Michael Brownstein
Chapter for the *Routledge Handbook of the Social Mind*

Implicit Attitudes, Social Learning, and Moral Credibility[1]

1. Spontaneity and credibility

Dichotomous frameworks for understanding human decision-making often distinguish between spontaneous or intuitive judgments, on the one hand, and deliberative, reasoned judgments, on the other. The precise qualities thought to characterize these two kinds of judgments—sometimes aggregated under the headings "System I" and "System II"—change from theory to theory.[2] Recent years, however, have seen a shift in dual systems theorizing from attempts to specify the precise qualities that characterize these two kinds of judgments to descriptions of the distinct neural and computational mechanisms that underlie them. In turn, these mechanisms are coming to be a focal point for the current incarnation of a long-standing debate about whether and why spontaneous judgments are ever good guides for decision-making and action.[3] Do our intuitions, emotional reactions, and unreasoned judgments ever have authority for us? Are they morally credible? On the one hand, one might think that the nature of the neural and computational systems underlying spontaneous judgments demonstrates that they are paradigmatically short-sighted and morally untrustworthy. On the other hand, the nature of these mechanisms might vindicate at least a defeasible authority accorded to our spontaneous judgments in some circumstances.

This debate can be articulated in terms of the mechanisms of evaluative learning. What kinds of processes are involved in forming our spontaneous judgments? Are those processes responsive to past experience, situationally flexible, etc. in ways that make them good guides for decision-making and action? In what follows, first, I'll describe a general argument others have made for the defeasible moral credibility of spontaneous judgments, based on the neural and computational mechanisms that underlie them (§2).[4] I'll then focus on the particular case of implicit

[1] Many thanks to Jason D'Cruz, Bryce Huebner, Julian Kiverstein, Victor Kumar, members of the Minorities and Philosophy chapter at SUNY-Albany, and members of the Manhattan College philosophy department for invaluable feedback on this chapter. I am also grateful to the Leverhulme Trust for funding the Implicit Bias and Philosophy workshops at the University of Sheffield, from which some of the ideas in this chapter sprung.

[2] See Stanovich (1999) and Stanovich & West (2000) for review.

[3] See Railton (2009, 2014, 2015); Brownstein & Madva (2012a,b); Paxton et al. (2012); Greene (2013); Seligman et al. (2013); Kumar & Campbell (forthcoming); Kumar (ms); Martin & Cushman (ms).

[4] My concern is epistemic, not moral as such. I take it for granted that some spontaneous judgments are morally good and others are morally bad, and that much of the time, our normative theories will agree about which are which. Notwithstanding particular salient (and important) cases of moral disagreement, in other words, we tend to agree that judgments that promote happiness, pro-sociality, cooperativeness, and so on are morally good, and that judgments that promote suffering, anti-sociality, selfishness, and so on are morally bad. My concern is epistemic in the sense that it is about knowing which of these moral ends our spontaneous judgments are going to promote when we are in the flow of

social attitudes (§3). I do so for two reasons. First, these attitudes can be understood as subserved in large part by the same neural and computational learning mechanisms that I describe in §2. This shows that implicit social attitudes count as an instance of spontaneous judgments in the relevant sense, and can in principle be good guides for decision-making and action. Second, the cases in which implicit social attitudes are good guides for action are extremely hard to distinguish on the ground, so to speak, from the cases in which they amount to morally deplorable *implicit biases*. This illuminates what I call the "credibility question:" under what conditions are one's spontaneous judgments good guides for action, compared with the conditions under which one ought to override one's immediate inclinations?[5] Peter Railton (2014) identifies certain traits and experiences of people whose spontaneous judgments have putative moral authority. Focusing on implicit social attitudes as a test case, I consider the evidence for his suggestion in the final section of the chapter (§4).

2. Value and attunement

Research in computational neuroscience suggests the existence of two distinct mechanisms for value-based decision-making. These are typically called "model-based" and "model-free" systems (Blair, 2004, 2013; Crockett, 2013; Cushman, 2013).[6] I briefly describe each (§2.1), then consider evidence suggesting that model-free systems can produce morally credible spontaneous judgments (§2.2).

2.1 Model-based and model-free evaluative learning

Model-based evaluative learning systems produce map-like representations of the world. They represent the actions that are available to an agent, along with the potential outcomes of those actions and the values of those outcomes. This comprises what is often called a "causal model" of the agent's world. In evaluating an action, a model-based system runs through this map, calculating and comparing the values of the outcomes of different choices, based on the agent's past experiences, as well as the agent's abstract knowledge. A simple example imagines a person navigating through a city, computing and comparing the outcomes of taking one route vs. another, and then choosing the fastest route to their destination.[7] The agent's internal map of the city comprises a causal model of the agent's action-relevant "world." This model can also be thought of as a decision-tree. Running through the "branches" of a decision-tree is what many commonly refer

---

action and decision-making. This is to say that my concern is about practical reason in precisely those cases of judgment and behavior when explicit reason*ing* does not or cannot happen.

[5] In asking this question, I take up the worry left unresolved in Brownstein and Madva (2012a). Note also that I use "credible" in the broad sense of being trusted, not the narrower epistemic sense of being believable.

[6] See Kishida et al. (2015) and Pezzulo et al. (2015) for evidence of information flow between model-free and model-based evaluative learning systems. I discuss some of the upshots of integration between these learning systems in §3.1.

[7] I am indebted to Fiery Cushman for this example.

to as "reasoning" (Cushman, 2013). That is, model-based systems are thought to subserve the process in which agents consider the outcomes of various possible actions, and compare those outcomes, in light of what the agent cares about or desires. This sort of reasoning is inherently prospective, since it requires projecting into the likely outcomes of hypothetical actions. For this reason, model-based systems are sometimes referred to as "forward-models."

In contrast, a model-free system computes the value of one particular action, based on the agent's past experience in similar situations. Model-free systems enable value-based decision-making without representing complex map-like causal links between actions and outcomes. Essentially, model-free systems compute the value of one particular action, without modeling the "world" as such.[8] The computation is done on the basis of a comparison between the agent's past experience in similar situations and whether the current action turns out better or worse than expected. For example, faced with a decision of turning left or right at a familiar corner, a model-free system can generate a prediction that turning one direction (e.g., left) will be valuable to the agent. The system does this based on calculations of the size of any discrepancy between how valuable turning left was in the past and whether turning left this time turns out better than expected, worse than expected, or as expected. Suppose that in the past, when the agent turned left, the traffic was better than she expected. The agent's "prior" in this case for turning left would be high. But suppose this time, the agent turns left, and the traffic is worse than expected. This generates a discrepancy between the agent's past reinforcement history and her current experience, which is negative given that turning left turned out worse than expected. This discrepancy will feed into her future predictions; her prediction about the value of turning left at this corner will now be lower than it previously was. Model-free systems rely on this "prediction-error" signaling, basing new predictions on comparisons between past reinforcement history and the agent's current actions. While there is no obvious folk psychological analogue for model-free processing, the outputs of this system are commonly described as gut feelings, spontaneous inclinations, and the like. This is because these kinds of judgments do not involve reasoning about alternate possibilities, but rather, they offer agents an immediate positive or negative sense about what to do.

Model-based systems are often described as flexible, but computationally costly, while model-free systems are described as inflexible, but computationally cheap. Navigating with a map enables flexible decision-making, in the sense that one can shift strategies, envision sequences of choices several steps ahead, utilize "tower of Hanoi" like tactics of taking one step back in order to take two more forward, etc. But this sort of navigation is costly in the sense that it involves computing many action-outcome pairs, the number of which expand algorithmically even in seemingly simple situations. On the other hand, navigating without a map, based on the information provided by past experiences for each particular decision, is comparatively inflexible. Model-free systems only enable one to evaluate one's current action, without considering options in light of alternatives or future consequences. But navigating without a map is easy and cheap. The number of options to compute are severely constrained, such that one can make on-the-fly

[8] But see Kishida et al. (2015) for evidence of fluctuations in dopamine concentration in the striatum in response to both actual and counterfactual information.

decisions, informed by past experience, without having to consult the map (and without risking tripping over one's feet while one tries to read the map, so to speak).

If one accepts the rough generalization that the outputs of model-based processing are deliberative judgments and the outputs of model-free processing are spontaneous judgments,[9] then one might also think that deliberative judgments are flexible but inefficient and spontaneous judgments are inflexible but efficient. Indeed, this is what many people think, whether they are focused on the more general level of System I and System II or whether they are focused on the specific model-free and model-based learning systems that appear to subserve System I and II.[10] But there is reason to question the putative inflexibility of spontaneous judgments and to do so based on what model-free systems can do. Moreover, it is not just situationally flexible behavior that model-free learning can support, but socially attuned, experience-tested behavior and decision-making as well.

2.2 Wide competence and model-free learning

Consider three cases in which agents' spontaneous reactions tend to outperform their deliberative judgments. In the "Iowa Gambling Task" (Bechara et al., 1997), participants are presented with four decks of cards and $2000 in pretend gambling money. They must choose facedown cards, one at a time, from any of the decks. Two of the decks are "good" in the sense that choosing from them offers an overall pattern of reward, despite only small rewards offered by the cards at the top of the deck. Two of the decks are "bad" in the sense that picking from them gives the participant a net loss, despite large initial gains. It takes subjects on average about 80 card-turns before they can say why they prefer to pick from the good decks. After about 50 turns, most participants can say that they prefer the good decks, even if they aren't sure why. But even before this, during what Bechara and colleagues call the "pre-hunch" phase, most participants prefer the good decks (as revealed by their actual choices). But when stopped and asked about their preferences and beliefs about the decks every 10 turns, participants don't report having any preferences or strategic beliefs during this phase. Most intriguingly, after about only 20 turns, most participants have higher anticipatory skin conductance responses before picking from the bad decks.[11]

It's not just in tracking statistical regularities that spontaneous judgments have the potential to outperform deliberative judgments. A second set of examples stem from cases of expert action, particularly in sports. Oftentimes, expert athletes do *not* have greater conscious or declarative access to the reasons for which they make the choices they do, such as playing a particular shot or swinging

[9] But see §3.1 for discussion.
[10] For instance, see Greene (2013).
[11] Maia and McClelland (2004) argue that participants in the Iowa Gambling Task may, in fact, have conscious knowledge of the most advantageous strategies as soon as they behave according to these strategies. See also Dunn et al. (2005) for critique of Bechara et al. (1997). See Bechara et al. (2005) for response, who show, for example, that anticipatory skin conductance response occurs before participants have conscious knowledge of advantageous strategies.

at a particular pitch (Beilock, 2010; Brownstein, 2014; Michaelson & Brownstein, 2015). This is perhaps why the best athletes don't necessarily make good coaches; while experts' performances can be extraordinary, their understanding of what distinguishes their abilities is often just ordinary. Instead, expert athletes appear to have a special ability to make nearly instantaneous action-guiding predictions about the relevant variables in the sport (Yarrow et al, 2009). In ball sports, for example, like baseball, experts' motor performance (e.g., hitting) is tied to their ability to accurately predict when and where the ball will cross the plate.[12] Baseball expertise, it seems, is not just determined by greater physical strength, a more determined will, and better coordination, but also by the ability to make spontaneous and accurate on the fly predictions about the outcomes of valued events under ambiguous conditions.[13]

A final set of examples have to do with interpersonal social fluency, which is commonly recognized as "people skills" or "tact." Interpersonal social fluency requires one's spontaneous gestures and "micro-expressions" to be attuned to others and to the general demands of the situation.[14] Consider, for example, then President-Elect Obama's inauguration in 2009. In front of millions of viewers, Obama and Chief Justice of the Supreme Court John Roberts both fumbled the lines of the Oath, opening the possibility for a disastrously awkward moment.[15] But after hesitating for a moment, Obama smiled widely and nodded slightly to Roberts, as if to say, "it's okay, go on." These gestures received little explicit attention, but they defused the awkwardness of the moment, enabling the ceremony to go on in a positive atmosphere. Despite his nervousness and mistakes, Obama's social fluency was on display. Most of us know people with similar skills, which require real-time, fluid spontaneity, and can lead to socially valuable ends (Manzini et al., 2009).[16]

There is reason to think that model-free learning mechanisms can do substantial work in explaining agents' decisions and behavior in cases like these—in estimating statistics, in expert athletics, and in interpersonal interaction. In some cases, there is direct evidence. Statistical competence, for example, has been traced to model-free learning mechanisms (Daw et al., 2011).[17] These findings are consistent with wide-ranging research suggesting that agents'—even nonhuman

---

[12] This is shown using two related experimental scenarios. In a temporal occlusion scenario, athletes are shown the first part of a scenario—for example, a pitcher winding up and releasing the pitch—but then the action is cut off (Müller et al., 2006). In a spatial occlusion scenario, athletes' vision is obscured (Müller & Abernathy, 2006).

[13] I am not advancing the strong claim that model-free learning mechanisms are sufficient for explaining how agents make good spontaneous judgments, as in the case of the Iowa Gambling task, or perform skilled spontaneous behavior, in cases like batting in baseball. My claim is that model-free mechanisms are surprisingly explanatory. Moreover, in some cases, model-free learning can help to explain what distinguishes experts from novices. It is likely that competent but not expert baseball players have complex model-based representations of potential outcomes of potential actions, perhaps even as complex as experts' model-based representations. What seems to distinguish experts, however, is the quality of their model-free representations of the value of particular actions.

[14] The following example was originally discussed in Brownstein and Madva (2012a). See Madva (2012) for further discussion of the concept of interpersonal fluency.

[15] For a clip of the event, and an "analysis" of the miscues by CNN's Jeanne Moos, see: http://www.youtube.com/watch?v=EyYdZrGLRDs.

[16] Moreover, research suggests that poor social sensitivity is, in some cases, a result of "choking," much as in athletics and high-stakes testing. See Knowles et al. (2015).

[17] However, as Daw and colleagues (2011) find, statistical learning typically involves integration of predictions made by both model-based and model-free systems. See discussion in §3.1 on the integration of multiple learning systems.

agents'—spontaneous judgments are surprisingly competent at tracking regularities in the world (e.g., Kolling, 2012; Preuschoff et al., 2006). Yarrow and colleagues (2009) combine this with research on motor control to understand expertise in athletics. They focus on experts' ability to make predictive, rather than reactive, decisions, on the basis of values generated for particular actions. More research is clearly needed, but this is suggestive that model-free learning is essential to the skilled spontaneous judgment that distinguishes experts in sports from beginners and even skilled amateurs.

It is relatively uncontroversial, however, to say that model-free learning helps to explain spontaneous judgment and behavior in cases in which the relevant variables are repeatedly presented to the agent in a relatively stable and familiar environment. In cases like batting in baseball, this repeated presentation of outcomes in familiar situations enables the agent to update her predictions on the basis of discrepancies between previous predictions and current actions. But what about cases in which an agent spontaneously displays appropriate, and even skilled, behavior in unfamiliar environments? Interpersonal social fluency requires this. Offering a comforting smile can go terribly awry in the wrong circumstance; interpersonal *fluency* requires deploying the right reaction in changing and novel circumstances. The question then is whether model-free systems can explain a kind of "wide" competence in spontaneous decision-making and behavior.[18] Wide competencies are not limited to a particular familiar domain of action. Rather, they can manifest across a diverse set of relatively unfamiliar environments. One reason to think that model-free systems *can* subserve wide, rather than narrow (i.e., context-bound) abilities is that these systems can treat novel cues which are not rewarding as predictive of other cues which *are* rewarding. Huebner (2016) describes this process:

> 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 . . ..[19]

This kind of "scaffolding" of reward prediction is known as "temporal difference reinforcement learning" (TDRL; Sutton, 1988; Cushman, 2013). It enables model-free systems to treat cues in the environment which are themselves not rewarding, but are predictive of rewards, as intrinsically rewarding. The chocolate shop is not itself rewarding, but is predictive of other outcomes (eating chocolate) that are rewarding. (Better: the chocolate shop is predictive of buying chocolate which is predictive of eating chocolate which is predictive of reward.) The key point is that the chocolate

[18] So far as I know, Railton (2014) first discussed wide competence in spontaneous decision-making in this sense.
[19] See also Cushman (2013, 280).

shop itself comes to be treated as rewarding. The agent need not rely on a causal map that abstractly represents A leading to B, B leading to C, and C leading to D.

This helps to explain how a spontaneous and socially attuned gesture like Obama's grin can be generated by a model-free learning system. Smiling-at-Chief-Justices-during-Presidential-Inaugurations is not itself rewarding. Or, in any case, Obama did not have past experiences that would have reinforced the value of this particular action in this particular context. But presumably Obama did have many experiences that contributed to the fine-tuning of his micro-expressions, such that these spontaneous gestures have come to be highly adaptive. Indeed, this adaptiveness is unusually salient in Obama's case, where he displayed a high degree of interpersonal social fluency. And yet he might have very little abstract knowledge *that* he should smile in situations like this one (Brownstein & Madva, 2012). As Cushman (2013) puts it, a model-free algorithm knows that some choice feels good, but it has no idea why.

The upshot is that the outputs of model-free learning ought to be accorded some kind of defeasible authority for us. That is, the wide competence and experience tested qualities of model-free learning suggest that there are times when we ought to trust our spontaneous judgments. Seligman and colleagues (2013) count four related reasons for thinking that the learning system subserving paradigmatic spontaneous judgments should be accorded some practical authority in decision-making. First, these systems enable agents to learn from experience, given some prior expectation or bias. Second, they enable prior expectations to be overcome by experience over time, through the "washing out" of priors. Third, they are set up such that expected values will, in principle, converge on the "real" frequencies found in the environment, so that agents really do come to be attuned to the world. And fourth, they adapt to variance when frequencies found in the environment change, enabling relatively successful decision-making in both familiar and relatively novel contexts. Together, these features of model-free learning underlie what I mean by the defeasible moral credibility of our spontaneous judgments.

Of course, our spontaneous judgments are only *defeasibly* good guides for action, and for several reasons. As Huebner (2009, 2016) emphasizes, these learning systems will only be as good as the environment in which they are trained. An agent whose past experiences are morally blinkered, perhaps due to being raised in an isolated and xenophobic environment, is likely to have morally blinkered spontaneous judgments. Likewise, people who live in an unjust world (like us), suffused with prejudice and negative stereotypes, are likely to become attuned to common prejudicial attitudes and to reflect social stereotypes in their reward predictions. In cases like these, agents' spontaneous social judgments may still be thought of as attuned to the social world, but just to the wrong features of it. Worries like these give rise to the idea that a decision-making system that ought to hold authority for us must do more than just represent first-order values. One might think, for example, that a morally credible action-guidance system must be responsive to values that an agent reflectively endorses, and not just to predictors of good feelings. The fact that our spontaneous judgments *lack* features like these gives rise to the credibility question, that is, the

question for practical agents of knowing when the defeasible authority of their spontaneous judgments has indeed been defeated.[20]

3. Implicit Attitudes

Proponents of the view that model-free learning systems can have moral credibility have developed this claim using examples such as interpersonal social fluency (Railton, 2014), judgments involving moral luck (Kumar, ms; Martin & Cushman, forthcoming), and even the "Statistical Victim Effect" (Railton, 2015). Here I consider the claim in light of a distinct but related set of phenomena. Research on "implicit attitudes" has grown rapidly over the past 25 years, and there is good reason to believe that model-free learning can explain substantial features of how these states function (§3.1). If this is right, then implicit attitudes should be defeasibly credible guides to action. They should act as both valuable social tuning devices, that is, but also be highly susceptible to bias (§3.2). Given this, how can we tell when to trust them? I raise some worries about the difficulty of answering this question (§3.3), then consider one kind of solution (§4).

3.1 Implicit attitudes and model-free learning

People hold implicit attitudes toward food, clothing, brands, alcohol, and, most notably, social groups. Implicit attitudes are generally understood as preferences that need not enter into focal awareness and are relatively difficult to control. Virtually all theoretical models of implicit attitudes understand them to be the product of a complex mix of cognitive and affective processes.[21] Elsewhere I have offered an account of how these cognitive and affective processes cohere into a particular kind of (implicit) mental state (Madva & Brownstein, ms). Here I consider what kind of learning mechanisms might subserve these states. I suggest that model-free evaluative learning systems explain more about implicit attitudes that others might suppose. This makes possible the idea that implicit attitudes are defeasibly good guides to action. Of course, amassing sufficient evidence for this claim would require a paper of its own. My aim instead is to sketch a conceptual architecture on the basis of which this is plausible and then to consider the credibility question about implicit attitudes.

My view is much indebted to Huebner (2016), who argues that implicit attitudes are constructed by the aggregate "votes" cast by basic "Pavlovian" stimulus-reward associations, model-free reward predictors, and model-based decision-trees. Pavlovian stimulus-reward associations are distinguished from model-free reward predictors in that the former passively bind innate responses

[20] Philosophers enamored of the concept of "reasons-responsiveness" are invited to understand me as saying that model-free learning helps to explain why our spontaneous judgments often seem to be responsive to reasons, but can nevertheless run afoul of our overriding moral reasoning.

[21] See, for instance, Fazio (1990); Gawronski & Bodenhausen (2006); Amodio & Ratner (2011).

to biologically salient rewards, whereas the latter compute decisions based on the likelihood of outcomes and incorporate predictors of predictors of rewards, as described in §2. This process of decision-system voting is substantiated by previous research (Crockett 2013; Daw et al 2011; Huys et al 2012). In short, on Huebner's view, Pavlovian associative mechanisms track cues in the environment that are biologically salient, such as signs of danger or sexual reward. In a world such as ours, which is suffused with images and stories tying particular social groups to signs of danger, sex, etc., these basic associative mechanisms will attune agents to these racialized and sexualized representations. These socially saturated stimulus-response reactions aren't tantamount to implicit attitudes just as such, however. This is because implicit attitudes aren't only responsive to threats and biological rewards, but also to social norms. Tracking and updating according to the demands of social norms is the work of model-free systems, which can treat social norms as stand-ins for expected rewards. Huebner here draws on research showing that model-free prediction-error signaling is largely responsible for norm-conformity (e.g., Klucharev et al 2009). Finally, Huebner draws on evidence showing that, in some cases, implicit attitudes are responsive to things like inferential processing and argument strength (Mandelbaum, 2015). This responsiveness relies upon model-based representations of alternate possibilities, competing goals, and abstract values. Implicit attitudes, then, reflect the potentially conflicting pull of these three decision-making systems. Huebner summarizes:

> 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 . . .

This account is compelling, particularly for its ability to accommodate a large range of otherwise seemingly conflicting data. There are two different ways to interpret Huebner's view. One is that it represents a computational explanation of implicit attitudes as such. On this interpretation, Pavlovian, model-free, and model-based learning mechanisms cast "votes," the aggregated outcome of which represents the content of an agent's particular implicit attitude. Implicit attitudes are the *product* of the competition between these three evaluative systems, in other words. A second interpretation is that these three learning systems cast votes, the aggregated outcome of which determines an agent's *behavior*. On this second interpretation, implicit attitudes represent a *component* of the competition between these three systems. It is possible on this second interpretation that the agent's implicit attitude is exclusively or primarily the product of one kind of learning system, the output of which then competes with the outputs of the agent's other learning systems. Huebner seems to have both of these interpretations in mind. His stated aim is to provide a computational account of implicit biases (51), but he also suggests that both our implicit *and* our explicit attitudes represent the combined influences of these three types of evaluative systems (58), and that our implicit attitudes are themselves regulated *by* model-based evaluations (64).

My investigation into the defeasible credibility of model-free learning systems might seem wrongheaded if implicit attitudes as such are the product of competition between Pavlovian, model-free, and model-based evaluative learning systems. For why focus on model-free learning alone if implicit attitudes are the product of multiple overlapping systems? However, if behavior is the result of the competition between these systems, and implicit attitudes represent the output of one component of this competition, then it might not be so wrongheaded to think that we can learn about the potential moral credibility of implicit attitudes by considering model-free learning in particular. On this interpretation, on which behavior is the result of the competition between learning systems, implicit attitudes may be thought of as the output of model-free learning mechanisms in particular (though perhaps not exclusively). That is, behavior is the result of the combined influence of our reflexive reactions to biologically salient stimuli (which are paradigmatically subserved by Pavlovian mechanisms), our implicit attitudes (paradigmatically subserved by model-free mechanisms), and our explicit attitudes (paradigmatically subserved by model-based mechanisms). Now, this picture is surely too simplistic. As Huebner rightly emphasizes, these processes mutually influence each other. For example, one's implicit attitudes are likely to be affected one's "desire to produce and sustain egalitarian values," a desire which Huebner suggests is the product of model-based mechanisms.[22] But to accept this mutual penetration of cognitive and affective processes is not tantamount to the view that these systems mutually *constitute* one's implicit attitudes. It is one thing to say that implicit attitudes are mental states paradigmatically produced by model-free evaluative learning systems, which are in important ways influenced by other learning systems. It is another thing to say that implicit attitudes are mental states produced by the competition between these learning systems themselves.

Huebner can in fact embrace both of these interpretations—that implicit attitudes are the product of these three evaluative learning systems and also that implicit attitudes are a component of the competition between these three systems—because he holds a dispositional view of attitudes. On this dispositional view, attitudes (in the psychological sense of likings and dislikings) are not mental states that can occur; rather, they are multi-track dispositions to behave in particular ways across varied situations.[23] Since the dispositional view denies that implicit attitudes are a unified cognitive state, and are better understood as stable dispositional traits, then there is no problem in saying that implicit attitudes are both a product and a component of the competition between evaluative learning systems. There is no problem, in other words, in saying that the competition between learning mechanisms issues in both attitudes and behavior, on the dispositional view, since this view denies that there is a meaningful distinction between attitudes and behavior.

As mentioned above, I have argued elsewhere for a particular conception of implicit attitudes as a relatively unified kind of mental state. I won't focus here on the debate between mental state and dispositional views of attitudes. Rather, I consider the right way to think of the

[22] For example, one's motivation to act in egalitarian ways "for its own sake" (rather than to appear unprejudiced in the eyes of others) strongly moderates the effects of the implicit attitude on one's behavior and judgment (Plant & Devine, 1998). See also Glaser and Knowles (2008).

[23] Huebner (2016) endorses Machery's (2016) dispositional account of implicit attitudes.

competition between learning systems *given* a mental state view of implicit attitudes. In short, it is hard to understand how a competition between learning mechanisms could issue in both attitudes and behavior on a mental state view. Rather, it is more parsimonious, on the assumption that implicit attitudes are mental states, to think that what the competition between learning systems helps to explain is how particular decisions and behavior are produced by the interaction of biologically attuned reflexes, implicit attitudes, and explicit attitudes. What then remains open is how best to understand biologically attuned reflexes, implicit attitudes, and explicit attitudes in terms of the learning mechanisms that subserve them. My view is that model-free learning explains more about implicit attitudes than Pavlovian or model-free mechanisms do. Architecturally, a reasonable, albeit loose, way of thinking (as described above) is that biologically attuned reflexes are the paradigmatic causal outputs of Pavlovian mechanisms, implicit attitudes are the paradigmatic outputs of model-free learning mechanism, and explicit attitudes are the paradigmatic outputs of model-based learning mechanisms. Implicit attitudes are certainly affected by biologically salient stimuli—for example, those that elicit aversive fear—as well as by an agent's explicit values, but the attitude itself is the association between two attitude objects. Of course, much more would need to be said to substantiate this. My aim, though, is to establish sufficient centrality of model-free systems in understanding implicit attitudes in order to show that implicit attitudes have a defeasible moral credibility, due to the learning mechanisms that subserve them. I now turn to give an example of how the defeasible credibility of implicit attitudes works in practice. In some situations, one and the same set of implicit attitudes seem to be both authoritative and morally disastrous. This leads to what I call the credibility question.

3.2 The gift and the curse of fear

Victims of violent assault often say, after the fact, that "something just felt wrong" about the person walking on the other side of the street or offering to help carry the groceries into their apartment. But to their great regret, they dismissed these feelings, thinking that they were just being paranoid or suspicious. In *The Gift of Fear*, Gavin de Becker argues that the most important thing a person can do to avoid becoming a victim of violent assault is to trust their intuition when something about a person or situation seems amiss. He writes,

> A woman is waiting for an elevator, and when the doors open she sees a man inside who causes her apprehension. Since she is not usually afraid, it may be the late hour, his size, the way he looks at her, the rate of attacks in the neighborhood, an article she read a year ago—it doesn't matter why. The point is, she gets a feeling of fear. How does she respond to nature's strongest survival signal? She suppresses it, telling herself: 'I'm not going to live like that; I'm not going to insult this guy by letting the door close in his face.' When the fear doesn't go away, she tells herself not to be so silly, and she gets into the elevator. Now, which is sillier: waiting a moment for the next elevator, or getting into a soundproofed steel chamber with a stranger she is afraid of? (1998, 30-31)

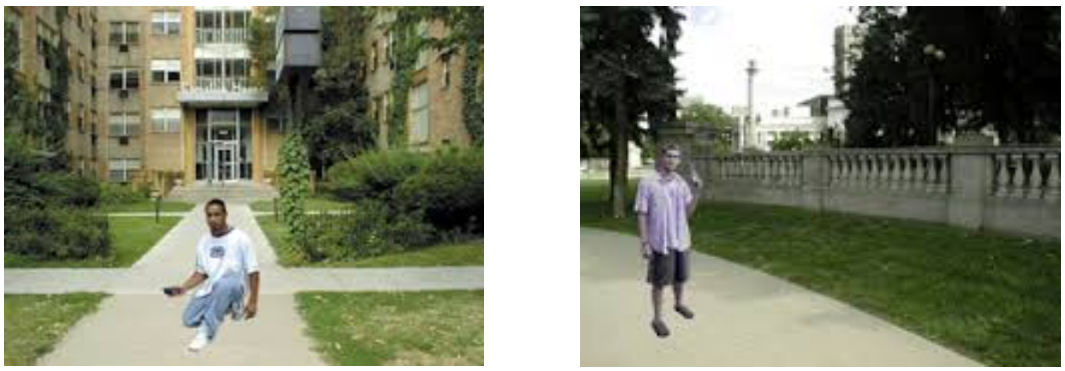
De Becker offers trainings promising to teach people how to notice their own often very subtle feelings of fear and unease—their "Pre-Incident Indicators"—in potentially dangerous situations. These indicators, he argues, are responsive to nonverbal signals of what other people are thinking or planning. For example, we may feel unease when another's "micro-expressions," like a quick sideways glance, or rapid eye-blinking, or slightly downturned lips, signal their intentions, even though we might not notice these cues consciously. De Becker's trainings have been adapted for police officers too, who also often say, after violent encounters, that they could tell that something was wrong in a situation, but they overrode those feelings because they didn't seem justified at the time.

De Becker's Pre-Incident Indicators are good candidates for valuable social tuning devices that are produced by implicit attitudes.[24] They typically emerge into an agent's peripheral, rather than focal awareness, which is a hallmark of implicit attitudes (Gawronski et al., 2006; Brownstein & Madva, 2012b). They are also relatively automatic, in the same way in which outcomes of measures of implicit attitudes, like the Implicit Association Test (IAT; Greenwald et al., 1998), are relatively automatic. This is evident in the way in which de Becker describes one's intuitions as often in conflict with one's reflective judgments (as in the case in which a person feels that something is amiss but can't find any reason to justify the feeling). Finally, these Pre-Incident Indicators seem likely to be generated by model-free learning systems. Agents who are enculturated in typical ways make and refine predictions about subtle social signaling, such as posture and eye gaze, and presumably update these predictions on the basis of discrepancies between those predictions and outcomes.

Assuming this is right, and that de Becker's approach is indeed a valuable tool for protecting oneself, then we seem to have reason to treat our pre-incident indicators as good guides for decision-making.[25] But there is a problem—perhaps an obvious one—with de Becker's advice. Consider, for example, research on "shooter bias." In a computer simulation, participants are quickly shown images like these and must try to shoot all and only those people shown holding guns:

[24] In discussing model-free learning, Railton makes a similar claim, writing that the "core" of "spontaneous yet apt responsiveness to reasons for belief and action has at its core the operation of implicit affective processes" (2014, 847). This is a telling remark, suggesting that the model-free learning structures I've been discussion are akin to the processes that generate what social psychologists call implicit attitudes. But in the same paper, Railton also speaks of implicit knowledge and understanding, implicit social and cultural competence, implicit models of social situations, and implicit attentional and motivational processes. This suggests that his use of the term "implicit" is not specifically meant to refer to what social psychologists call implicit attitudes.

[25] This approach has been influential. De Becker designed the MOSAIC Threat Assessment System that is used by many police departments to screen threats of spousal abuse, and is also used to screen threats to members of the United States Congress, the CIA, and Federal Justices, including the Justices of the Supreme Court. <https://en.wikipedia.org/wiki/Gavin_de_Becker>



Sample Images from Correll et al. (2002)

The results are deeply unsettling. Participants are more likely to shoot an unarmed black man than an unarmed white man and are more likely to fail to shoot an armed white man than an armed black man (Correll et al., 2002). Measures of implicit bias like the IAT predict these results. People who demonstrate strong implicit racial biases (in particular, strong implicit associations between "black" and "weapons") are more likely to make these race-based mistakes than people who demonstrate weaker or no implicit racial biases (Glaser & Knowles, 2008). Moreover, while some experimental results have been mixed, a recent meta-analysis finds that police officers fare no better on the shooter bias simulations compared to civilians in terms of unbiased performance (Mekawi & Bresin, 2015). These findings are ominous in light of continued and recent police shootings of unarmed black men in the United States.

Findings like these show that the way we perceive, and act upon our perceptions of, micro-expressions and subtle social signals is often influenced by stereotypes and prejudices that most of us disavow. Shooter bias involves acting on the basis of subtle feelings of fear that most white Americans are more likely to feel (but not necessarily notice themselves feeling) when they are in the presence of a black man compared to a white man.[26] Research shows that these feelings are indeed race-based. For example, shooter bias is exacerbated after participants read newspaper stories about black criminals, but not after they read newspaper stories about white criminals (Correll et al., 2007). These subtle feelings of fear pervade many mundane situations, too, often in ways that only victims of prejudice notice. George Yancy (2008), for example, describes the purse-clutching, averted gazes, and general unease of some white women when they are in an elevator with a black man, such as himself. In commenting on the death of Trayvon Martin, Barack Obama made a similar point:

> . . . there are very few African-American men who haven't had the experience of walking across the street and hearing the locks click on the doors of cars. That happens to me, at least before I was a senator. There are very few African-Americans who haven't had the

[26] One source of evidence for the fact that shooter bias involves acting on the basis of subtle feelings of fear stems from the fact that shooter bias can be mitigated by practicing the plan, "if I see a black face, I will think 'safe!'" (Stewart & Payne, 2008). But planning to think "quick!" or "accurate!" doesn't have the same effect on shooter bias.

experience of getting on an elevator and a woman clutching her purse nervously and holding her breath until she had a chance to get off. That happens often.[27]

So while one's Pre-Incident Indicators might be justifiably set off by a potential assailant's posture or gestures, they might also be set off by an innocent person's skin color, hoodie, or turban. This means that it might be both true that subtle feelings and intuitions can act like social antennae, tuning us into what's happening around us, and also true that these very same feelings can be profoundly affected by prejudice and stereotypes. Our Pre-Incident Indicators might be a valuable source of attunement to the world, in other words, but they might also be a tragic source of moral failing. This is a grave point, particularly given de Becker's recommendations to police officers to trust their intuition about potential criminal suspects.

3.3 The difficulty of the credibility question

How, then, can we tell our morally good implicit attitudes from our morally bad ones? Are there conditions under which one's implicit attitudes are likely to have moral credibility? This is harder to answer than it might at first seem. There are three reasons for this difficulty: a feasibility worry, a conceptual worry, and a "relearning" worry.[28]

The feasibility worry is that, even if we could recognize, through deliberation, the properties of morally good implicit attitudes, in mundane, real-time social interaction, we often have to rely on our spontaneous judgment. It is simply not feasible for creatures like us to rely on deliberation to evaluate our spontaneous actions and reactions most of the time. There are several reasons for this. One is that real time interaction does not offer agents the time required to deliberate about what to say or do. This is evident when you think of a witty comeback to an insult, but only once it's too late. Most of us have neither Oscar Wilde's spontaneous wit nor George Costanza's dogged willingness to fly to Ohio to deliver a desired retort. To be witty requires rapid and fluent assessments of the right thing to say in the moment. In addition to time pressure, social action inevitably relies upon implicit attitudes because *explicit* thinking exhausts us. We are simply not efficient enough users of cognitive resources to reflective check our spontaneous actions and reactions all the time. When we try to do this, we become "cognitively depleted" (i.e., we become mentally tired; Baumeister et al., 1998). And when we become cognitively depleted, the quality of our social interactions is likely to suffer. People who are cognitively depleted are quicker to act aggressively, for example, and are more likely to act on the basis of implicit biases.[29] This points to a final element of the feasibility worry. The minor actions and reactions subserved by our implicit attitudes promote positive and prosocial outcomes when they are fluently executed. Hagop

[27] http://www.huffingtonpost.com/2013/07/19/obama-racial-profiling_n_3624881.html

[28] For original discussion of the feasibility and conceptual worries, see Tamar Gendler's talk "Moral Psychology for People with Brains." See also my discussion of these worries in Brownstein (2016).

[29] For cognitive depletion and anger, see Stucke and Baumeister (2006), Finkel et al. (2009), and Gal and Liu (2011). For cognitive depletion and implicit bias, see Richeson and Shelton (2003) and Govorun and Payne (2005).

Sarkissian (2010, 10) describes some of the ways in which these seemingly minor gesture can have major ethical payoffs:

> For example, verbal tone can sometimes outstrip verbal content in affecting how others interpret verbal expressions (Argyle et al. 1971); a slightly negative tone of voice can significantly shift how others judge the friendliness of one's statements, even when the content of those statements are judged as polite (Laplante & Ambady 2003). In game-theoretic situations with real financial stakes, smiling can positively affect levels of trust among strangers, leading to increased cooperation (Scharlemann et al. 2001). Other subtle cues, such as winks and handshakes, can enable individuals to trust one another and coordinate their efforts to maximize payoffs while pursuing riskier strategies (Manzini et al. 2009).

The conceptual worry is that some prosocial and even ethical actions are constitutively tied to acting spontaneously on the basis of one's implicit attitudes. If this is the case, then we are really stuck with the credibility question, since it is not just that acting deliberatively is sometimes not feasible, but that sometimes acting deliberatively undermines things we value. Some of the phenomena Sarkissian describes fall into this category. A smile that reads as calculated or intentional—a so-called "Pan-Am" smile—is less likely than a genuine—or "Duchenne" smile—to positively affect levels of trust among strangers. Here it seems as if it is the very spontaneity of the Duchenne smile—precisely that it is *not* a product of explicit thinking—which we value. More broadly, research suggests that people value ethically positive actions more highly when those actions are performed spontaneously rather than deliberatively (Critcher et al., 2013). Related to this are cases in which it seems as if the *only* way to act well in a given situation is to act spontaneously, on the basis of implicit, rather than explicit, attitudes. Bernard Williams' (1981) famous example of saving one's drowning spouse without having "one thought too many" can be interpreted this way.[30] Jason D'Cruz (2013) has more directly described cases of what he calls "deliberation-volatility." For example, one might have reasons to eat ice cream for dinner on a whim every once in a while. One of the reasons to do this is that doing things spontaneously can be joyous. Were one to deliberate about whether to eat ice cream for dinner tonight, one would no longer be acting spontaneously. Thus one would no longer have reasons to eat ice cream for dinner on a whim. Deliberating itself is reasons-destroying, in a case like this, in which the distinctive value of the act is tied to the action being done spontaneously. Of course, eating ice cream for dinner too often is a bad policy. The risk of just acting without deliberation is that one will act irresponsibly. This illustrates the conceptual worry, writ small. The value of acting on a whim is sometimes constitutively tied to acting on a whim. The agent can either deliberate about whether to be spontaneous, and thus risk forfeiting what is valuable about being spontaneous, or the agent can just act spontaneously, and thus risk acting poorly.[31]

[30] Williams' direct target is the idea that moral principles are required to justify actions such as preferring to save one's drowning wife before saving a drowning stranger.

[31] See Brownstein (ms) for further discussion of deliberation-volatility.

Finally, the relearning worry is that, were we so fortunate as to have morally credible implicit attitudes, we must find a way to maintain their moral status over time in a world in which they are constantly threatened by exposure to injustice. One's implicit attitudes might come to reflect egalitarian values through practice and effort, for example, but in pretty much any society, one will constantly be bombarded with sexist and racist images, slogans, narratives, and so on that push one's implicit attitudes back toward the status quo.[32] So the practical difficulty for agents in real time is not only knowing how to act on the basis of their morally good implicit attitudes while avoiding acting on the basis of their morally bad ones, but also, recognizing when one's formerly morally credible implicit attitudes have become comprised by living in an unjust world. The difficult task is ongoing.[33]

4. Moral Authority

Of course, it's not as if scholars in the 20th and 21st centuries have all of a sudden woken up and identified a heretofore unrecognized challenge of acting virtuously yet spontaneously. Railton, for instance, proposes an answer to the credibility question by drawing upon classic Aristotelian virtue ethics. He suggests that we look to the conditions that give rise to ethical exemplars:

> With the help of anecdotes, supplemented by some evidence from genuine research done by others, I have made a few, tentative suggestions about when intuitive moral assessments might be expected to have greater credibility—even when they oppose one's own considered judgment: for example, when individuals have wider and more representative experience, a better-developed ability to imagine what things would be like from the standpoints of others, a better 'feel' for the underlying dynamics in personal and social situations, or greater foresight in imagining alternatives. These are also, I think, characteristics of those people whose intuitive moral responses we especially value or trust. What is it about these people that gives their intuitive responses greater authority for us? Is it that they hold moral principles we share? Many people who share our principles are decidedly not individuals to whom we would turn in difficult decisions. I suspect that we seek out people who strike us as having well-developed implicit social and emotional competencies in virtue of which they are better attuned to the evaluative landscape of concerns, values, risks, and potentialities inherent in the actual, messy situations we face. These are individuals whose intuitive

[32] See Huebner (2009) for elaboration of this worry. See also Huebner's reply to Railton (2014) on the Pea Soup blog (<http://peasoup.typepad.com/peasoup/2014/08/ethics-discussions-at-pea-soup-peter-railtons-the-affective-dog-and-its-rational-tale-intuition-and-.html>). Also see Madva (ms) for discussion.

[33] Victor Kumar suggests that people might be able to structure their environments in such a way as to facilitate good implicit judgment, perhaps thereby avoiding at least the feasibility and relearning worries. To whatever extent it is possible to restructure one's environment in this way, I support it. I am not so sure how possible it is, however. Without becoming a hermit, it is hard to see how one could insulate oneself from being bombarded by racist and sexist stereotypes in a culture like ours. Perhaps rather than hiding from the world, then, the solution is to change it. Here the likelihood of success—of sufficiently changing the world around us so that our implicit attitudes are better—seems even slimmer (although I don't take this as a reason not to try). See Brownstein (2016) for more discussion of this question.

assessments are, by our own lights, likely to be more reasons-responsive than our own. (2014, 858)

I find this answer to the credibility question appealing. It is clearly influenced by long-standing views—in both Western and Eastern virtue ethics traditions—about how to cultivate virtuous dispositions. I also think that it comprises a straightforwardly empirical claim.[34] *Do* people with wider and more representative experience, better-developed ability to take the perspective of others, better "feel" for social dynamics, and greater foresight in imagining alternatives really have more morally credible implicit attitudes? This is a broad claim, perhaps too broad to assess just as such.[35] In order to make the question more tractable, in what remains, I'll examine whether the evidence supports Railton's claim with respect to implicit attitudes in particular. Should we expect agents' "intuitive moral assessments"—that is, their implicit attitudes—to have greater moral credibility under these four conditions?

The four characteristics Railton proposes must be "operationalized" in order to locate the relevant experimental data (where it exists). This is relatively straightforward in the case of agents who have "wider and more representative experience" and "a better-developed ability to imagine what things would be like from the standpoints of others." In intergroup psychology, wide and representative experience is known as "social contact," and the claim that intergroup social contact promotes moral ends is known as the "contact hypothesis." Seeing things from the standpoint of others also has a relatively clear analogue in intergroup psychology; it is called "perspective-taking." Researchers have examined whether, and under what conditions, both social contact and perspective-taking change and/or improve agents' implicit biases. Things are murkier, however, when it comes to "social feel" and "imagining alternatives."

4.1 Social Contact

The contact hypothesis has a long history in the study of intergroup prejudice. Originally proposed by Gordon Allport (1954), the core claim of the contact hypothesis is that intergroup contact promotes positive social relationships and helps to undo prejudice. The contact hypothesis is well-supported in the study of *explicit* prejudice.[36] There is also evidence supporting the notion that intergroup contact promotes unprejudiced implicit attitudes. Higher levels of social contact with members of the LGBTQ community is associated with lower levels of implicit anti-gay bias, for example (Dasgupta & Rivera, 2008). Even regardless of one's past experiences, mere exposure to

[34] Railton's view actually comprises two empirical claims. One is that people with these characteristics have implicit attitudes with greater moral credibility. Another is that people with these characteristics have moral authority (i.e., they are, in fact, treated as moral exemplars). I will ignore this second empirical claim. Instead, I will treat it as a normative upshot of the first claim. That is, I will interpret Railton to be saying that people with these four characteristics will tend to have more moral implicit attitudes, and for that reason, we ought to treat people with these characteristics as having moral authority (not to make moral commandments, of course, but to lead by example as moral exemplars).

[35] Although the wave of philosophical literature on the "situationist" critique of virtue ethics can be seen as looking for empirical validation (or a lack thereof) for related broad claims about ethical exemplars. See, for instance, Doris (2002).

[36] For a review of the evidence, see Pettigrew and Tropp (2006) and Pettigrew et al. (2011).

(i.e., contact with) pictures, stories, and information about admired gay men and women appears to lower anti-gay implicit biases (Dasgupta & Rivera, 2008; Dasgupta, 2013). In the domain of implicit racial attitudes, Shook and Fazio (2008) found in a field study that random assignment to a black roommate led white college students to have more positive implicit attitudes towards black people.

There are, however, seemingly necessary conditions under which social contact promotes intergroup harmony. Pettigrew and Tropp's (2006) review of the literature on explicit intergroup attitudes suggests that, for social contact to work, members of different groups must be of relatively equal status, involved in pursuing shared goals, cooperative, and engaged in activities that are sanctioned by a mutually recognized source of authority.[37] One important point is that it is not yet clear whether these same conditions are necessary for social contact to promote unprejudiced implicit attitudes. It is often not a safe assumption that the observed effects on explicit attitudes will translate to equivalent changes on implicit attitudes.[38] A second important point may be obvious: intergroup contact is not always—or perhaps not often—experienced under these conditions. Due to vast stratification in occupations, wealth, and socio-economic status, racial intergroup contact is very often experienced under conditions of inequality. In some cases, "local" status may be more salient. College roommates, for example, may perceive each other as of equal status despite coming from different socio-economic backgrounds. But not always. And when the conditions Pettigrew & Tropp identify for positive intergroup contact *don't* obtain, or aren't salient, things can backfire.[39]

Having wider and more representative experience, understood as having greater intergroup social contact, is a promising path toward promoting unbiased implicit attitudes. The two central open questions for future research are (1) whether the evidence for the conditions under which social contact promotes desirable explicit attitudes are equivalent to the conditions under which social contact promotes desirable implicit attitudes; and (2) whether these necessary conditions are problematically rare or relatively attainable.

4.2 Perspective-Taking

Perspective-taking involves actively contemplating the psychological experiences of others. Like for the contact hypothesis, there is evidence suggesting that perspective-taking leads to more unbiased attitudes. Blatt and colleagues (2010), for example, asked white physician-assistant students to contemplate the experience of black patients prior to clinical interactions, and found that patient satisfaction with these interactions increased. Todd and colleagues (2012) report that perspective-taking leads to more "approach-oriented" behavior toward black people (e.g., placing one's chair closer to a black partner in an interracial interaction) and causes people to be less likely to deny the existence of discrimination. Todd and colleagues (2012) also demonstrate an effect of perspective-

[37] Pettigrew and Tropp inherit these conditions from Allport (1954). Pettigrew and colleagues (2011) report an updated view that anxiety and empathy are the major mediators of the effects of intergroup contact on social attitudes. Conditions that diminish anxiety and promote empathy appear to improve intergroup relations.

[38] For discussion, see Bodenhausen and Gawronski (2014).

[39] See Al Ramiah and Hewstone (2013). Calvin Lai (p.c.) also reports ironic effects of intergroup contact in unpublished data.

taking on implicit attitudes: diminished bias on the standard race-evaluation IAT. Follow-up studies suggest that this effect lasts, at least up to 24 hours after intervention (Todd & Burgmer, 2013).

Todd and Galinsky (2014) have suggested that perspective-taking might work by creating a self-outgroup associative merger. By actively contemplating others' psychological experiences, we strengthen the link between our self-concept and our conception of the outgroup, thereby making members of outgroups more "self-like." This proposal helps to explain a noteworthy finding. The positive effects of perspective-taking appear to be limited to people with positive self-evaluations. People who like themselves, in other words, are more likely to benefit from perspective-taking. Because these people have positive self-evaluations, when their self and outgroup concepts merge, the outgroup is thought to take on the positive evaluation that one holds of oneself.

It is not yet clear if this proposal about a self-outgroup associative merger is correct. If it is, it suggests that perspective-taking promotes more morally credible implicit attitudes only for relatively confident people. Other studies on perspective-taking also warrant caution. Bruneau and Saxe (2012) find that perspective-taking can have negative effects in situations of long-standing intergroup conflict, for instance. Given that many intergroup conflicts—such as relations between black and white Americans, or relations between Israelis and Palestinians—are indeed long-standing, we must be hesitant to endorse perspective-taking whole hog.

My point in raising these worries isn't to cast doubt on the importance of perspective-taking as such. As above, in the discussion of social contact, the point is that further conditions seem to be required in order for Railton's proposal to be secure. Moreover, simply more research is needed.

4.3 Social Feel

There is no extant research (to my knowledge) examining the relationship between people who have a "feel" for the underlying dynamics of social relations and implicit intergroup bias. "Social feel" seems to be not one particular thing, but rather a collection of skills and traits. One thought is that people who are extroverted are likely to have high levels of social feel. Unfortunately, the relationship between extroversion and implicit bias is unclear in the current empirical literature. Another possibility is that social feel is related to "social tuning," or acting in such a way as to create a shared reality with another. One study by Sinclair and colleagues (2005) found that people who are more likely to exhibit social tuning are also more likely to have diminished implicit biases. But the effect was found only when the person being tuned into exhibited explicitly egalitarian attitudes. Other possibilities for getting an empirical grip on social feel include considering people with large friend networks or people who score high on social sensitivity tasks. The relationship between these skills and traits and implicit intergroup attitudes should be studied.

As with social contact and perspective-taking, I think there is reason to be cautious when endorsing the effects of social feel on implicit attitudes. There is the possibility of a double-edged

sword.[40] Social skill may result, in part, from a greater-than-usual ability to recognize common social attitudes and subtle social stereotypes. People with social feel, in other words, may have a feel for "picking up on" social biases. They may be more in touch with, or attuned to, these biases. And why presume that, if they are, this will result in rejecting those biases rather than endorsing them or acting upon them?[41] Indeed, evidence suggests that the relative accessibility of social stereotypes—how easily they come to mind—plays a central role in the likelihood that they will affect one's behavior (Madva, 2016). A related thought is that social feel, while tuning one into the dynamics of social relations between others, may have little to do with the ability to recognize one's own biases. Thus those with social feel might succumb just as much as others to what is known as the "blind spot bias" (i.e., the fact that it is easier to spot others' biases than one's own; Pronin et al., 2002).

Perhaps social feel is a necessary but not sufficient element of having pro-social implicit attitudes. Again, more research is clearly needed.

4.4 Imagining Alternatives

Finally, it is also difficult to know exactly how to test the relationship between imaginative foresight and implicit attitudes. The ability to imagine alternatives is perhaps related to fluid intelligence, but no research (of which I am aware) speaks to the relationship between fluid intelligence and implicit bias. Hofmann and colleagues (2008) report that automatic (i.e., implicit) attitudes toward temptations, such as unhealthy food, have a stronger influence on participants with high working memory capacity. But this finding doesn't necessarily translate to either imaginative foresight or to implicit social attitudes. A related alternative is that imaginative foresight is expressed in the "need for cognition," or the tendency to engage in and enjoy thinking (Cacioppo & Petty, 1982). A few studies have considered the relationship between need for cognition and implicit bias, but none have done so directly.[42] A final possibility is that imaginative foresight is related to creativity. And at least one study suggests that inducing a creativity mindset lowers implicit stereotype activation (Sassenberg & Moskowitz, 2005).

5. Conclusion

[40] Thanks to Alex Madva for this suggestion.

[41] The professional poker player Annie Duke talks about using other players' gender biases against them in this sense, by picking up on the predictions they make about her decisions. For example, some men—almost all the other plays are men—seem to expect that she will play meekly, while others seem to expect that she will go easy on them if they flirt with her. Duke can then use these expectations to her advantage. See <http://www.npr.org/2015/09/28/444236895/how-poker-player-annie-duke-used-gender-stereotypes-to-win-matches>

[42] Florack and colleagues (2001) show that implicit and explicit prejudiced judgments are more likely to correlate in participants who score low in need for cognition. Briñol and colleagues (2002) find that argument strength affects the implicit attitudes of people who score high in need for cognition, compared with people who score low in need for cognition.

I've argued that the neural and computational mechanisms underlying our implicit attitudes gives us reasons to think that the spontaneous judgments these attitudes create can have moral credibility. By examining research on implicit bias, I've shown how the defeasible moral authority of our spontaneous judgments leads to what I've called the credibility question. This question focuses on identifying conditions under which our implicit attitudes ought to have authority for us. Using Peter Railton's proposal as a launching pad, I've considered four plausible conditions. At present, evidence for the salutary effects of these conditions is mostly incomplete, and is at times mixed. There is significant support for the first two of Railton's suggestions—social contact and perspective-taking—although the role of significant mediating and moderating conditions must be examined. There is less evidence for the second two of his suggestions. Of course this does not mean that these suggestions are wrong, but rather that more research is needed. This research must also include longitudinal studies with multiple attitude and behavioral measures, in order to see how durable and "multi-track" the effects of the traits and skills presumably engendered by these conditions are. Future research must also examine in general whether what works for improving the moral credibility of our explicit attitudes also works for improving the moral credibility of our implicit attitudes.

Ultimately, I am in agreement with Railton (2015) when he likens the cultivation of moral implicit attitudes to skill learning: "People, we know, can acquire greater competency in a given domain when they gain more extensive and variegated experience, can make use of what they learn, and benefit from clear feedback. That is the moral of skill-learning generally, from language acquisition to playing championship bridge." Skill learning *does* provide a good model for improving the credibility of our implicit attitudes. Future research must focus on the particulars of which skills we must learn and how best to acquire them.

Works Cited

Allport, G. 1954. *The Nature of Prejudice*. Reading: Addison-Wesley.

Al Ramiah, A. & Hewstone, M. 2013. Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: Evidence, limitations, and potential. *American Psychologist* 68(7), 527-542.

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

Argyle, M., Alkema, F., and Gilmour, R. 1971. The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology*, 1(3): 385–402.

Baumeister, R., Bratslavsky, E., Muraven, M., & Tice, D. 1998. Ego depletion: is the active self a limited resource? *Journal of personality and social psychology*, 74(5), 1252.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. 1997. Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293-1295.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in cognitive sciences*, *9*(4), 159-162.

Beilock, S. 2010. *Choke: What the secrets of the brain reveal about getting it right when you have to*. New York: Free Press.

Blair, J, Mitchell, D., Leonard, A., Budhani, S., Peschardt, K., & Newman, C. 2004. Passive avoidance learning in individuals with psychopathy: modulation by reward but not by punishment. *Personality and Individual Differences*, *37*(6), 1179-1192.

---. 2013. The neurobiology of psychopathic traits in youths. *Nature Reviews Neuroscience*, *14*(11), 786-799.

Blatt, B., LeLacheur, S., Galinsky, A., Simmens, S., & Greenberg, L. 2010. Does perspective-taking increase satisfaction in medical encounters? Academic Medicine, 85, 1445–1452.

Bodenhausen, G. and Gawronski, B. 2014. Attitude Change. In *The Oxford Handbook of Cognitive Psychology*, D. Reisberg (ed.). New York: Oxford University Press.

Briñol, P., Horcajo, J., Becerra, A., Falces, C., & Sierra, B. 2002. Implicit Attitude Change. *Psicothema* 14(4), 771-775.

Brownstein, M. 2014. Rationalizing Flow: agency in skilled unreflective action. 168, 545-568.

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

Brownstein, M. and Madva, A. 2012a. Ethical Automaticity. *Philosophy of the Social Sciences*, 42(1): 67-97.

---, 2012b. The Normativity of Automaticity. *Mind and Language*, 27(4): 410-434.

Bruneau, E., & Saxe, R. 2012. The power of being heard: The benefits of 'perspective-giving' in the context of intergroup conflict. *Journal of Experimental Social Psychology*, 48, 855–866.

Cacioppo, J. & Petty, R. 1982. The need for cognition. *Journal of personality and social psychology* 42:1, 116.

Correll, J., Park, B., Judd, C., and Wittenbrink, B. 2002. The police officer's dilemma: Using race to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83: 1314–1329.

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

Critcher, C., Inbar, Y., & Pizarro, D. 2013. How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*(3), 308-315.

Crockett, M. 2013. Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

Cushman, F. 2013. Action, outcome, and value a dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

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

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.

D'Cruz, J. 2013. Volatile Reasons. *Australasian Journal of Philosophy* 91:1, 31-40.

De Becker, G. 1998. *The Gift of Fear*. New York: Dell.

Doris, J. 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Dunn, B., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience & Biobehavioral Reviews*, *30*(2), 239-271.

Fazio, R. 1990. Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in experimental social psychology*, 23: 75-109.

Finkel, E., DeWall, C. Slotter, E., Oaten, M. and Foshee, V. 2009. Self-Regulatory Failure and Intimate Partner Violence Perpetration. *Journal of Personality and Social Psychology* 97:3, 483–99.

Florack, A., Scarabis, M., & Bless, H. 2001. When do associations matter? The use of automatic associations toward ethnic groups in person judgments. *Journal of Experimental Social Psychology*, *37*(6), 518-524.

Gal, D., & Liu, W. 2011. Grapes of wrath: The angry effects of self-control. *Journal of Consumer Research* 38:3, 445-458.

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

Gawronski, B., Hofmann, W., and Wilbur, C. 2006. Are "implicit" attitudes unconscious? *Consciousness and Cognition*, 15: 485–499.

Glaser, J. and Knowles, E. 2008. Implicit motivation to control prejudice." *Journal of Experimental Social Psychology*, 44: 164-172.

Govorun, O., & Payne, B. K. 2006. Ego—depletion and prejudice: separating automatic and controlled components. *Social Cognition*, *24*(2), 111-136.

Greene, J.D. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin.

Greenwald, A., McGhee, D., and Schwartz, J. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74: 1464-1480.

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. 2008. Working memory capacity and self-regulatory behavior: toward an individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of personality and social psychology*, *95*(4), 962.

Huebner, B. 2009. Trouble with Stereotypes for Spinozan Minds. *Philosophy of the Social Sciences*, 39: 63-92.

---. 2016. Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. In *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, M. Brownstein and J. Saul (eds.). Oxford: Oxford University Press.

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.

Kishida, K., Saez, I., Lohrenz, T., Witcher, M., Laxton, A., Tatter, S., White, J., Ellis, T., Phillips, P., and Montague, P.R. 2015. Subsecond dopamine fluctuations in huam striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*. Doi: 10.1073/pnas.1513619112.

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

Knowles, M., Lucas, G., Baumeister, R., & Gardner, W. 2015. Choking Under Social Pressure Social Monitoring Among the Lonely. *Personality and Social Psychology Bulletin*, *41*(6), 805-821.

Kolling, N., Behrens, T.E.J., Mars, R.B., and Rushworth, M.F.S. 2012. Neural mechanisms of foraging. *Science, 366*: 95-98.

Kumar, V. Manuscript. Empirical Vindication of Moral Luck.

Kumar, V. & Campbell, R. Forthcoming. Honor and Moral Revolution. *Ethical Theory and Moral Practice*.

Laplante, D., & Ambady, N. 2003. On How Things Are Said Voice Tone, Voice Intensity, Verbal Content, and Perceptions of Politeness. *Journal of Language and Social Psychology*, *22*(4), 434-441.

Machery, E. 2016. DeFreuding Implicit Attitudes. In *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, M. Brownstein and J. Saul (eds.). Oxford: Oxford University Press.

Madva, A. 2012. *The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency*, PhD dissertation, Columbia University.

---. 2016. Virtue, Social Knowledge, and Implicit Bias. In *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, M. Brownstein and J. Saul (eds.). Oxford: Oxford University Press.

---. Manuscript. Biased Against De-Biasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle Against Prejudice.

Madva, A. and Brownstein, M. Manuscript. The Blurry Boundary Between Stereotyping and Evaluation in Implicit Cognition.

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 16075-16080.

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

Manzini, P., Sadrieh, A., and Vriend, N. 2009. On smiles, winks and handshakes as coordination devices. *The Economic Journal*, 119:537, 826–854.

Martin, J., & Cushman, F. Forthcoming. The adaptive logic of moral luck. *The Blackwell Companion to Experimental Philosophy*.

Mekawi, Y., and K. Bresin. 2015. Is the evidence from racial bias shooting task studies a smoking gun? Results From a Meta-Analysis. *Journal of Experimental Social Psychology*. Doi:10.1016/j.jesp.2015.08.002.

Michaelson, E. & Brownstein, M. Manuscript. Doing without believing: Intellectualism, Knowledge-How and Belief-Attribution

Müller, S., & Abernethy, B. 2006. Batting with occluded vision: An in situ examination of the information pick-up and interceptive skills of high-and low-skilled cricket batsmen. *Journal of Science and Medicine in Sport*, *9*(6), 446-458.

Müller, S., Abernethy, B., & Farrow, D. 2006. How do world-class cricket batsmen anticipate a bowler's intention?. *The quarterly journal of experimental psychology*, *59*(12), 2162-2186.

Paxton, J., Ungar, L., & Greene, J. 2012. Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163-177.

Pettigrew, T. and Tropp, L. 2006. A Meta-Analytic Test of Intergroup Contact Theory. *Journal of Personality and Social Psychology*, 90: 751-83.

Pettigrew, T., Tropp, L., Wagner, U., & Christ, O. 2011. Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, *35*(3), 271-280.

Pezzulo, G., Rigoli, F., Friston, K. 2015. Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17-35.

Plant, E., & Devine, P. 1998. Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, *75*(3), 811.

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

Pronin, E., Lin, D. Y., & Ross, L. 2002. The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381.

Railton, P. 2009. Practical Competence and Fluent Agency. In D. Sobel and S. Wall (Eds.) *Reasons for Action*. Cambridge: Cambridge University Press, 81-115.

---. 2014. The Affective Dog and its Rational Tale: Intuition and Attunement. *Ethics* 124:4, 813-859.

---. 2015. Dual-process models of the mind and the "Identifiable Victim Effect." In Cohen, I., Daniels, N., & Eyal, N. (Eds.) *Identified versus Statistical Lives: An Interdisciplinary Perspective*. Oxford: Oxford University Press.

Richeson, J. and Shelton, J, 2003. When prejudice does not pay effects of interracial contact on executive function. *Psychological Science*, 14(3): 287-290.

Sarkissian, H. 2010. Minor tweaks, major payoffs: The problems and promise of situationalism in moral philosophy. *Philosopher's Imprint*, 10(9): 1-15.

Sassenberg, K., & Moskowitz, G. 2005. Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology*, *41*(5), 506-514.

Scharlemann, J., Eckel, C., Kacelnik, A., and Wilson, R. 2001. The value of a smile: game theory with a human face. *Journal of Economic Psychology* 22, 617–640.

Seligman, M., Railton, P., Baumeister, R., & Sripada, C. 2013. Navigating into the future or driven by the past. *Perspectives on Psychological Science*, *8*(2), 119-141.

Shook, N., & Fazio, R. 2008. Interracial Roommate Relationships An Experimental Field Test of the Contact Hypothesis. *Psychological Science*,*19*(7), 717-723.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. 2005. Social tuning of automatic racial attitudes: the role of affiliative motivation. *Journal of personality and social psychology*, *89*(4), 583.

Stanovich, K. 1999. *Who is rational? Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K., & West, R. 2000. Advancing the rationality debate. *Behavioral and brain sciences*, *23*(05), 701-717.

Stewart, B., and Payne, B. 2008. Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control. *Personality and Social Psychology Bulletin*, 34: 1332-1345.

Stucke, T. & Baumeister, R. 2006. Ego depletion and aggressive behavior: Is the inhibition of aggression a limited resource? *European Journal of Social Psychology* 36:1, 1-13.

Sutton, R. 1988. Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9-44.

Todd, A., Bodenhausen, G., Richeson, J., & Galinsky, A. 2011. Perspective taking combats automatic expressions of racial bias. *Journal of personality and social psychology*, *100*(6), 1027.

Todd, A., & Burgmer, P. 2013. Perspective taking and automatic intergroup evaluation change: Testing an associative self-anchoring account. *Journal of personality and social psychology*, *104*(5), 786.

Todd, A., & Galinsky, A. 2014. Perspective-Taking as a Strategy for Improving Intergroup Relations: Evidence, Mechanisms, and Qualifications.*Social and Personality Psychology Compass*, *8*(7), 374-387.

Williams, B. 1981. *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Yancy, G. 2008. *Black bodies, white gazes: The continuing significance of race*. Rowman & Littlefield.

Yarrow, K., Brown, P., & Krakauer, J. 2009. Inside the brain of an elite athlete: The neural processes that support high achievement in sports. *Nature Reviews: Neuroscience, 10*, 585-596.