develop a more plausible account of where and when strategies for intervening on implicit biases are likely to prove successful, and my aim in the final section of this paper is to sketch an account of the most plausible types of interventions on implicit bias.

4. Three types of imagination

Suppose that our experience of the world is filtered through the lens of a massively parallel computational system that continually integrates the outputs from numerous representational mechanisms, each of which only produces representations that are salient to the tasks they have evolved and developed to tackle. Some of these mechanisms process information relevant to our current goals and values, others process information that arises through episodic memories or counterfactual reasoning, and still others generate low-level predictions about reward-based or action-outcome contingencies. On this view, it "is as if the mind constantly explodes the outside world into multiple parallel possibilities for action, but must then reduce and distill these back for use in a world in which you can only do one thing at a time" (Bargh 2006, 158). We rely on integrative mechanisms that obscure the massively parallel processing that lies behind our decisions and motivations. As Early Modern philosophers commonly recognized, this yields many confused representations of the world in which we live and act (*confundere:* mingled together). In many cases, these representations make it easy to assume that interventions focused on our immediate experience will have a significant effect on our future behavior. So, we imagine better ways to live and act, we commit to egalitarian worldviews, and we try to be better people. But as the literature on implicit bias suggests, such strategies are often insufficient to modify our behavioral dispositions; and to the extent that we fail to moderate or suppress our implicit attitudes, we often find ourselves acting on the basis of habit, or reactive fear—and often, as a matter of our encounters with the world, we find ourselves compelled to pursue worse things though we see that better ones are available (Spinoza 2002, 320).

In part, our failures to moderate and suppress our implicit biases derive from the fact that we are unaware of the causes of our behavior, and our attempted interventions leave many of the model-free and Pavlovian systems that guide our behavior untouched. They do little more than add another voice to the cacophony of representations that collectively guide our behavior. In some cases, this helps. An extra voice sometimes tips the balance toward more egalitarian behavior. But since model-free and Pavlovian systems are calibrated against the pervasive stereotypes and oppressive power relations that permeate our world, they continually push us back toward bad habits and problematic ideologies, an effect that is more pronounced when we are tired, stressed, annoyed, or under a high cognitive load—and we live in a world that fosters these states. When the awareness of our biases fades, we often find ourselves—as a result of our encounters with the world—compelled to act in ways that we cannot reflectively avow. So I maintain that the only way to effectively intervene on our implicit biases, and to free us from the problematic constraints imposed by Pavlovian and model-free processing, is to modify the internal and external environments to which our automatic behavior is attuned.

4.1 Just use your imagination

Many implicit attitudes are situational adaptations that are attuned to features of the racist, sexist, and heteronormative communities in which we are immersed (Dasgupta 2013, 240). But if error-driven learning mechanisms make such attunement possible, they will also make it possible to shift these attitudes by changing the patterns of socially salient stimuli to which we are exposed. In a striking confirmation of this fact, Dasgupta & Greenwald (2001) presented people with images and short biographies of admired and respected African Americans (e.g., Martin Luther King Jr. and Michael Jordan), and deeply despised White Americans (e.g., Jeffrey Dahmer and Timothy McVeigh); they then used an IAT (which participants tended to view treated as a hand-eye coordination task) to show that implicit racial bias was significantly decreased in people exposed to these stimuli—more strikingly, the effect persisted 24 hours later. Gawronski & Boddenhausen (2006, 698) claim that this effect is likely to have occurred because task demands led to enhanced activation of preexisting associations related to Black people; and this hypothesis gains support from the fact that evaluations of well-known people depend, at least in part, on whether they are categorized on the basis of their race or another socially salient feature: Michael Jordan elicits positive evaluations when he is categorized as an athlete, but negative evaluations when he is categorized as an African-American (Mitchel, Nozek, & Banaji 2003). There is also a great deal of evidence demonstrating that priming a person with a socially salient category facilitates associative processing in an IAT (Gawronski & Boddenhausen 2006). As I noted above, the situation in which such decisions are made does have a significant biasing effect on the integration of associative representations. But something is not quite right with this hypothesis. Why should thinking about Michael Jordan and Martin Luther King Jr. affect the categorization of unfamiliar faces, and why should it have this effect 24 hours later when people are not reminded of the primes and believe that the IAT is really a hand-eye coordination task? There is every reason to believe that these participants would have had other preexisting associations related to black people, and that many of these associations would have affect response time in the other direction. On the assumption that they did not see the racial primes as related to the initial IAT, there is little reason to suppose that the same preexisting associations that were deployed in the first phase of the task would be triggered the next day.

The account of implicit bias I advanced earlier offers insight on this point. Gawronski & Boddenhausen (2006, 698) are right that using priming tasks can impact the competitions between multiple systems. But they do not see the possibility of using imaginative representations to co-opt associative learning mechanisms, and to adjust our reactions in ways that align them with our reflectively held ideals. The model-free systems implicated in some types of associative processing are sensitive to both actual and fictive-error signals, and they are responsive to simulated experiences as well as more abstract forms of social instruction and inferential processing—though precisely how successful these types of interventions are remains an open question. Reading a brief narrative and briefly imagining the life of Martin Luther King Jr. may therefore modulate existing associations by updating the stored value of race-relevant contingencies. Put schematically, prefrontal working memory systems operate in parallel to the midbrain systems that guide model-free learning, and can amplify or dampen reward signals by way of bi-directional circuits linking these areas; imaginative engagements with out-groups can trigger the operation of these systems in ways that have a significant impact on our biases. As Johnson and his colleagues (2013) show, reading a brief narrative about a *strong* Muslim woman's response to being assaulted in a subway station effectively reduces implicit bias, especially in people who are not antecedently disposed to engage in perspective-taking.

In a sense, narrative engagement is a familiar strategy for intervening on implicit attitudes, and it shares much in common with the kind of cognitive and dialectical behavioral therapy that has been designed to re-train our implicit attitudes (cf. Huebner 2009). There is even evidence that we can rely on repeated and directed exposure to the non-stereotypical properties of a stereotype-target to influence automatic judgments (cf., Kawakami et al 2000). But this is not the only way that we can intervene on the internal environment to which our biases are attuned. Implementation intentions have also been shown to have a similar effect, using only a brief rehearsal of an *in-then* action plan that specifies a trigger-cue and an outcome (e.g., *If I see a person, then I will ignore his race!*). Mentally rehearsing an implementation intention thee times yields a form of *reflexive action control*, much like the kind of control evoked by chronic egalitarian values and other shifts in value-driven processing suggested above, and it can significantly decrease shooter bias and modulate the latency of responses in IATs and GNATs (Mendoza, Gollwitzer, Amodio 2010; Webb, Sheeran, & Pepper 2012). Such results initially seem surprising, but I maintain that they are likely to be the result of encoding novel action-plans in working memory, which can be used to up-regulate or down-regulate the salience of existing associations (Miller & Cohen 2001). The operation of mechanisms in the ventromedial PFC, which translate the reward and avoidance values computed by multiple systems into a common currency, can be strongly impacted by the presence of an action plan (Levy & Glimcher 2011; Montague & Berns 2002). Through bi-directional connections between the PFC and the midbrain, the outputs of model-based computations up-regulate and down-regulate the salience of existing associations in light of our goals and values.

This all seems like good news, as such interventions are simple, and they could have a strong impact on the computations that guide biased judgments and biased behavior. But these strategies are limited, and each depends on a relatively local intervention. It would take super-human cognitive resources to moderate and suppress all of our biases in this way, especially where ongoing feedback from the racist world in which we live continually pushes against this system. As long as we are able to focus on egalitarian goals, prevent them from decaying over the course of our everyday experience, and activate implementation intentions when they are called for, such strategies will operate in ways that help to re-shape our habits and reflexive evaluations. But if my model is approximately right, we should also expect there to be other systems pushing back, and since the patterns of associative relations we experience in our world are often difficult to track it will be difficult to moderate and suppress the impact of all of these factors on our decisions and behavior. This, I take it, is the main insight of contemporary Spinozist theories of belief fixation (e.g., Gilbert, Tafarodi, & Malone 1993; Huebner 2009; Mandelbaum submitted). The systems that guide our implicit attitudes are likely to "reflect whatever local environments they are chronically immersed in", and while brief exposures to counter-stereotypical situations or local adjustments to our goals and values produce a brief reduction of bias, we will eventually revert to local norms and prior biases (Dasgupta 2013 271). The effectiveness of these strategies are thus likely to be compromised whenever computational resources are slim, when we are distracted, and when we are experiencing a stressful situation—and all of these situational variables can cause us to backslide into a reliance on Pavlovian and model-free associations that are attuned to the biased structure of our world.

4.2 Niche construction & prefigurative imagination

My account of implicit attitudes suggests that biased stereotypes arise because there is something 'right' about them; but in many cases, the only thing that is 'right' about them is that they pick out statistical regularities in our experience, which are produced and sustained by patterns of systematic, deeply entrenched, institutionalized bias (e.g., there *are* fewer women in STEM fields, Blacks *are* more likely to be arrested, and racialized images pervade the media images we are most likely to encounter). Unfortunately, this is all that backward-looking error-driven learning systems can see, and it is all that they can care about. So as we watch or read the news, watch films, rely on tacit assumptions about what is likely to happen in particular neighborhoods, or draw elicit inferences on the basis of the way in which a person is dressed, we cause ourselves to backslide into our implicit biases. No matter how calm, vigilant, and attentive to our biases we try to be, I maintain that we will be unable to moderate or suppress all of our problematic implicit biases until we eliminate the conditions under which they arise. But with an understanding of the mechanisms that guide reinforcement learning, we can begin to see a way *forward* to developing more robust strategies for intervening on our implicit biases.

To my mind, Nilanjana Dasgupta (2013, 247) has made the most important *ethical* contribution to the rapidly growing literature on implicit bias; as she notes, "environments that facilitate positive contact with members of stereotyped groups create and reinforce positive implicit associations, thereby counteracting implicit bias". Those who live in diverse environments, where members of out-groups are encountered in a diverse range of situations, show lower rates of implicit bias and more egalitarian reflective attitudes (Dasgupta & Rivera 2008). Such environments not only reduce implicit bias, "they also increase people's support for public policies and legislation focused on fixing structural bias and extending equal rights to all groups" (Dasgupta 2013, 247). But the core insight is that we should not simply attempt to eliminate biases, we should attempt to develop more egalitarian attitudes. To do this we must first live in a world that institutes more egalitarian practices. Of course, we do not inhabit such a world, and many people will only experience the members of racialized out-groups through the distorting and one-dimensional lens of the mainstream media, or through encounters that are easily coded as interactions with 'co-workers' or 'friends-of-friends' instead of interactions with 'black people'. Since a person may elicit a positive evaluation when he is categorized in some ways, while still triggering a negative evaluation when he is categorized as an African-American, such encounters are unlikely to impact our behavioral dispositions in their full generality (Mitchel, Nozek, & Banaji 2003). This is why we cannot be content with a world dominated by racist associations—in such worlds bias will often seep into our thoughts and behavior.

Fortunately, we are rapid niche constructors, and we can manipulate *our world* in ways that make it represent new things for us. We have long worked to make the world smart so that we can be dumb in peace (Clark 1998, 80), and the strategies that we have used to do this must be deployed in the context of our morally and socially significant attitudes as well if we want to eliminate implicit and explicit bias. We can cultivate egalitarian commitments; and we can construct relationships with like-minded people who will help us to defend and promote egalitarian attitudes. On this ground we can work to build a world that is not grounded of racist (sexist, heteronormative, or ableist) attitudes or beliefs (cf., Haslanger 2000). To build such a world, we must attempt to reject dominant social norms, challenge existing social institutions, and develop practices that are better than those we have come to expect. But we face a real difficulty in this regard. Error-driven learning helps us to calibrate our behavior against existing norms, and we find norm compliance intrinsically rewarding and norm violation intrinsically aversive. Over the course of our evolutionary history, these facts have had an undeniable benefit (Zawidzki 2013); but our propensity toward norm conformity and over-imitation have also made us conservative organisms that are willing to work to replicate the oppressive forms of social organization that dominate our world. So the type of niche construction that is required to overcome our racial biases calls for another type of imagination, which capitalizes on our ability to transform one another's attitudes, beliefs, and behavioral dispositions.

If we want to *overcome* implicit bias, and if we want to become the sorts of agents who are not dominated by reactions that we cannot reflectively avow, we must engage in collective prefigurative practices designed to create a world where our reflexive reactions are already calibrated against our reflectively held goals and values. We cannot do this on our own, as we always need others to nudge us back toward better practices when we backslide into the racist (sexist, heteronormative, and ableist) thoughts and behaviors that are *statistically normal* in the world that surrounds us. We need to build a new and better world in the shell of the old one, we need to build relationships and forms of social engagement that embody the world we want to live in. This type of practice is forward-looking, and it requires making creative use of the model-based mechanisms that allow us to imagine alternative possibilities; but it also requires trying to live as if an egalitarian world exists before one actually does. These prefigurative practices can become a collective form of *mindshaping* (Mameli 2001; Zawidzki 2013): as we build the world we want to inhabit, we will thereby generate new expectations about how we should live, and these expectations will cause our reflexive learning mechanisms to attune to the work we are trying to bring about. Put differently, I contend that we can only develop egalitarian attitudes by living in an egalitarian world—as paradoxical as it seems, this is the most promising strategy for overcoming implicit bias.

Acknowledgments

The argument in this paper has greatly from a number of conversations with Cassie Herbert, Dan Kelly, Rebecca Kukla, Nabina Liebow, Eric Mandelbaum, and Susanne Sreedhar; I am also grateful for the helpful feedback on earlier drafts of the paper from Michael Brownstein, Ruth Kramer, and Jennifer Saul.

Works cited

Adolphs, R. (2010). What does the amygdala contribute to social cognition?.*Annals of the New York Academy of Sciences*, *1191*, 1, 42-61.

Adolphs, R., Tranel, D., & Damasio, A. (1998). The human amygdala in social judgment. *Nature*, *393*, 6684, 470-474.

Akins, K. (1996). "Of Sensory Systems and the 'Aboutness' of Mental States," *The Journal of*

Philosophy, 93, 337-72.

Balleine, B., Daw, N., O'Doherty, J. (2008), Multiple forms of value learning and the function of dopamine. *Neuroeconomics* (pp.367-387). London: Academic Press.

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

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, *36*, 2, 147-168.

Baron, A., Dunham, Y., Banaji, M. & Carey, S. (in press) Constraints on the Acquisition of Social Category Concepts. *Journal of Cognition and Development.*

Bechara, A (2001). Neurobiology of decision-making. Seminar in Clinical Neuropsychiatry 6: 205–216.

Blair, I., Judd, M., Sadler, M., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. Journal of Personality and Social Psychology, 83, 5–25.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, 713-737.

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference?. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), Attitudes: Insights from the new implicit measures (pp. 285-326). New York: Psychology Press.

Chiu, P., Lohrenz, T., & Montague, P. (2008). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature neuroscience*, *11*, 4, 514-520.

Clark, A. (1998). Being there: Putting brain, body, and world together again. Cambridge: MIT press.

Clark, L,, Bechara, A., Damasio, H., Aitken, M., Sahakian, B., & Robbins, T. (2008). Differential effects of insular andventromedial prefrontal cortex lesions on risky decision-making. Brain, 131, 5, 1311-1322

Cooper, J., Dunne, S., Furey, T., & O'Doherty, J. (2012). Dorsomedial prefrontal cortex mediates rapid evaluations predicting the outcome of romantic interactions. *The Journal of Neuroscience*, *32*, 45, 15647-15656.

Correll, J., G. Urland, & T. Ito (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology* 42,1: 120-128.

Critchley, H., & Rolls, E. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. Journal of Neurophysiology, 75 (1996), pp. 1673–1686

Crockett, M. (2013). Models of morality. Trends in Cognitive Science, 17, 8, 363-6.

Cunningham, W., Johnson, M., Raye, C., Gatenby, J., Gore, J., & Banaji, M. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, *15*(12), 806-813.

Cushman, F. (2013). Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review,* 17 (3), 273-292.

D'Ardenne K, N Eshel, J Luka, A Lenartowicz, L Nystrom, & J Cohen (2012) Role of prefrontal cortex and the midbrain dopamine system in working memory updating. PNAS, 109:19900–19909.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to Situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology*, 47, 233-279.

Dasgupta, N., & Greenwald, A. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of personality and social psychology*, *81*, 5, 800-814.

Dasgupta, N., & Rivera, L. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, *26*, 1, 112-123.

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.

Dayan, P. (2012) How to set the switches on this thing. Current Opinion in Neurobiology, 22, 1068–1074

Dayan P., Niv Y., Seymour B., Daw N. (2006). The misbehavior of value and the discipline of the will. Neural Networks 19: 1153–1160.

De Martino, B., Camerer, C. F., & Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *PNAS*, *107*, 8, 3788-3792.

Dennett, D. (1991) Consciousness explained. Boston: Little Brown Books.

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

Diesendruck, G., & Eldror, E. (2011). What children infer from social categories. Cognitive Development, 26, 118-126.

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

Dunham, Y., Chen, E. & Banaji, M. (in press). Two Signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. Psychological Science.

Evans, J. & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8, 263-271.

Ferguson M., & Bargh J. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. Journal of Personality and Social Psychology, 87, 557–572.

Frank M. & Claus, E. (2006). Anatomy of a decision. Psychological Review, 113, 300–326.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. Psychonomic Science, 4, 123-124.

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.

Gilbert, D., & Malone, P. (1995). The correspondence bias. *Psychologica. bulletin*, *117*, 1, 21.

Gilbert, D., Tafarodi, R. & Malone, P. (1993). You can't not believe everything that you read. Journal of personality and social psychology 65, 221-233.

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*, 1, 164-172.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences, 108 (Supplement 3), 15647-15654.

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

Greenwald, A., Poehlman, T., Uhlmann, E., & Banaji, M. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology, 97, 17–41.

Gregg, A., Seibt, B., & Banaji, M. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. Journal of Personality and Social Psychology, 90, 1-20.

Haslanger, S. (2000). Gender and race: (what) are they? (What) do we want them to be?. *Nous*, *34*, 1, 31-55.

Heyes, C. (2012) Simple minds: A qualified defence of associative learning. Philosophical Transactions of the Royal Society B, 367, 2695-2703.

Huang, J., & Bargh, J. (2011). The selfish goal: Self-deception occurs naturally from autonomous goal operation. Brain and Behavioral Sciences.

Huebner, B. (2009). Troubles with stereotypes for Spinozan minds. Philosophy of the social sciences, 39, 1, 63-92.

Huebner, B. (2013). Do emotions play a constitutive role in moral cognition? *Topoi.*

Huebner, B. (2014). Macrocognition. New York: Oxford University Press.

Huebner, B & R Rupert (2014). Massively representational minds are not always driven by goals, conscious or otherwise. Behavioral and Brain Science, 37 (02), 145-146.

Huys, Q., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*, 3, e1002410.

Johnson, D., Jasper, D. Griffin, S., & Huffman, B. (2013). Reading narrative fiction Reduces Arab-Muslim prejudice and offers a Safe haven from intergroup anxiety. Social Cognition, 31, 5, 578–598

Kawakami, K., Dovidio, J. Moll, J., Hermsen, S. and Russin, A. (2000). Just say no (to stereotyping). Journal of personality and social psychology 78, 5, 871-88.

Kim, M., Loucks, R., Palmer, A., Brown, A., Solomon, K., Marchante, A., & Whalen, P. (2011). The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. *Behavioural brain research*, *223*, 2, 403-410.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. Neuron 61, 140–151.

Klucharev, V., Munneke, M., Smidts, A., and Fernández, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. Journal of Neuroscience, 31, 11934–11940.

Knoch, D, M Nitsche, U Fischbacher, C Eisenegger, A Pascual-Leone & E Fehr (2008). Studying the Neurobiology of Social Interaction with Transcranial Direct Current Stimulation. Cerebral Cortex,18 (9):1987-1990.

Koenigs M., & Tranel, D. (2007) Irrational economic decision-making after ventromedial prefrontal damage. Journal of Neuroscience, 27: 951-956.

Kubota, J., Banaji, M., & Phelps, E. (2012). The neuroscience of race. *Nature neuroscience*, *15*, 7, 940-948.

Levy D & P Glimcher (2011) Comparing apples and oranges. Journal of Neuroscience, 31:14693–14707.

Lohrenz, T., McCabe, K., Camerer, C., and Montague, P. (2007). Neural signature of fictive learning signals in a sequential investment task. PNAS 104, 9493–9498.

Machery, E. (This volume). De-Freuding Implicit Attitudes.

Mahajan, N., Martinez, M., Gutierrez, N., Diesendruck, G., Banaji, M., & Santos, L. (2011). The evolution of intergroup bias: Perceptions and attitudes in rhesus macaques. *Journal of Personality and Social Psychology*, *100*, 387-405

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy*, *16*, 5, 595-626.

Mandelbaum, E. (in prep). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias.

Mandelbaum, E. (submitted). Thinking is believing.

Mendoza, S., Gollwitzer, P., & Amodio, D. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin*, *36*(4), 512-523.

Miller E. & J. Cohen (2001) An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24:167–202.

Milne, E. & Grafman, J. (2001). Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. *Journal of Neuroscience,* 21,12, 1-6.

Mitchell, J., Macrae, C., & Banaji, M. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*, 4, 655-663.

Mitchell, J., Nosek, B., Banaji, M. (2003). Contextual variations in implicit evaluation. Journal of experimental psychology—General, 132, 455-469.

Moll J, F Krueger, R Zahn, M Pardini, R de Oliveira-Souza & J Grafman (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. PNAS 103:15623-15628

Montague P. & G. Berns (2002). Neural economics and the biological substrates of valuation. Neuron **36**: 265–284.

Montague, P., Hyman, S., & Cohen, J. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431,* 7010, 760-767.

Montague P., Dayan P., Sejnowski T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 16, 1936–1947.

Moskowitz, G., Gollwitzer, P., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. Journal of Personality and Social Psychology, 77, 1, 167-184.

Niv, Y. (2009) Reinforcement learning in the brain. The Journal of Mathematical Psychology 53. 3, 139-154.

Nosek, B., & Banaji, M. (2001). The go/no-go association task. Social Cognition 19 (6): 625–664.

Preuschoff, K., Bossaerts, P., and Quartz, S. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51, 381–390.

Preuschoff, K., Quartz, S., and Bossaerts, P. (2008). Human insula reflects risk predictions errors as well as risk. Journal of Neuroscience 28, 2745–2752.

Prinz, J. (2002). Furnishing the mind: concepts and their conceptual basis. Cambridge: MIT Press

Quadflieg, S., & Macrae, C. (2011). Stereotypes and stereotyping: What's the brain got to do with it? *European Review of Social Psychology, 22*, 215-273

Quartz, S. (2009). Reason, emotion, and decision-making. Trends in Cognitive Sciencem 13, 209–215.

Railton, P. (2014). The affective dog and its rational tale. *Ethics*, 124, (4), 813-859.

Rescorla, R. (1988). Pavlovian conditioning: It's not what you think it is. American Psychologist, 43, 151-160.

Rupert, R. (2011). Embodiment, Consciousness, and the Massively Representational Mind. *Philosophical Topics*, 39 (1): 99-120.

Saver, J & A Damasio (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. Neuropsychologia, 2912.

Schultz,W.(1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1–27.

Schultz, W. (2010). Dopamine signals for reward value and risk. Behavior and brain Function, 6, 24.

Schultz, W., Apicella, P., & Ljungberg, T. (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience, 13, 900–913.

Schultz, W., Dayan, P., & Montague, P. (1997). A Neural Substrate of Prediction and Reward. Science, 275, 1593-1599.

Schwabe, L., & Wolf, O. (2013). Stress and multiple memory systems: from 'thinking'to 'doing'. *Trends in cognitive sciences*.

Sechrist, G. & Stengor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. Journal of Personality and Social Psychology 80, 4. 645-54.

Shenhav, A & J Greene (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron, 67, 667-677.

Sherman S., Rose J., Koch K., Presson C., Chassin L. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. Journal of Social and Clinical Psychology, 22, 13–39.

Spinoza, B. (2002). Complete works. S. Shirley (trans). London: Hackett.

Stanley, D., Phelps, E., & Banaji, M. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science*, *17*(2), 164-170.

Sutton, R. & Barto, A. (1998). Reinforcement Learning: An Introduction. Cambridge: MIT Press.

Webb, T., Sheeran, P., & Pepper, J. (2012). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. British

Journal of Social Psychology, 51, 13-32.

Wegner, D. (1994). "Ironic processes of mental control." Psychological review, 101: 34-52.

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

Zawidzki, T. (2013). Mindshaping: A New Framework for Understanding Human Social Cognition. Cambridge: MIT Press.