NOÛS 00:0 (2014) 1–30
doi: 10.1111/nous.12089

Attitude, Inference, Association: On the Propositional Structure of Implicit Bias

ERIC MANDELBAUM
Baruch College, CUNY

The overwhelming majority of those who theorize about implicit biases posit that these biases are caused by some sort of association. However, what exactly this claim amounts to is rarely specified. In this paper, I distinguish between different understandings of association, and I argue that the crucial senses of association for elucidating implicit bias are the cognitive structure and mental process senses. A hypothesis is subsequently derived: if associations really underpin implicit biases, then implicit biases should be modulated by counterconditioning or extinction but should not be modulated by rational argumentation or logical interventions. This hypothesis is false; implicit biases are not predicated on any associative structures or associative processes but instead arise because of unconscious propositionally structured beliefs. I conclude by discussing how the case study of implicit bias illuminates problems with popular dual-process models of cognitive architecture.

Implicit biases have received much attention, and for good reason: many pernicious and ubiquitous forms of prejudice are perpetuated because of them. A person with a strong implicit bias against African Americans is apt to smile less at them and to cut off conversations with them sooner (McConnell and Leibold 2001). Such a person also rates African Americans lower than Caucasians on a host of social-status scales. These generalizations hold even when the subject who harbors such a bias has explicitly egalitarian attitudes toward all racial groups. Implicit biases also appear to be a major determinant of institutional bias, since they can explain how a group of explicitly egalitarian people can still make biased group decisions. These are serious problems, and the phenomenon causing them demands serious attention.

Enormous amounts of data have been collected in order to verify the psychological reality of implicit biases. Yet these investigations have been largely atheoretical; comparatively little has been written about the cognitive causes of implicit bias, even by the data-driven, theory-wary standards of social psychology. This is an unfortunate oversight. Examining the workings of implicit bias can illuminate a host of foundational issues in cognitive science, such as the entities that populate the unconscious mind, and how rationally responsive unconscious thought can be. The study of implicit bias is deeply intertwined with questions of how learning interacts with cognitive structure. This relationship can aid in building

© 2014 Wiley Periodicals, Inc.

theories about the development of cognitive structure and can help constrain psychologically realistic models of inferential patterns of thought. Because questions of learning, cognitive structure, and inference interact in investigations of implicit bias, such inquiries are well positioned to inform our models of cognitive architecture, particularly concerning the validity of dual-process models of cognitive architecture.

Of course, there is a practical payoff too. Few attempts to modulate implicit biases have been successful. No current theories explain why these attempts have failed or how future attempts should differ. Investigating the structure underlying implicit bias can help us to identify why previous attempts have been inefficacious and can help direct future interventions to a more productive path. To that end, this essay aims to explicate the psychological structures and processes that underlie implicit bias.

***

Let's begin with a near truism: implicit biases are caused by implicit attitudes. This claim is not wholly trivial—for example, radical behaviorists might quarrel with it. Be that as it may, I will ignore the radical behaviorist position and assume that implicit biases have some cognitive causes.

A far less trivial, but no less widely accepted claim is that implicit bias is, in an important sense, caused by some associative process/associative structure. The associative process/structure is generally assumed but not discussed in any depth. Insofar as its meaning is analyzed, it is usually glossed as some type of evaluative association, such as an association between a valence (e.g., negative affect) and a concept (e.g., BLACK MALE).[1] In order to make the associative hypothesis as strong as possible, I will also consider cases where the putative association is a purely cognitive one that joins two concepts (e.g., BLACK MALE and UNPLEASANT), and cases where the association is a hybrid type, yoking two distinct concepts with each other and a valence.

Call the hypothesis that implicit bias is caused by some sort of associative process or structure AIB (for 'associative implicit bias'). The majority of cognitive scientists, especially social psychologists (those most apt to research implicit biases), believe something like AIB. However, I suspect they may have done so uncritically. I will argue that AIB is at best misleading and at worst false root and branch.

The structure of the essay proceeds as follows. In Section 1, I'll offer some evidence that AIB is indeed ubiquitous. In Section 4, I'll canvass different types of associations in order to examine what exactly AIB commits one to. Then I will introduce a theory of implicit bias that opposes AIB. In Section 5, I will present the main arguments against AIB, arguments that are based on modulating implicit attitudes via reasoning. Section 7 will close the essay with a short examination of how AIB became so popular, and discuss how the downfall of AIB affects the plausibility of certain dual-process theories of cognition.

1. Caveats and Idealizations

Let's begin with some caveats. First, I assume that there is a monolithic phenomenon to be investigated; that is, I assume that there is some causally implicated cognitive structure involved in many, if not all, cases of implicit bias. This is an idealization, and it's probably false, strictly speaking. Implicit *bias* is a normative notion at heart. Implicit biases are a subset of behaviors caused by implicit attitudes, and they are behaviors that most think are normatively suspect. To put it mildly, there's no a priori reason to think that the normative notion covering behaviors should map neatly onto a single cognitive structure.

Second, I assume that a host of standard though distinct tests can be used to uncover implicit biases, including the Implicit Association Test (IAT; Greenwald et al., 1998); the Affect Misattribution Procedure (AMP; Payne 2009); the Go/No-Go Association Task (GNAT; Nosek & Banaji 2001); the Sorting Paired Feature Task (SPFT; Bar-Anan, et al. 2009); the Weapon Identification Task (Payne 2001, Correll et al. 2002); and the Affective Lexical Priming Score (ALPS; Lebrecht et al. 2009). Glossing over the differences between these tests has its pitfalls. All of the tests are reasonably internally reliable, but they are not highly correlated with one another (Fazio & Olson 2003; Nosek et al. 2007a). The ultimate discussion of implicit bias should discuss the cognitive structures that each test reveals, and these needn't be the same structures. That said, AIB is assumed so widely that almost no one denies that associative structures are the things that are, at their base, tested in each paradigm. So I too will use these different tests interchangeably.

Lastly, I will vacillate between discussing different types of implicit biases, treating (e.g.,) implicit racism as sufficiently similar to implicit ageism, at least for the current discussion. Again, this is de rigueur in the literature, but it may ultimately be misleading. For example, one's implicit racism scores on one test tend to be poorly correlated with other tests (note for example the weak correlations between one's race IAT scores and one's error rate on Payne's Weapon Bias task, Payne personal communication), while other biases (say ageism) correlate quite well across tasks (see Nosek et al. 2007a for more discussion of the lack of correlations). Nevertheless, because AIB is assumed across not just experimental contexts but also content-based differences, I too will feel free to go back and forth between different implicit biases.

In the fullness of time these idealizations will probably have to be looked at more skeptically, but for now there are bigger problems afoot.

2. The Associative Theory of Implicit Bias

My goal in this section is to show that the vast majority of theorists who discuss implicit bias believe in AIB. I suspect the prevalence AIB is untendentious, so I'll keep defense of it minimal.

The main instrument for uncovering implicit bias is the IAT. 'IAT' stands for 'Implicit *Association* Test.' Millions of people have taken some form of IAT or other. Of course there are not only some variations on the IAT (e.g., the personalized IAT), there are also other paradigms for uncovering implicit bias. Nonetheless, because of the overwhelming success of the Project Implicit website far more people have taken the IAT, which certainly hasn't hindered the growth of AIB.

But it's not just an etymological artifact that people assume AIB. Almost any paper one can read on the topic expresses explicit (if off-handed) support for AIB. A sampling (italics mine): Nosek et al. say, "The IAT (Greenwald et al. 1998) assesses *associations* between two concepts (e.g., Black people and White people) and two attributes (e.g., good and bad)" (Nosek et al. 2007b, p8). The first sentence of Blair et al. (2001) reads: "Implicit stereotypes are social category *associations* that become activated without the perceiver's intention or awareness when he or she is presented with a category cue" (p828). Rydell and McConnell (2006, p995) write: "Implicit attitudes reflected an *associative system* characterized by a slower process of repeated pairings between an attitude object and related evaluations . . . [implicit attitudes were] unaffected by explicit processing goals, uniquely predicted spontaneous behaviors, and were exclusively affected by *associative* information about the attitude object that was not available for higher order cognition." Here is Lowery et al. (2001): "The IAT can reveal *associations* participants either do not consciously endorse or do not consciously desire to endorse" (p844). Lastly, here is Dasgupta and Greenwald (2001): "Response facilitation is interpreted as a measure of the *strength of association* between object and evaluation" (p801). These quotations represent but a few of the numerous examples I could cite.

The reader might be concerned that theorists sometimes speak of implicit 'attitudes' as opposed to 'associations.' However, this point does little to challenge AIB, for 'attitude' is never used to commit oneself to something more ontologically committal than associations. For better or worse, implicit attitudes are certainly not being conceptualized as propositional attitudes in the philosophical sense, since these aren't hypothesized to be relations to propositions at all.[2] 'Attitude' is most frequently just used as a synonym for 'association.' For example, the FAQ page of Project Implicit explains that an *attitude* "is a positive or negative evaluation of some object. An *implicit attitude* is an attitude that can rub off on associated objects".[3]

Though it is easy to find theorists flying the associative flag, there is scant discussion of what exactly the associationist's story amounts to. In the next section, I distinguish between some varieties of associationism, so that we can sidestep the fallacious inferences that have plagued the literature and hone in on the relations that are most germane to evaluating theories of implicit bias.

3. Varieties of Association

Associationism's historical prevalence and continued support is due in part to its promise to do a lot of empiricist work with very little machinery. Consider the question of how many mental processes there are. For a faculty psychologist or modularity theory aficionado, the question cannot be answered a priori. No matter the stripe of faculty psychologist, there are multiple mental processes to be posited and which (and how many) mental processes there are cannot be discerned without some empirical observation. In contrast, if you are sympathetic to empiricism, then you should find the proposal that there are multiple mental processes irksome. The more mental processes there are, the worse off one is from an empiricist standpoint because everyone, empiricists included, have assumed that mental processes are innate.[4] So the more mental processes we have, the more innate machinery we're stuck with.

Associationism allows for an empiricist-friendly ontology by using the process of association to cover a suite of disparate phenomena, in particular how certain information is learned (associative learning); the structure of certain mental states (associative structure); and the way in which certain thoughts relate to other thoughts (associative transitions in thought, also known as associative [Jamesian] 'trains of thought').[5] So, at a minimum, we have one use describing a type of world-to-mind causation (in learning), another for mental structures that have a particular type of non-propositional form, and a third for non-computational mental transitions of a certain sort (a type of thinking). In order to assess AIB it is important to understand the relation between associationist treatments of these different phenomena. This is because the inference from one sense of association to another is invalid without further argument and evidence.

Learning, thinking, and the structure of thought are very different phenomena, so it should strike one as curious that they are lumped together. Outside of associationists such groupings are non-existent. For example, theorists who are inclined towards a Quinean bootstrapping theory of concept acquisition (e.g., Carey 2009) would never posit bootstrapping as an account of *thinking*; to do so would be a category mistake.[6] But part of the beauty of associationism is its parsimony: associationists can posit just one mental process, the ability to associate ideas, and assume that this process can serve as a theory of thinking, learning, and cognitive structure. However, there is no a priori reason to have a theory of association hold over all three categories. For example, there is no logical inconsistency in having associative transitions between propositions. No doubt propositional structures allow for logical inference, e.g., they support inferences to THERE IS A CAT from THERE IS A BLACK CAT. But one's train of thought can contain propositions yet still unfold in an associative manner. If in the past I have usually gone to the pub at 10pm I may end up associating the thought IT IS 10PM with I SHOULD GO TO THE PUB without their being any further inferential (or even cognitive) relationship between the two thoughts. In this instance we have an associative transition holding between propositional structures.

As for associative structures, they are just concatenations of mental states such as the coactivation of two concepts, or the coactivation of a concept and a valence. Associative structures can be doubly dissociated from associative learning: one can gain associative structures from non-associative learning and associative learning can directly lead to the acquisition of propositional structures. As evidence for the former, note that one might gain an associative structure without any particular reinforcement pattern (such as our associations for SALT and PEPPER or ULTERIOR and MOTIVE) and from one shot learning (such as in cases of 'taste aversions', when one has gastrointestinal distress after eating a particular type of food, Logue et al.

1981). As evidence for the latter, note that one can learn to associate propositions through classical or operant conditioning (see, e.g., De Houwer 2009 and citations therein). Merely having a behavior reinforced in traditional ways does not ensure that any associative structure will be acquired. Take the (presumably apocryphal) tale about Skinner in the classroom. Supposedly one of his classes decided to condition him in the following way: every time he wrote on the right side of the board, the students shuffled their papers, and the shuffling papers acted as negative reinforcers. It's possible that this could have caused Skinner to associate the thoughts I AM WRITING ON THE RIGHT SIDE OF THE BOARD and THE STUDENTS ARE RESTLESS. If so, Skinner would have acquired these propositional thoughts through associative learning, and the transition between the two thoughts would be an associative one.

This last dissociation is of particular importance for the case of implicit bias. Theorists may think that implicit biases are acquired through some form of associative learning, but to infer from that to the idea that an associative structure has been acquired is unwarranted.[7] Few investigations of implicit bias explicitly concern themselves with the acquisition of the biases. Therefore, it is hard to read AIB as pertaining to associative learning, or any particular acquisition story for that matter. Consequently, I will understand AIB as the thesis that implicit biases a) have associative structure and b) enter into associative transitions (and do not enter into logical ones). Thus, questions about the acquisition of biases are beyond the scope of the current discussion; my focus will instead trace AIB's scope and only analyze associative structures and transitions.

Some may see this as problematic, since reinforcement patterns are directly observable and associative structures aren't (see, e.g., Greenwald & Nosek 2009, see also Ferguson et al. 2014). If associative learning doesn't entail the acquisition of associative structures, how can we tell if we are dealing with an associative structure? Though associative learning, transitions and structures can be dissociated from one another, there is a connection between them that will be of much probative value: how to modify associations. We can infer whether a given cognitive structure is associative by seeing how certain types of information modify (or fail to modify) behaviors under the control of the cognitive structures. In associative learning, one can condition stimuli and responses, or stimuli and stimuli, through certain patterns of reinforcement. Learning to associate two stimuli means tending to have the representations of those stimuli co-activate. Just as we'd destroy stimulus/response associations through changing certain external contingencies, so too can we change stimulus/stimulus associations, the co-occurrence of certain representations, through changes in the external stimuli. For a concrete example, imagine you've taught a pigeon to associate a peck of a bar with receiving some food. If you want to extinguish the connection between the bar peck and food, you allow the bar peck to proceed without pairing it with the food—sooner or later the organism will pick up on the lack of contingency. Accordingly, if you want to extinguish a person's association of salt and pepper, a similar technique should be used: whenever one is around, make sure the other is quite scarce (and vice versa). Sooner or later the association will be weakened, hopefully even extinguished.

To put it abstractly, mechanisms for successful extinction work either by presenting one of the relata without the other, which is the normal extinction paradigm, or via counterconditioning, which can only happen when a valence is involved in an associative structure. Either way, if you want to modify an associative structure, you have to change certain contingency structures. But note what would be silly to do: throw some rational argumentation at an associative structure. If every time a bell rings I get hungry because of my history with bells and feeding, then giving me good arguments for not getting hungry when the bell rings won't in fact affect my hunger. Likewise, say I associate 10PM and BEER. Being told not to associate the two together is all fine and well, and may even be sage advice, but nonetheless it won't do much to break my associative structure. I'll still end up thinking about beer when I realize it's 10pm (except now maybe I'll also feel bad about it).

In sum, it is invalid to infer that because something was associatively learned, an associative structure has been acquired. Likewise, one cannot baldly assert that an associative structure must have been acquired through associative learning. Furthermore, it is invalid to infer that because a certain associative transition has occurred, the elements involved in that transition are associative structures. Although all of these inferences are invalid, there is a relation in the area that is integral to understanding implicit bias and the nature of unconscious thought in general, which is predicated on how one modulates an associative structure: if you want to break apart an associative structure your options are limited; you can extinguish it by presenting one of the relata without the other or you can countercondition it, by changing the valence of the relata. Those are the only routes to modulating an associative structure. In other words, if rational argumentation (or any logical or evidential intervention) can be used to modulate an implicit attitude, then that implicit attitude does not have associative structure. This insight will be used as the central principle for the arguments that follow.

> **AIB Principle**: Implicit biases (a) can be changed by changing certain environmental contingencies and (b) can only be changed by changing certain environmental contingencies, i.e., by extinction or counterconditioning.

If the AIB principle is true, then no logical or evidential interventions should directly work to change implicit attitudes.

In the next section, I aim to show that (b) is false.[8] But before I get there it may be worth introducing a positive picture that opposes AIB, one that I will refer to as the Structured Belief hypothesis. The hypothesis isn't particularly revolutionary, so it's a bit surprising how much it's been overlooked. The idea is that implicit bias is underwritten by unconscious beliefs. These beliefs are not just mere associations, but they are honest-to-god propositionally structured mental representations that we bear the belief relation to. So instead of maintaining that implicit racists merely associate (say) BLACK MALE and DANGEROUS, the hypothesis is that implicit racists have a belief with the structure BLACK MALES ARE DANGEROUS. This may appear to be a small tweak, but it's not. Because of the prevalence of dual-process theories, having unconscious structured representations that enter into logical relations with one another is anathema to many cognitive scientists. For example, many dual-process theories (e.g., Sloman 1996, Smith & Decoster 1999, Frankish 2010, Kahneman 2011, Evans & Stanovich 2013) rule out unconscious structured, logical processes.

Structured beliefs can be reason-responsive: there's a logic for how to transition from one structured belief state to another.[9] A toy example: if one has the belief IF THERE IS SMOKE THERE IS FIRE and then comes to believe THERE IS SMOKE then, ceteris paribus, one will infer THERE IS FIRE. In other words, Structured Beliefs allow for the possibility that the causal roles of certain beliefs mirror the implicational structure of the contents of the beliefs.[10]

By hypothesis, Structured Beliefs are unconscious. Perhaps they are necessarily unconscious, or perhaps they are just frequently unconscious. In particular, what is unconscious is the structure of thought undergirding implicit bias. What needn't be unconscious though, are valences that can be associated with Structured Beliefs. Valences can be attached to either concepts (as in, e.g., a microvalence story, Lebrecht et al. 2012) or even whole propositions—a thought can have a certain valence that emerges even though its constituents don't have that valence.[11] These valences might be conscious even though the structures that they are associated with are not.[12]

A commitment to Structured Beliefs doesn't preclude commitments to other entities in the mind's unconscious. I have no reason to deny that there are also free-standing associations that are not connected to Structured Beliefs; in fact, I suspect that there are free-standing associations, but that such associations do far less causal work than is often supposed, especially in the implicit bias literature.[13] Furthermore, I also suppose that there are prototypes, exemplars, probability distributions, mental images, and a whole host of other flora and fauna that lurks in the background of the mind. Again though, I suspect that none of those phenomena matter for assessing AIB or uncovering the underlying structure of implicit bias.

A caveat about Structured Beliefs: one might think that something can't be a belief unless it has a certain normative profile (e.g., it "aims at the true", Velleman 2000, or falls under some other preferred epistemic constraints, such as probability coherence, Christensen 2004). Philosophers of such a persuasion would be hesitant (to say the least) to call unconscious, non-endorsed thoughts beliefs. This paper is not the place to argue about exactly what counts as a belief. To put my cards on the table, I think these states are beliefs and yet do not belong to the same natural kind as the conscious judgments that are often also called 'beliefs' (for discussion see Mandelbaum 2014). If you think there are some heavy-duty normative requirements that hold over belief such that these states don't deserve the belief appellation, then feel free to see the alternative to AIB as the 'Structured Thought' alternative. Structured Thoughts will do just as well as Structured Beliefs for serving as a reasonable alternative to AIB.

If you are sympathetic to a theory on which reasons can be causes, or where thinking consists of making logical inferences, you should be sympathetic to this sort of propositional view. More to the point, if you think that implicit attitudes themselves can serve as premises of inferences or are responsive to reason, then you should very sympathetic to the Structured Belief view and more than skeptical of AIB (for associations aren't truth apt, so can neither be reasons nor serve as premises in inferences). Nevertheless, one could reasonably reject AIB and still not accept the Structured Belief view. My main goal will be to cultivate a healthy skepticism towards AIB based around the idea that associations don't have the right properties to explain why certain interventions are successful at changing implicit bias. I think that the evidence we'll peruse is consistent with the Structured Belief hypothesis, but I won't argue at length that the hypothesis is true. Showing that AIB is false is enough of a challenge for one paper, but it would be unsporting not to offer a replacement theory, even if I haven't the space to argue for that theory in great detail (though see section 6 for further discussion).

Now that we are clearer on what AIB entails and what an alternative to AIB might look like, we can get down to the business of assessing AIB.

4. The Logical Basis of Implicit Attitudes Interventions

The structure of my main argument follows from the Principle espoused above: if AIB is true, then the only interventions that should be reliably successful can take the form of an extinction process (presenting the 'conditioned stimulus' [or its putative equivalent] without any reinforcer) or a counterconditioning one (presenting the CS with a reinforcer that takes an opposite valence than the one the CS currently has). If there are interventions that reliably work to counteract implicit bias and if these interventions are not reducible to extinction or counterconditioning, then we have evidence that the structure of implicit bias is not, after all, underwritten by associations. In which case AIB is false. So, for every datum about to be canvassed, the goal of the defender of AIB is to show how the datum could be understood as an associative intervention.

To make the associationist picture as strong as possible, I'll relax the restriction on what can be associated with what. Pure associationism has no propositional structures at all, so the question about whether the cognitive representations (of phrases, clauses, propositions, etc.) can be associated never arises. I'll ignore the pure associationist position in what follows and assume that associations can hold between any types of mental structure, for if I can show that an AIB picture with more degrees of freedom cannot explain the data, I thereby falsify the more restrictive pure associationist picture too.

Additionally, some widely held but overly strict forms of AIB (e.g., Rydell & McConnell 2006) claim that implicit attitudes are handled by a purely associative processor. This idea stems from a type of dual-process theory, one that disallows crosstalk between propositional processors and associative ones. Again, if I can cast doubt on a position that allows for such crosstalk, I thereby cast doubt on the more restrictive position too.

Lastly, I'll also grant AIB the ability to inhibit certain associations. Associationism taken alone only specifies when the relevant entities (stimuli, responses, concepts, valences, etc.) become associated; it does not provide a framework for understanding what situations inhibit other entities from becoming active. But in order to make AIB as strong as possible, I'll assume it has the capacity to inhibit activations (even when the principles and instances of inhibition appear to be ad hoc).

With the groundwork now established, let's turn to the data.

4.1 Associative additivity and balance theory

A venerable social psychological hypothesis is that people interpret the enemy of their enemy to be their friend. This posit is made clear in Heider's Balance Theory (Heider 1958) and various forms of dissonance theory (e.g., Aronson & Cope 1968). In other words, if I dislike Assad and I know that Assad dislikes Szymborska, then I'll be apt to like Szymborska. However, if my mental transitions weren't inferential but were instead merely associative, then we should find evidence of the normal second-order conditioning effects, which would supply the opposite prediction from that of Balance Theory (Walther 2002). The associationist predicts that since I have a negative association with Assad and since I know he has a negative association with Szymborska (and assuming I have no other information about Szymborska), then I should in fact have a negative association with Szymborska, for Szymborska has been paired with two negative stimuli. In other words, if we can find support for something like Balance Theory among implicit attitudes, we can be reasonably sure that implicit attitudes aren't partaking in an associative process but instead have some sort of logic operating over them.

Previous results from dissonance theory lend support to this system of appraisals. For example Aronson and Cope (1968) found that a subject tends to like a person who is mean to someone who had previously mistreated the subject. In other words, if someone has been mean to you, you will tend to like people who are mean to your antagonizer; thus, two negatives can make a positive. Assuming dissonance theory is evidence of propositional reasoning (a reasonably untendentious assumption), this datum shouldn't be surprising.[14] But can we find a similar case of where two negatively valenced implicit attitudes somehow lead to a positive evaluation?

In fact we can. Gawronski et al. (2005) examined the effects of cognitive balance on implicit attitudes. The experimenters first introduced participants to a photo of an unfamiliar individual (CS1). The CS1 was then paired with statements that were either consistently positive or consistently negative, thus conditioning the subjects to respond to the CS1 with the designated evaluation. A different unfamiliar individual (CS2) was subsequently introduced and subjects were told whether the CS1 liked or disliked the CS2. Finally, subjects were given an affective priming task to assess the subjects' implicit attitudes toward both the CS1 and the CS2. The procedure was then replicated for five other novel CS1 and CS2 pairs.

A cursory inspection of the data might give the impression that an associative process is at work, since the experimenters found that the subjects had positive implicit attitudes toward the CS2s whom the positively valenced CS1s liked. In other words, if you thought someone was good and were told that the good person liked someone else, then you would form a positive implicit attitude toward that third person. An associative account can handle this datum, as it just shows that positive valence + positive valence = positive valence, which is exactly what an associative (and, for that matter, propositional) account would predict.

However, upon further inspection, the other findings are quite destructive to an associative account, for the experimenters also found exactly what Heider would've predicted, but not at all what AIB predicts. A negatively valenced CS1 who disliked a CS2 caused the subjects to *like* the CS2. In other words, if you were originally taught that a person was bad and subsequently learned that this person dislikes another person, you then would like that second person. This makes perfect sense to those who subscribe to the logic of 'the enemy of my enemy is my friend,' but it is anathema to proponents of AIB. Note first that 'the enemy of my enemy is my friend,'—whether normatively respectable or not as a moral theory—at least has a rational basis, and associations have no such basis. But more damagingly, AIB predicts the exact opposite effect from the one found. AIB predicts that you should have enhanced negative reactions toward the CS2 because you a) are encountering the CS2 as yoked to negatively valenced CS1 and b) are activating another negative valence because you are told that the CS1 *dislikes* the CS2. I have no opinion on whether two wrongs make a right, but I'm confident that if you find two negatives making a positive, what you've found is a propositional, and not an associative, process.

Perhaps unsurprisingly, similar anti-associative effects were found for individuals that the person you disliked liked. If you were conditioned to dislike a CS1 and learned that CS1 liked a CS2, then you would tend to have negative implicit attitudes toward that CS2. If the associative story were true, then you would only be able to sum associations, and the combination of a negative and positive association should (ceteris paribus) predict a near neutral response, not a negative one. AIB fails where a little (folk psychological) reasoning succeeds.[15]

At this point the reader might be wondering whether associationism doesn't have a few more degrees of theoretical freedom. In particular, one might wonder whether some sort of inhibition couldn't be used to help out AIB. Maybe associative processes not only activate chains of associations but also can serve to inhibit certain concepts (or combinations of concepts) from becoming active.

However, inhibition is of little use to AIB for dealing with the issue at hand. What inhibition can do is stop a concept from activating. What it can't do is work as a negation. AIB needs to explain a certain type of negation-like structure, one that can immediately transform (e.g.,) two negatives into a positive. Even if we allow the AIB proponent to claim (in an ad hoc fashion) that a negative response is being inhibited in the double negative set up, that still would not explain how the combination of two negative valences turns into a positive valence.

More concretely, cognitive balance research tells us that if you dislike person A, and A dislikes person B, you'll like B. Allowing inhibition in the picture would mean that the associative system can inhibit the connection between A and B. But inhibiting this connection still doesn't explain why B takes on a positive valence. What we need is a way to explain not just why B isn't negatively valenced (which inhibition can help do) but also why B is positively valenced and AIB-plus-inhibition hasn't the resources to do so.

To reiterate the main point of the subsection: implicit attitudes are sensitive to *rational* relations, and implicit attitudes do not only combine in an associative manner. When a person you don't like dislikes another, you tend to like that other person. So a negative valence when combined with a negative valence somehow results in a positive valence. The 'somehow' is an utter mystery on an associative theory but is sensible on a propositional theory—just as you might consciously reason that you should probably like those that Hitler hates and dislike those that Hitler likes, so too it appears that we unconsciously reason this way. This isn't at all surprising if we have unconscious logical operations, but to have such operations we need some propositional structures. Furthermore, we've seen that in order to explain the data, we need logical, propositional processes—*inferences* and not just associative transitions—working over propositional *structures*. The balance theory data thus do double duty. First, they cause trouble for any AIB-based theory by showing that there are implicit attitudes that have neither associative structure nor enter into associative transitions. Second, insofar as the data demand propositional processes and inferential structures, they lend support to a Structured Belief-type view.[16]

4.2 Argument strength and implicit bias

Brinol et al. (2009) provide data that show how implicit attitudes are sensitive to the strength of arguments. The experimenters examined how argument strength affected race IATs and vegetable IATs (which pit vegetable vs. animals and good vs. bad). In the race IAT intervention, one group of participants read strong arguments encouraging the hiring of African American professors. The strong messages said that the subjects' university should hire more black professors for doing so would increase the number and quality of professors without any corresponding tuition increase; the number of students per class would be reduced 25%, etc. A separate group of subjects received weak arguments, with statements saying that the subjects' university should hire black professors because doing so would allow it to appear to be trendy, allow the current professors to have more time for themselves, etc. After reading these arguments, both groups were given a race IAT. The subjects in the strong argument group had more positive implicit attitudes toward African Americans than those in the weak argument group. Note that this result held even though both groups of subjects had read paragraphs about African American professors and the messages mentioned 'African American professors' the same amount of times. In other words, the associations were controlled across both groups, thereby nullifying the mechanism that AIB dictates would have to be active in order to get a change in implicit attitudes.

More concretely, if AIB were true, then the mere coactivation of AFRICAN AMERICAN and PROFESSOR should improve bias scores on a race IAT (assuming that PROFESSOR takes a positive connotation). Furthermore, the weak arguments still contained positively valenced mentions of African Americans—if implicit biases were purely associative, then this manipulation should be effective, for the manipulation is just a form of counterconditioning. Nevertheless, the logical force of the argument appears to be the critical causal variable: strong arguments alleviated the bias, weak arguments did not have any effect.[17] Whereas AIB is damaged by these data, the Structured Belief hypothesis is bolstered: argument strength is exactly the sort of thing that Structured Beliefs can interact with. Structured Beliefs have the right causal powers to be affected by reasoning and inference, and of course inferences are sensitive to the strength of arguments and weights of evidence.

A similar result was also found when using vegetables as stimuli. Subjects either read highly persuasive messages (e.g., vegetables have more vitamins than most supplements, which is particularly beneficial during exam and work out periods) or unpersuasive messages (e.g., vegetables are becoming more popular at weddings because they are colorful and look beautiful on plates). Again, both messages contained positive valences associated with vegetables; for instance, the weak messages mentioned that vegetables are beautiful and associated them with weddings. Nonetheless, only the strong messages had any effect on participants' scores on a vegetable IAT. Giving participants good arguments in favor of vegetables increased the participants' opinions of vegetables; contra AIB, merely associating vegetables with positive valences had no such affect.[18]

4.3 Evidential adjustment to peer attitudes

Sechrist and Stangor (2001) uncovered another effect of logical reasoning on implicit attitudes by showing that implicit attitudes are adjustable in light of what one's peers think of the topic at hand. Participants were first given a test to survey their racial attitudes (the Pro-Black Scale; Katz & Hass 1988), and they were subsequently split into high and low prejudice groups. These groups were then told that the same test had been given to their peers (other students at the same college). Half of both the high and low prejudice groups were then told that the vast majority of their peers (81%) agreed with them, thus forming two high-consensus groups. The other half was told that the majority of their peers disagreed with them, creating two low-consensus groups. One by one, the (all-white) subjects were then asked to take a seat in a waiting room, where an African American confederate awaited. The dependent variable simply tracked how closely each participant sat to the confederate. Seating distance has been shown to be highly predictable by implicit attitudes and has thus become used as a diagnostic measure of implicit attitudes (Macrae et al. 1994; Nosek et al. 2007b).

The main finding was that conformity to one's peer group affected seating distance. Specifically, people who rated as having low-prejudice on the pre-test measure sat closer to the African American confederate when receiving high-consensus feedback than did the low-prejudice group members who were told they were outliers (the low-consensus feedback group). In other words, if you had low prejudice and you found out that your peers agreed with you then you sat closer to the confederate than if you found out that your peers disagreed with you. Likewise, if you had high prejudice and found out that your peers agreed with you (i.e., that they too were highly prejudiced against African Americans), then you sat farther away from the confederate than did the high-prejudiced subjects that did not have their beliefs subjectively validated by their peers (ibid. p649).

It's difficult to see how any associative story can explain away this data. For example, suppose that you are a high-prejudiced person in the experiment. In that case, activating the concept AFRICAN AMERICAN should produce negative affect, for we have been assuming that the associative story is, in part, an evaluative one where negative valences get attached to categories corresponding to the stereotyped groups. After having the negative valence activated, you now receive feedback. We know from dissonance theory that it feels good when people agree with us, and it hurts when people don't (see, e.g., Elliot and Devine, 1994). In other words, finding out that people agree with us creates positive affect. So on a purely associative story, when the high prejudiced person gets positive feedback he should in fact feel better and experience more overall positive affect. In which case there's little reason to expect him to sit further away from the confederate—his fear response should be inhibited, not activated. Likewise, the high-prejudice person who receives negative feedback, finding out that his peers disagree with him, should now have his negative affect exacerbated. Yet this exacerbation of negative affect causes him to move closer, not further away from the experimental compatriot. The associative story makes the wrong predictions in this case, but the Structured Belief story can explain the behavior because it doesn't just rely on affect: it assumes that the representations underlying implicit bias can be adjusted in the face of countervailing (subjective) evidence.

The AIB proponent may feel inclined to retort that happiness increases stereotyping, and though there is evidence for this (e.g., Bodenhausen et al. 1994), it's of little use for AIB. For one thing, associationism just takes it for granted that this is the case and doesn't explain why this is. But even if one did assume that happiness increases stereotyping as a basic law it still wouldn't explain this data, for the low-prejudice subjects move their seats *closer* upon receiving confirmatory feedback; if happiness increases stereotyping the subjects should be seating themselves farther away from the target, not closer.

In case the reader is skeptical of this data because of the behavioral measure, it is worth noting that the same effect was reproducible on a more straightforward implicit measure: a lexical decision task where AFRICAN AMERICAN was primed and then stereotypical or non-stereotypical traits served as targets. If, before taking the lexical decision task, a subject was told that their peer group agreed with their highly prejudiced views, then the subject was quicker in responding to stereotypical traits after encountering an African American prime than if the subject was told their peer group disagreed with their views. It's difficult to see how an associative story can handle this data, for all the associations here are controlled for—the only real change is in informational consensus: finding out that people disagree with you can cause you to loosen the connection between stereotype and target. It's hard to understand how this very non-extinction based paradigm can help one to loosen the association between stereotype and target when *all the intervention does is put stereotype and target in even closer contiguity in all of the conditions*. This contiguity of stereotype and target should facilitate, not inhibit, the association.

What appears to be happening in both conditions of this study is that subjects are adjusting the strength of their implicit attitudes in virtue of what they took to be germane evidence, the opinions of their peers. Such evidential adjustment is, of course, quite consistent with the Structured Belief view. No doubt associations can be strengthened or weakened; however they are not changed in this way based on incoming evidence, but instead based on conditioning or valence summing. If anything, this experimental paradigm should've strengthened associations in all conditions, since the experiment heightens the contiguity between the category and its stereotyped attributes. That this isn't what occurred is a serious problem for the proponent of AIB.

4.4 Logical effects: Abstract supposition and concrete learning

The last piece of evidence I'll discuss directly examines the effects of conditioning versus reasoning on implicit attitudes. Gregg et al. (2006) ran a series of experiments using the evaluative conditioning paradigm in order to probe dual-process views that accept AIB and in doing so uncovered strong anti-AIB evidence. Here's their take on the type of dual-process model they have in mind: "Smith and DeCoster (1999) have postulated the existence of two complementary representational systems: a rule-based one, in which sudden transformations of serial representations (or symbols) occur, and an associative one, in which gradual transformations of connectionist representations (or weights) occur" (Gregg et al. 2006 p2). Gregg et al. were attempting to see whether different types of learning, "abstract" and "concrete" forms, differentially affected implicit attitudes. They defined concrete learning as "the act of cognitively assimilating multiple pieces of information about the characteristics of an object or, alternatively, of assimilating the same piece of information multiple times. Thus, reading a detailed descriptive account of some object or undergoing a session of intensive associative conditioning (De Houwer, Thomas, and Baeyens, 2001) would both qualify as instances of concrete learning" (p4). On the other hand, abstract supposition was taken to be "hypothetically assuming that an object possesses particular characteristics. Thus, entertaining the idea, out of the blue, that a novel object is X or −X or that an existing object known to be X is in fact −X (or vice versa), both qualify as instances of abstract supposition (p4)." For Gregg et al. the critical differences between the two types of learning is that in concrete learning you continually encounter instances of the object to be learned, whereas in abstract learning no such exemplars are available (p4). Thus, "In other words, the act of abstractly supposing that some state of affairs is the case involves entertaining cognitions that are purely *formal* and *symbolic*. Consequently, abstract supposition should be particularly well-suited to activating explicit representations—namely, those that are 'rule-based', 'rational', and 'constructed'—whereas concrete learning should be particularly well-suited to activating implicit representations—namely, those that are 'association-based', 'experiential', and 'dispositional'" (p4; italics mine). Accordingly, what should not happen, at least as far as AIB is concerned, is having any 'abstract learning' variable affect implicit attitudes, at least insofar as those attitudes are wholly associationistic.

The stimuli used by Gregg et al. consisted of two fictitious tribes, the Niffites and the Luupites. Participants were split into a 'concrete learning' group and an 'abstract learning' group. In the concrete condition the two tribes were paired with highly valenced words, in a traditional evaluative conditioning paradigm (e.g., Niffites would be paired with 'barbaric' and Luupites 'benevolent'). There were 240 rounds of the conditioning induction. In the abstract supposition group, participants were merely asked to suppose that there were two tribes, one of which was peaceful and civilized and the other savage and barbaric. Note that the abstract learners underwent no conditioning at all. All participants were then given a Niffites/Luupites good/bad IAT and no differences were found between the abstract learners and either of the concrete learners; all showed the same (e.g.,) anti-Niffite biases to the same degree.

This should be quite surprising to supporters of AIB. The concrete learners underwent an intensive evaluative conditioning paradigm,[19] whereas the abstract learners were only asked to suppose that one group was benevolent, the other malevolent. Even if one presumed that an association could be created from an encounter with a hypothetical supposition, the associative strength should at least be stronger for the group that encountered 240 pairings as opposed to the group who received a single-sentence instruction. Supporters of AIB haven't the theoretical tools to explain away such data; what they need is some way to explain how 240 trials of evaluative conditioning can lead to the same level of valenced implicit attitudes as that of a group that undergoes no conditioning but merely supposes something to be true. Though this data is anathema to AIB, if instead we supposed that what was underlying implicit attitudes were Structured Beliefs, then we would have an answer for what was causing the observed effects: all groups formed the same (strong) belief that Niffites were bad while Luupites were good.

The finding was not accidental. In a separate experiment, the data were not only replicated but were also extended in a way that shows that the implicit attitudes were inferentially promiscuous. The experimenters re-ran the aforementioned procedure, but with one added twist: at the end of the first preference induction participants in the abstract learning condition were introduced to two new groups, the Jebbians and Haasians, and were told that the Jebbians were equivalent to the Niffites, and the Haasians equivalent to the Luupites. Participants were then given either a Jebbian/Haasian good/bad IAT or a Niffites/Luupites good/bad IAT. The findings were that the mere mention of equivalence between the groups was enough to make the IAT results indistinguishable across subjects. That is, identical implicit attitudes were formed whether one underwent extensive evaluative conditioning (as the concrete learners did) or was merely told a) that the Niffites are cruel and b) that the Jebbians are equivalent to the Niffites (as the abstract learners were). This finding shows that implicit attitudes can have cognitive effects that are not predicated on chains of conditioning, but are modulated based on acknowledgement of logical equivalence. Once again, it's easy to see how this logical inferential promiscuity could be the case if implicit attitudes were Structured Beliefs that were reason responsive—the participants just form the beliefs that the Jebbians are good and the Haasians are bad based on the knowledge that, e.g., the Jebbians are equivalent to another good group—but it's quite difficult to see how AIB could explain this data.

The problems for AIB become more acute when one looks at what happens when the experimenters attempted to expunge these implicit attitudes. After the subjects had formed their implicit attitudes and received an IAT, the experimenters then attempted to countercondition the attitudes away. Importantly, the subjects' attitudes would not extinguish via counterconditioning—counterconditioning just didn't work.[20] Moreover, *pace* AIB, a logical intervention did in fact have an impact on participants' implicit attitudes. The abstract learners were simply asked to now suppose that the good-natured tribe was evil and vice versa. Merely entertaining this piece of information caused the abstract learner's implicit attitudes to become less extreme.[21] Thus, the logical intervention, being told that what they previously had learned was in fact backwards, was more effective than intensive counterconditioning. It's hard to see how AIB could explain this datum, since AIB posits that a) counterconditioning is the route to changing implicit attitudes and b) that reasoning should have no effect on implicit attitudes.

The experimenters take their data to indeed be problematic for dual-process theories: "Our first two experiments therefore empirically contradict what dual-process models can plausibly be taken as implying, namely, that automatic attitudes are relatively immune to sophisticated symbolic cognition" (p9). Yet, the experimenters do not end up supposing that implicit attitudes are underwritten by anything like structured mental representations. I suspect the reason for this is that they were impressed by the fact that although the counterfactual supposition changed the subjects' implicit attitudes in the expected direction, it did not completely reverse or eliminate their implicit attitudes. In other words, though the implicit attitudes were fungible, they were not instantaneously destroyed. This immunity to complete destruction may mislead some theorists into thinking that Structured Beliefs cannot be implicit attitudes.

It is important to keep in mind that just because a state is relatively immune to change does not mean that that state is associative. Structured Beliefs are responsive to logic, but that does not entail that they will always be rationally revised in response to evidence in normatively respectable ways. The venerable tradition of cognitive dissonance is forever uncovering Structured Beliefs that are not all that amenable to change even by quite persuasive evidence (see, for example Festinger et al. 1956 for a particularly harrowing example of this). One might harbor strong beliefs that a) aliens exist and b) if aliens exist then they will come down tomorrow and dance on the stairs of the Washington monument, yet still not only keep, but in fact increase, one's belief that aliens exist even after an uneventful tomorrow comes and goes.

To see if one is dealing with a Structured Belief we only need to see if it is modifiable by logical reasoning. Mere intransigence or irrationality cuts no diagnostic ice. Irrationality is an error, but to paraphrase Ryle, errors are a sort of cognitive achievement; to err, you must first be playing the game, since errors assume a background of minimal competence. Structured Beliefs can be updated rationally (when things go right) or irrationally (when they don't); in contrast, associations are wholly arational—they can't even be updated irrationally because they are not rationally responsive, whereas Structured Beliefs are at least capable of being rationally responsive.[22]

The takeaway from this subsection is this: if AIB were true, then intensive evaluative conditioning should create stronger attitudes than merely giving subjects a single piece of counterattitudinal information. But what we've seen is that this is not in fact the case. Moreover, though counterconditioning had no effect on subjects' implicit attitudes, telling the subjects that their beliefs were now baseless did in fact have an effect on subjects' attitudes, showing that implicit attitudes are inferentially promiscuous and reason responsive. A Structured Belief hypothesis can explain how implicit attitudes are so responsive, while an associative hypothesis hasn't the tools to do so.

***

I could continue walking through interventions that may initially appear to be associative but on further inspection have logical structure, but I hope by now it is clear that implicit attitudes have more structure than mere associations.[23] I'd like to think that we have a handle on what structure that is: mental representations with propositional structure that function as unconscious beliefs. But I'll settle for less: I'll be happy enough if the reader leaves feeling more skeptical of AIB. Assumptions so widespread don't just die overnight, but perhaps it's time to treat the associative story with a greater skepticism.

5. Associationism and Dual-Process Theories

Earlier I distinguished three kinds of association: association as a type of transition between thoughts, as a mental structure, and as a learning procedure, and argued that it was invalid to simply infer from one sense to another. Distinguishing between these senses of 'association' is important because they illuminate reasonable theoretical possibilities. For instance, one can still hold that implicit attitudes are the products of long-term exposure to particular concatenations of properties, i.e., ambient 'associations' (Lowery et al. 2001, p842), while denying that implicit attitudes themselves are in any interesting sense associative.

Although we've only been discussing the relation between associations and implicit attitudes, it's worth noting that the pure associative story has been long dead in other parts of cognitive science. Not many psycholinguists take associative structure to be the only type of representational structure. This is because one really can't do psycholinguistics (never mind generative semantics or syntax) without, at a minimum, structures that take truth-values, and because associations aren't truth-apt, they cannot serve that role. It should strike the reader as odd that elsewhere in our cognitive theorizing we are happy to quantify over propositional structures, yet for some reason they have seemingly been banished in social psychology.

This oddity has metastasized in the literature, being reified in the proliferation of a certain type of dual-process model, the 'system 1'/ 'system 2' model originally popularized in Sloman (1996) and expanded in (e.g.,) Smith and Decoster (1999), Wilson et al. (2000), Evans (2003), Kahneman (2011), and Evans and Stanovich (2013). This type of dual-process model posits that 'system 1', or as Sloman helpfully calls it, "The Associative System" (p7) is fast, automatic, intuitive, arational, *unconscious*, and of course *associative*. System 1 is often additionally thought to be phylogenetically ancient and ontogenetically antecedent to system

2, the "Rule Based System", which is hypothesized to be involved in classic symbol manipulation of propositional structures, and is slow, logical, rational, and conscious.

The support for this type of dual-process theory has not only pervaded the implicit attitude literature (e.g., Gawronski & Bodenhausen 2006; Grumm et al. 2009) but has also grown outside of social psychology (see, e.g., Gendler 2008; Frankish 2010). Most damagingly, the popularity of this dual-system talk has made theorists assume that we can infer from any system 1 property to any other; thus if a certain process is automatic we can infer that it's arational (or 'heuristic'); if it's fast, then it must be automatic; and most importantly for current purposes, if it's unconscious, then it's associative.

A moral of this essay is that this set of inferences is unjustified. Contra dual-process theories, propositional structures—the putative hallmark of system 2 processes—can affect supposedly proprietary (and supposedly associative) system 1 processes. Moreover, propositional processes and structures not only affect unconscious states, but the propositional structures can *be* unconscious states and their corresponding logical processes can operate unconsciously. The types of structures needed in order to explain just about any of the effects canvassed, are unconscious propositional structures. These propositional structures often have causal roles that can mimic their inferential roles (that is, the entailment relations of its contents), thereby making them conducive to being reason-responsive—they are sensitive to information in ways that exhibit the structure of rationalizations and reasoning, and not just subject to the mere contingencies of ambient associations.

Those who champion these types of dual-process models rule out the possibility of unconscious, reason-responsive propositional structures. This is problematic, because even setting aside the arguments given here, we presuppose these structures elsewhere in cognitive science. For decades both linguistics and vision science have made progress by positing complex, structured, unconscious states. But somehow this progress hasn't bled over into much of social psychology. The situation is even more perplexing because social psychologists have been uncovering logical unconscious processes ever since at least the heyday of Festinger, and this tradition has continued through Dijksterhuis and the theorists cited above.[24] In fact, one of the must frequently used paradigms in social psychology, the misattribution paradigm, continually implicates unconscious logical processes. For instance, take Storms and Nisbett (1970). They gave insomniacs placebos that subjects were told had different effects. One group was given a placebo that its members were told was a stimulant, the other given one that was supposedly a depressant. Intuitively, one might have thought that the latter group would have an easier time going to sleep: after all, a stimulant should keep you awake longer and a depressant should make you sleep easier. However, the opposite finding was uncovered. Because of their chronic insomnia, subjects in both conditions were apt to feel heightened anxiety when going to bed. The subjects given the putative stimulant now had a cogent reason for their anxiety: they could attribute it to the pill and not to trying to fall asleep. In contrast, the group that was given the putative depressant had no such attributional base: since they were given something that was supposed to ease anxiety, they reasoned that their anxiety must have been wholly due to the sleep situation and their insomnia in fact increased.[25] This is complicated and impressive unconscious logic, and such logic appears to be involved in many, if not all, misattribution paradigms (for a similar reading of such logic in priming paradigms, see Loersch & Payne 2011). The best explanation we have of this type of complicated reasoning involves quantifying over unconscious propositional processes, processes that are amenable to reasoning. But to do so is to reject the current dual-process theory zeitgeist. By barring unconscious propositional reasoning we not only make a hash of the implicit bias literature, we also lose the only explanations we have of a host of important psychological research.

It's time to stop assuming that since a process is unconscious, or automatic, or fast, then it must be associative. Unconscious, automatic, and fast processes can be logical and can operate over propositional structures. The system 1/system 2 idiom (as opposed to, say, the language of modularity theory) invites these inferences to proceed without argument. But such inferences misrepresent the contours of our cognitive architecture and blind us to ways of changing the more uncouth parts of our cognitive underbelly, such as the structures underlying implicit bias.

The mind is a complicated system and different processes are arranged in a multitude of different ways. In particular, unconscious processes can be associative, but they can also be logical and can operate over propositional structures. There may be much room in cognitive science for dual-process theories, but there shouldn't be any room for the type of dual-process theory that infers from 'automatic' or 'unconscious' to 'associative.'

6. Explaining Attitudinal Divergence: Fragmentation and Redundant Representation

After focusing on all the negatives of dual-process theory, it is worth spending a moment to ponder why theorists are drawn to it in the first place. Despite all of its serious shortcomings, dual-process theories are tailor-made to explain an inconvenient fact: that our implicit attitudes and explicit attitudes are often misaligned. Part of what is so surprising about implicit bias is that, often enough, people profess to not have the corresponding explicit belief. The most extreme (and not widely held) versions of dual-process theories posit no causal interaction between propositional and associationist processes. By drawing a hard boundary, these theories can explain why we find attitudinal divergence: different types of attitudes are the products of different types of systems.

In rejecting dual-process theory, the Structured Belief view thus faces the uncomfortable question of why there appears to be such divergence between explicit and implicit attitudes. Why don't our implicit attitudes line up with our explicit ones (or for that matter, why don't our explicit ones line up with our implicit ones)? Since the Structured Belief view proposes that evidence can affect implicit attitudes one might expect an overall coherence in beliefs. And if the attitudes don't line up, then why think implicit attitudes are beliefs?

The idea that implicit attitudes have propositional structure but aren't belief-like is a live theoretical option, and has recently been proposed by Levy (2014). In closing, I'll say a bit about why I think Levy's criticisms are surmountable, and then explain why I think we see such attitudinal divergence.

First, a note on why I think implicit attitudes are beliefs. It's just for the simple reason that they appear to function as beliefs (as outlined in Mandelbaum 2014). In particular, they are inferentially promiscuous, interact with motivational states to cause behavior, and have the capacity to be sensitive to evidential considerations. The basic disagreement between Levy's view and the Structured Beliefs view concerns the functional profile of beliefs. In particular Levy thinks beliefs shouldn't control low-level behaviors and should control assertion. I take these criticisms in turn.

Levy argues that implicit attitudes control 'microbehaviors,' such as seating distance and hand touching, and he claims that these behaviors cannot be explained by propositional attitudes. Levy asks, "What belief might motivate these behaviors?" (p14). Although this is posed as a rhetorical question, I think it's answerable. If, for example, subjects unconsciously believe that African Americans are dangerous, then it is this belief (and its related inferences) that help control the subjects' microbehaviors. Subjects who believe African Americans to be dangerous and desire not to be in danger will, ceteris paribus, keep their distance from African Americans. In an experiment where seating distance is the dependent variable, positing this belief allows us to predict that the subjects will sit further from the African American confederate. Similar reasoning applies to the other cases Levy discusses.

Levy might be assuming that these so-called 'microbehaviors' have different psychological explanations than other mundane behaviors. But this separation is unmotivated. Microbehaviors (along with typical actions) look decidedly different than paradigmatically reflexive behaviors, such as the deep tendon reflex that controls knee jerks. One has to *decide* where to sit, and even if such decisions are unconscious, they are subject to decision-theoretic processes in a way knee-jerk reflexes aren't.

Levy also notes that people are inapt to assert their implicit attitudes (p16). But to assume that the lack of assertion makes them something other than beliefs is to build an intrinsic connection between belief and assertion that seems too strong, and inadequately responsive to empirical data (for discussion see Mandelbaum 2014).

But even if Levy's specific criticisms are surmountable, one is still left with the question: if implicit attitudes are beliefs, then why do we find such an attitudinal divergence? First, it is worth noting that claims of vast attitudinal divergence may be overblown. For instance, Payne et al. (2008) found that a good deal of the variance between explicit and implicit attitudes appeared to be due to the different modes of testing. The more structural fit between the tests, the closer the attitudes became. Nonetheless, Payne never uncovered anything close to perfect correlations between the attitudes, so it's safe to assume that there is some variance between explicit and implicit attitudes that is not due to mere differences in dependent variables. And that variance still needs to be explained, a task I won't attempt to do here. But there are two yet unmentioned factors dealing with the structure of central cognition and the storage of belief which can help ease the theoretical tension.

Contra the picture of central cognition offered in Fodor (1983), I'm inclined to think that central cognition is fragmented and contains redundant representations. By 'fragmented' I mean that some of our beliefs are causally isolated from other beliefs. The picture that emerges is one where we cannot, strictly speaking, talk of a person's single stock of beliefs. Rather, each believer will have multiple, synchronously encapsulated webs of belief, but no single overriding web as envisioned in Quine and Fodor.[26]

A fragmented system is one that can explain how we can harbor contradictory beliefs while still using beliefs to explain rational action (Lewis 1982, Egan 2008). One might believe that one is a terrible person in a given context, while believing that one is a good person in another. These beliefs might continue persisting, and in such a case there is no single answer for what believes about oneself simpliciter. As long as the beliefs are in their own causally isolated networks there is no simple fact of the matter what one believes.

A consequence of positing a fragmented view for dealing with contradictory beliefs is that we are apt to end up with redundant representations: multiple representations of the same information embedded in distinct belief networks. Add to this the ease with which new representations can be acquired and 'tagged' to the environment in which they were acquired and we end up with the following picture: people might have a large number of token beliefs that p, beliefs which are causally isolated from one another. And there's no a priori way to tell whether new information about p will be updated into an old network in which a token belief that p is embedded, or whether a new token of p is introduced into a new network.

This type of picture has in fact been offered by psychologists on both sides of the propositional/associationist divide. Here are proponents of a propositional picture, Mitchell et al, on multiple attitudes:

> [A]utomatic attitudes are defined within the context established by the situation. The appearance of stability or the existence of a single real attitude arises from the high consistency in environments that masks the fact the evaluations are continuously and actively being constructed against the backdrop of the current situation . . . the current findings cast doubt on the belief that there exist single, unitary attitudes awaiting authentic observation by implicit measures. With variation in context, multiple evaluations of an attitude object may be evoked, but none of those evaluations is more true than any other, even though some that are culturally privileged may be observed in the vacuum of the laboratory (2013, p467–8).

And here's Blair, a proponent of the dominant associationist view:

> Over the years of "failures" in attitude research, there have been periodic calls for the adoption of a more flexible, situation-specific definition of attitudes. As stated by Tesser (1978), "An attitude at a particular point in time is the result of a constructive process . . . .And, there is not a single attitude toward an object but, rather any number of attitudes depending on the number of schemas available for thinking about the objects" (2002, p. 256).

So if evidence can flip implicit attitudes, why don't aversive racists drop their implicitly biased attitudes? I think the answer is that they often enough do, but they are apt to have many different token representations, and some will have greater amounts of inferential connections than others. When trying to understand the updating of unconscious attitudes we have to take into account the amount of inferential connections that a given token attitudes has, and we have to take into account how many token beliefs with the same content we harbor. It's reasonable to suppose that the more important the content is, the more beliefs and connections it will have. To overturn implicit biases, it will be necessary to tackle all of these different representations, which will quite likely prove far more difficult than these laboratory demonstrations.[27]

Notes

[1] Small caps will be used throughout to denote structural descriptions of concepts. Thus BLACK MALE is taken to be a complex concept consisting of two meaningful parts: the concept BLACK and the concept MALE, which themselves are atomic. The structural descriptions are stipulated, but the stipulations will not affect the arguments in the text.

[2] Some theorists go so far as to refer to the structures underlying implicit bias as 'beliefs.' But the term does not carry either of its philosophical glosses: it refers to neither the pragmatic sense of belief, where beliefs are understood as behavioral dispositions (Dennett 1991) nor to the realist sense, where beliefs are understood as relations to propositionally structured mental representations (Fodor 1978; Mandelbaum 2013, 2014). Instead uses of 'belief' in the implicit attitude literature are rooted in AIB. For example, here is Blair (2002) on beliefs as associations: "The second definitional issue that must be addressed is the conventional distinction between stereotypes and prejudice, with the former referring to the beliefs (semantic associations) people have about social groups and the latter referring to their evaluations of groups" (p244). Likewise, Gregg et al. sound a similar theme: "Although [the] focus of our research is on automatic evaluative associations ('attitudes') many of the points we made are likely to apply equally to automatic semantic associations ('beliefs'). We use the catch-all term 'automatic attitude' to imply that our theorizing potentially straddles both types of association. (e.g., Banaji and Hardin, 1996)" (Gregg et al. 2006 p1).

[3] Retrieved from https://implicit.harvard.edu/implicit/demo/background/faqs.html#faq2, under the section "What is an 'implicit' attitude?", accessed on September 9, 2013.

[4] Constructivists, such as Karmiloff-Smith (1995), attempt to explain how some mental processes can be constructed through pre-existing mental processes. But this just pushes the problem back a level, because the pre-existing processes are themselves assumed, hence innate.

[5] Associationism has also been extended to cover and the implementation base of mental states (neural net implementation). Because of space restrictions, I will have next to nothing to say about debates over connectionism, except to remind the reader that the inference from 'associative implementation base' to 'associative mental process' is invalid. In principle, there is no reason that a classical propositional structure couldn't be housed in a connectionist network. One who thinks that there is a language of thought may also believe that brains are associative (I say 'may' but not 'must,' because one can take a position similar to Gallistel's and maintain that the implementational base of cognition is neurochemical, thus completely bypassing any level where nomological explanations involve pure associations, Gallistel & King 2009).

[6] Bootstrapping wouldn't be posited as a theory of mental structure either, though some might champion it as a causal determiner of cognitive structure (a two-factor theorist about conceptual content like Carey is in fact committed to such a picture).

[7] It being unwarranted sadly doesn't effect its prominence. For example, Olson and Fazio write, "Attitudes are thought to develop via classical conditioning through repeated pairings of potential attitude objects (conditioned stimuli, CSs) with positively and negatively valenced stimuli (unconditioned

stimuli, USs) and intuitively, one would expect this to be a ubiquitous means of attitude formation" (Olson and Fazio 2006, p413). Building on this picture in a later paper, Fazio moves from associative learning to associative structure: "In brief, the model views attitudes as associations between a given object and a given summary evaluation of the object-associations that can vary in strength and, hence, in their accessibility from memory" (Fazio 2007 p4).

[8] Because of space constraints, (a) will not be addressed at length, though see fn. 20.

[9] Of course, to say that Structured Beliefs can be reason-responsive is not to say that they are always responsive to reason, nor even that they are generally responsive to reason in normatively respectable ways. The extent and character of their reason responsiveness is an empirical question, though their ability to be reason responsive is relatively certain (for more on the reason responsiveness of unconscious beliefs, see fn. 22).

[10] A caveat: although I just used modus ponens in the above example there is no commitment to the idea that the logic of Structured Belief is isomorphic to any other well-known, reputable logical system. The logic of Structured Belief is, in essence, the logic of thought. I would tell you what this logic looks like, except I don't quite know what it looks like. Happily, I won't have to in order to argue for Structured Beliefs and against AIB. That said, a perusal of the Wason and Cheater Detection literature should make one reasonably certain that whatever the logic of thought is, it most probably isn't the first order logic taught to undergraduates (though it may end up sharing properties in common with something like the Mental Logic of Braine and O'Brien, see Lee et al. 1990).

[11] For example THE REPUBLICANS LOST THE ELECTION may carry a positive valence even if none of its constituents do. One can imagine a left-leaning person having a negative association attached to REPUBLICANS, LOSING, and ELECTIONS yet still generating a positive valence for the thought THE REPUBLICANS LOST THE ELECTION.

[12] Even though the belief might not be conscious, some of its constituents could still be. One might even have access to the contents of the belief but the fact that the structure is a belief might still be unconscious (note the similarity to Dretske 1993; 2004). This allows for the possibility that one might know, say, that one is thinking CATS ARE SWEET without knowing whether one believes or hopes that cats are sweet. There are enough degrees of freedom here to allow for one to buy into the Structured Belief hypothesis and still hold, as Loersch and Payne (2011) do, that people have some conscious access to their implicit attitudes.

[13] 'Free standing associations' are used in contrast with what I'll term 'piggybacking associations.' Structured Beliefs can create associations through the mere continued activation of the constituents of the beliefs. In other words, the concepts that make up structured beliefs can develop their own associations. If I often have the thought 10PM IS A GOOD TIME TO HAVE A BEER then, ceteris paribus, I'll start to associate 10PM and BEER. Likewise, the belief DOGS SLEEP ON TABLES has among its constituents DOGS and TABLES. Activating this thought will thus create a link between DOGS and TABLES, such that the activation of one concept might then facilitate the other, and the valence of one might evaluatively condition the other. Assume you have a positive valence assigned to DOGS and no particular valence assigned to TABLES. If you activate DOGS SLEEP ON TABLES enough then sooner or later the positive valence of DOGS will rub off on TABLES. These sorts of associations are what I'll call 'piggybacking associations.' Piggybacking associations supervene on the propositions that create them; take out the proposition, you'll take out the piggybackers. If one extinguishes the piggybacking association while leaving the proposition alone, then the association should be reinstated when the proposition is reactivated (for evidence of the ubiquity of reinstatement see Bouton 2002).

[14] It is important to be clear on what this conditional does and does not entail. What I am willing to suppose is that all dissonance effects are in fact produced by propositional structures partaking in unconscious reasoning. What this most surely does not entail is the conditional: if dissonance manipulations don't work, then we must not be dealing with propositional structures (i.e., the structures are associative). This is of particular importance because of the interesting work of Gawronski and Bodenhausen (see, e.g., Gawronski & Bodenhausen 2006). They have spilled much ink showing null effects of dissonance manipulations on implicit attitudes and then concluding that implicit attitudes are associative not propositional. Although I laud their efforts for at least discussing the near-invisible question of the structure of implicit attitudes, their inference is a non-sequitur. (For what it's worth I

suspect that dissonance manipulations would indeed work on implicit attitudes if we just had repeated dissonance exposures, but this line of evidence is for another time.)

[15] N.b., it's not that the inferences I'm making are part of folk psychology; rather, it's that the states underlying implicit attitudes look to be something 'spiritually similar' to folk psychological attitudes (Fodor 1987).

Perhaps it's also worth noting that this 'balance' effect of liking people who do bad things to people you dislike can be seen in 10-month old infants (Hamlin et al. 2011). This datum should stick in the craw of anyone who thinks that associative processes are ontogenetically prior to propositional ones (e.g., Gendler 2008).

[16] One may note that the second experiment in Garwonski et al. (2005) did not find balance theory effects but did find more AIB-friendly effects when presentation of the relation between CS1 and CS2 preceded the formation of a valenced attitude toward CS1 or CS2. But this datum is orthogonal to the discussion at hand. The main point is that balance theory effects exist and cannot be explained via AIB, whereas conditioning effects can be explained by a propositional theory (especially a 'piggybacking' one; see fn. 13). No doubt presentation of stimuli will affect the recall and salience properties which should also affect the sorts of learning on display. But none of these truisms address the main point in the text—how can AIB explain balance theory effects?

[17] Some recent data drive home the point even more acutely. Cone and Ferguson (forthcoming) conditioned subjects with either 100 positive or 100 negative statements about a fictional person. The 101st statement had the opposite valence of the previous 100. If the subjects took that last statement to be very diagnostic of personality (as seen on pre- and post-tests) then their implicit attitudes immediately matched the valence of the last statement, overriding the 101 previous conditioning events (which was all the subjects knew about the fictional person).

[18] A related effect was found in Maio et al. 2009 in regards to implicit attitudes pertaining to multiculturalism and racism. Subjects who had ambivalent attitudes towards multiculturalism increased their negative attitudes toward ethnic minorities after exposure to shoddy anti-racism arguments (e.g., "Young people generally rate their friendships with ethnic immigrants as 4% more satisfying than other friendships", p338) and slogans (e.g., "Multiculturalism = Prosperity + Progress") that were *in favor* of multiculturalism. Importantly, these arguments and slogans contained positively valenced words so an associative explanation cannot handle the datum. However, this datum makes sense if we assume that subjects infer that experimenters would present the best arguments they had for multiculturalism. Because the arguments given were shoddy, the subjects then infer that there aren't any good arguments for multiculturalism. Though AIB cannot handle such data the Structured Belief hypothesis is well positioned to explain this unconscious reasoning.

[19] To put into perspective just how intensive the evaluative conditioning was, consider that Gawronski et al. 2005 (discussed above) only needed three trials in order to solidify the evaluative conditioning, a routine amount of trials for an evaluative conditioning experiment, while here we are dealing with 240 trials.

[20] How frequently counterconditioning fails to work is a question of much interest. The documented successes of true counterconditioning (that is, changing a pre-existing valenced attitude, not a neutral or novel one) of implicit attitudes are rare (though for the successes, see Kawakami et al. 2000; Dasgupta & Greenwald 2001; Blair et al. 2001; Olson & Fazio 2006; Rydell et al. 2006). N.b., the claim of rarity holds over *counter*conditioning, not regular old conditioning of implicit attitudes (which isn't rare, but isn't necessarily caused by associative processes either, see, e.g., De Houwer 2009; Mitchell et al. 2009; Beckers 2006). To my eyes, the greatest successes have been in counterconditioning implicit self-esteem (Baccus et al. 2004, Gawronksi & LeBel 2008; Grumm et al. 2009). The successes of extinction are rarer still, though to be fair the data on extinction of evaluative associations in general are a bit ambiguous (De Houwer 2011). Were one to want to defend the efficacy of counterconditioning/extinction one would not only have to show that they are efficient techniques, but also show that successful extinction/counterconditioning effects have produced comparable curves to the normal learning and extinction curves found in classical conditioning, a risky gambit (e.g., Rescorla and Wagner 1972, Gallistel et al. 2004). Furthermore, if counterconditioning were effective then why aren't there gobs of documented instances of it? After all, making implicit attitudes disappear would be big news in academia and beyond. It is reasonable to suppose that the pragmatics of publication masks the inefficacy of countercondi-

tioning of implicit attitudes. Here's Gregg et al., mentioning the problem after reporting their own null effect: "If statistical significance is used as the sole criterion of malleability, then, by the conventional logic of hypothesis testing, only evidence for malleability can emerge, because the alternative would be a null result whose interpretation must remain equivocal" (2006, p15). In fact, Gregg et al. found associative extinction to be so lacking in efficacy, they exasperatedly write, "Concrete learning [i.e., associative extinction] did not emerge as a more effective means of undoing [implicit attitudes] than abstract supposition [logical intervention] did. It is remarkable that despite considerable theoretical precedent, concrete learning never exerted a larger impact on automatic attitudes than abstract supposition did" (p14).

[21] In Gregg et al. the bias didn't disappear completely, though it did recede. Jeremy Cone and Yarrow Dunham have replicated the main findings of the Gregg et al. and in their replication the supposition intervention didn't completely reverse the bias, though it did eliminate it (and counterconditioning did neither; personal communication).

[22] One might wonder if there are any generalizations to be had about the speed and normative respectability of Structured Belief updating. The rule of thumb seems to be that if the belief is one that has great value to a person, then the updating of that belief will happen gradually at best (and the updating won't be particularly rational, e.g., it won't be updated in any Bayesian fashion), whereas if the belief isn't one that a person self-identifies with, then it can sometimes be changed immediately in response to evidence (and this is the case even if one has a high credence in the belief). For more discussion see Mandelbaum 2014 or Levy and Mandelbaum 2014.

[23] There are scads of other data worth mentioning in a critique of AIB though space precludes me from doing so here. But the structure of the arguments given here should generate naturally on their own when applied to other work in the field, even when that work explicitly claims to be a pro-AIB study. For further experiments that sit poorly with AIB see Asgari et al. (2011); Dasgupta et al. (2009); and Maio et al. (2009). Additionally many studies in the implementation intentions literature, see, e.g., Stewart and Payne (2008), can also be easily interpreted as anti-AIB.

[24] See, for example, Dijksterhuis (2004). Note that my claim is very different from Dijksterhuis's. My argument doesn't depend on whether unconscious thought is in any sense 'better' or 'more efficient' than conscious thought; rather it's just that unconscious thought can be quite logical. For what it's worth, those who have failed to replicate Dijksterhuis (2004) (e.g., Thorsteinson & Withrow 2009) haven't found unconscious processes to be less logical (or less efficient) than conscious processes—they have just failed to find unconscious processes that are more efficient than conscious processes. What they've found instead is that unconscious processes are *as* efficient and logical as conscious ones, a datum that supports the current point in the text.

[25] Note that the effect wasn't just a between-group one: subjects in the placebo-stimulant condition went to sleep more quickly on night when they took the placebo than on nights when they didn't (and the same held, mutatis mutandis, for the placebo-depressant condition).

[26] I say 'synchronously encapsulated' because I don't think belief fragments are encapsulated in anything like a module. Rather, it is a porous form of encapsulation-at-a-moment. The principle behind it is that fragment 1 and fragment 2 are encapsulated from each other unless both fragments get activated at a time, in which case the fragments can merge.

[27] This essay has been long in production and over the years it has received much helpful criticism from a wide cast of people and audiences, not all of whom can be mentioned here. Particular thanks are due to audiences and seminar participants at Colgate, Columbia, Harvard, NYU, Wisconsin, Yale, the Jean Nicod Institut, the Institute for Advanced Study, Toulouse, and conference participants in Sheffield, Charleston, and Vancouver. I've also received extremely helpful feedback from Tim Bayne, Ned Block, Paul Bloom, Michael Brownstein, Brendon Dill, Fred Dretske, Yarrow Dunham, Jonathan Evans, Tamar Gendler, Bryce Huebner, Zoe Jenkin, Josh Knobe, Neil Levy, Shaun Nichols, Keith Payne, Jake Quilty-Dunn, Dave Ripley, Laurie Santos, Nick Shea, and Susanna Siegel.

References

Aronson, E., and Cope, C. (1968). My Enemy's Enemy is My Friend. *Journal of Personality and Social Psychology* 8 (1): 8–12.

Asgari, S., Dasgupta, N., and Stout, J. (2011). When Do Counterstereotypic Ingroup Members Inspire Versus Deflate? The Effect of Successful Professional Women on Young Women's Leadership Self-Concept. *Personality and Social Psychology Bulletin* 38 (3): 370–383.

Baccus, J., Baldwin, M., and Parker, D. (2004). "Increasing Implicit Self-Esteem Through Classical Conditioning." *Psychological Science* 15 (7): 498–502.

Banaji, M., and Hardin, C. (1996). Automatic Stereotyping. *Psychological Science* 7 (3), 136–141.

Bar-Anan Y., Nosek, B., and Vianello, M. (2009). The Sorting Paired Features Task: A Measure of Association Strengths. *Experimental Psychology*, 56, 329–43.

Beckers, T., Miller, R., De Houwer, J. and Urushihara, K. (2006). Reasoning Rats: Forward Blocking in Pavlovian Animal Conditioning Is Sensitive to Constraints of Causal Inference. *Journal of Experimental Psychology General* 135 (1): 92–102.

Blair, I., Ma, J., and Lenton, A. (2001). Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery. *Journal of Personality and Social Psychology* 81 (5): 828–841.

Blair, I. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, 6, 242–261.

Bodenhausen, G., Kramer, G., and Susser, K. (1994). Happiness and Stereotypic Thinking in Social Judgment. *Journal of Personality and Social Psychology*, 66 (4): 621–632.

Bouton, M. (2002). Context, Ambiguity, and Unlearning: Sources of Relapse after Behavioral Extinction. *Biological Psychiatry*, 52(10): 976–986.

Brinol, P., Petty, R., and McCaslin, M. (2009). Changing Attitudes on Implicit versus Explicit Measures: What is the Difference? In *Attitudes: Insights from the New Implicit Measures*, R. Petty, R. Fazio, and P. Brinol (Eds.). New York: Psychology Press.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Christensen, D. (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford: Oxford University Press.

Cone, J. and Ferguson, M. (Forthcoming). He Did *What*? The Role of Diagnosticity in Revising Implicit Evaluations. *Journal of Personality and Social Psychology*.

Correll, J., Park, B., Judd, C., and Wittenbrink, B. (2002). The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals. *Journal of Personality and Social Psychology* 83 (6): 1314–1329.

Cosmides, L., Barrett, H., and Tooby, J. (2010). Adaptive Specialization, Social Exchange and the Evolution of Human Intelligence. *Proceedings of the National Academy of Sciences*, 107 (2): 9007–9012.

Dasgupta, N., and Greenwald, A. (2001). On The Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals. *Journal of Personality and Social Psychology* 81 (5): 800–814.

Dasgupta, N. (2009). Mechanisms Underlying Malleability of Implicit Prejudice and Stereotypes: The Role of Automaticity versus Cognitive Control. In T. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination*. Mahwah, NJ: Erlbaum.

De Houwer, J., Thomas, S., Baeyens, F. (2001). Association Learning of Likes and Dislikes: A Review of 25 Years of Research on Human Evaluative Conditioning. *Psychological Bulletin* 127 (6): 853–869.

De Houwer, J. (2009). The Propositional Approach to Associative Learning as an Alternative for Association Formation models. *Learning and Behavior* 37 (1): 1–20.

De Houwer, J. (2011). Evaluative Conditioning: A Review of Functional Knowledge and Mental Process Theories, In T. Schachtman and S. Reilly (Eds.), *Associative Learning and Conditioning Theory: Human and Non-Human Applications*. NY: Oxford University Press.

Dennett, D. (1991). Real Patterns. *Journal of Philosophy* 88 (1): 27–51.

Dijksterhuis, A. (2004). Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making. *Journal of Personality and Social Psychology* 87 (5): 586–598.

Dretske, F. (1993). Conscious Experience. *Mind* 102 (406): 263–283.

Dretske, F. (2004). Knowing What You Think Vs. Knowing That You Think It. In R. Schantz (Ed.) *The Externalist Challenge*. Berlin: De Gruyter.

Elliot, A., and Devine, P. (1994). On the Motivational Nature of Cognitive Dissonance: Dissonance as Psychological Discomfort. *Journal of Personality and Social Psychology* 67(3): 382–394.

Egan, A. (2008). Seeing and Believing: Perception, Belief Formation and the Divided Mind. *Philosophical Studies*, 140(1): 47–63.

Evans, J. (2003). In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Science* 7 (10): 454–459.

Evans, J. and Stanovich, K. E. (2013). Dual-process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8, 223–241, 263–271.

Fazio, R. and Olson, M. (2003). Implicit Measure in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*, 54: 297–327.

Fazio, R. H. (2007). Attitudes as Object-evaluation Associations of Varying Strength.. *Social Cognition*, 25 (5), 603.

Festinger, L, Riecken, H., and Schacter S. (1956). *When Prophecy Fails*. MN: University of Minnesota Press.

Fodor, J. (1978). Propositional Attitudes. *The Monist*, 61, 501–523.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, Mass: MIT Press.

Fodor, (1987). *Psychosemantics* Oxford: Oxford University Press.

Ferguson, M., Mann, T, and Wojnowicz, M. (2014). Rethinking Duality: Criticisms and Ways Forward. In J. Sherman, B. Gawronski, and Y. Trope (Eds.), *dual-process theories of the social mind* (pp. 578–594). NY: Guilford Press.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass* 5 (10): 914–26.

Gallistel, C., Fairhurst, S., and Balsam, P. (2004). The Learning Curve: Implications of a Quantitative Analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36): 13124–13131.

Gallistel, R. and King, P. (2009). Memory and The Computational Brain: Why Cognitive Science Will Transform Neuroscience. *New York*: *Wiley/Blackwell*

Gawronski, B., Walther, E., and Blank, H. (2005). Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information. *Journal of Experimental Social Psychology*, 41, 618–626.

Gawronski, B., and Bodenhausen, G. (2006). Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change. *Psychological Bulletin*, 132, 692–731.

Gawronski, B., and LeBel, E. (2008). Understanding Patterns of Attitude Change: When Implicit Measures Show Change, but Explicit Measures Do Not. *Journal of Experimental Social Psychology*, 44 (5): 1355–1361.

Gendler, T. (2008). Alief and Belief. *Journal of Philosophy* 105 (10): 634–663.

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 (6): 1464–1480.

Greenwald, A., and Nosek, B. (2009). Attitudinal Dissociation: What Does It Mean? In *Attitudes: Insights from the New Implicit Measures*, R. Petty, R. Fazio, and P. Brinol (Eds.). New York: Psychology Press.

Gregg, A., Seibt, B., and Banaji, M. (2006). Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences. *Journal of Personality and Social Psychology* 90 (1): 1–20.

Grumm, M., Neslter, S., and von Collani, G. (2009). Changing Explicit and Implicit Attitudes: The Case of Self-Esteem. *Journal of Experimental Social Psychology*, 45: 327–335.

Hamlin, J., Wynn, K., Bloom, P., and Mahajan, N. (2011). How Infants and Toddlers React to Antisocial Others. *Proceedings of the National Academy of Sciences* 108 (50): 19931–199316.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Karmiloff-Smith, A. (1995). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, Mass.: MIT Press/Bradford Books.

Katz, I. and Hass, R. (1988). Racial Ambivalence and American Value Conflict: Correlational and Priming Studies of Dual Cognitive Structures. *Journal of Personality and Social Psychology* 55 (6): 893–905.

Kawakami, K., Dovidio, J., Moll, J., Hemsen, S., and Russin, A. (2000). Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation. *Journal of Personality and Social Psychology* 78 (5): 871–888.

Lebrecht, S., Pierce, L., Tarr, M. and Tanaka, J. (2009). Perceptual Other-Race Training Reduces Implicit Racial Bias. *PLoS ONE* 4(1): 4215.

Lebrecht, S. Bar, M., Feldman-Barrett, L., and Tarr, M. (2012). Micro-Valences: Affective Valence in "Neutral" Everyday Objects. *Frontiers in Perceptual Science* 3, (107): 1–5.

Lea, R., O'Brien, D., Fisch, S, Noveck, I., and Braine, M. (1990). Predicting Propositional Logic Inferences in Text Comprehension. *Journal of Memory and Language*, 29(3), 361–387.

Levy, N. (2014). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous*.

Levy, N., and Mandelbaum, E. (2014). In J. Matheson and R. Vitz (Eds.) The Powers that Bind: Doxastic Voluntarism and Epistemic Obligation in *The Ethics of Belief*. pp. 15–32.

Lewis, D. (1982). Logic for Equivocators. *Noûs*, 16 (3): 431–441.

Loersch, C. and Payne, B. (2011). The Situated Inference Model: An Integrative Account of the Effects of Primes on Perception, Behavior, and Motivation. *Perspectives on Psychological Science* 6 (3): 234–252.

Logue, A., Ophir, I., and Strauss, K. (1981). The Acquisition of Taste Aversion in Humans. *Behavioral Research and Therapy*, 19, 319–333.

Lowery, B., Hardin, C., and Sinclair, S. (2001). Social Influence Effects on Automatic Racial Prejudice. *Journal of Personality and Social Psychology*, 81 (5): 842–855. http://psycnet.apa.org/journals/psp/81/5/842/

Macrae, C., Bodenhausen, G., Milne, A., and Jetten, J. (1994). Out of Mind but Back in Sight: Stereotypes on The Rebound. *Journal of Personality and Social Psychology* 67 (5): 808–817.

Maio, G., Haddock, G., Watt, S., and Hewstone, M. (2009). Implicit Measures in Applied Contexts: An Illustrative Examination of Antiracism Advertising. In *Attitudes: Insights from the New Implicit Measures*, R. Petty, R. Fazio, and P. Brinol (Eds.). New York: Psychology Press.

Mandelbaum, E. (2013). Against Alief. *Philosophical Studies* 165 (1): 197–211.

Mandelbaum, E. (2014). Thinking is Believing. *Inquiry* 57 (1): 55–96.

McConnell, A. and Leibold, J. (2001). Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology*, 37 (5): 435–442.

Mitchell, C., De Houwer, J., and Lovibond, P. (2009). The Propositional Nature of Human Associative Learning. *Behavioral and Brain Sciences*, 32 (2): 183–246.

Mitchell, J. P., Nosek, B. A., and Banaji, M. R. (2003). Contextual Variations in Implicit Evaluation. *Journal of Experimental Psychology: General*, 132 (3), 455.

Nosek, B., and Banaji, M. (2001). The Go/No-Go Association Task. *Social Cognition* 19 (6): 625–666.

Nosek, B. Greenwald, A. and Banaji, M. (2007a) The Implicit Association Test at Age 7: A Methodological and Conceptual Review, in J. Bargh (Ed.) *Automatic Processes in Social Thinking and Behaviour*, Psychology Press, pp.265–292

Nosek, B., Smyth, F., Hansen, J., Devos, T., Lindner, N., Ratliff, K., Smith, C., Olson, K., Chugh, D., Greenwald, A., and Banaji, M. (2007b). Pervasiveness and Correlates of Implicit Attitudes and Stereotypes. *European Review of Social Psychology*, 18, 36–88.

Olson, M. and Fazio, R. (2006). Reducing Automatically-Activated Racial Prejudice through Implicit Evaluative Conditioning. *Personality and Social Psychology Bulletin*, 32 (4): 421–433.

Payne, B. (2001). Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon. *Journal of Personality and Social Psychology* 81 (2): 181–192.

Payne, B. (2009). Attitude Misattribution: Implications for Attitude Measurement and the Implicit-Explicit Relationship. In A. Black and W. Prokasy (Eds.) R. Petty, R. Fazio, and P. Briñol (Eds.), *Attitudes: Insights from the new wave of implicit measures*. Hillsdale, NJ: Erlbaum.

Payne, B., Burkley, M, and Stokes, M. (2008). Why Do Implicit and Explicit Attitude Tests diverge? The Role of Structural Fit. *Journal of Personality and Social Psychology*, 94(1): 16–31.

Rescorla, R. and Wagner A. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. Black and W. Prokasy (Eds.) *Classical Conditioning II: Current Research and Theory* New York: Appleton Century Crofts, pp. 64–99.

Rydell, R. and McConnell, A. (2006). Understanding Implicit and Explicit Attitude Change: A Systems of Reasoning Analysis. *Journal of Personality and Social Psychology* 91 (6): 995–1008.

Rydell, R., McConnell, A., Mackie, D., and Strain, L. (2006). Of Two Minds: Forming and Changing Valence-Inconsistent Implicit and Explicit Attitudes. *Psychological Science* 17 (11): 954–958.

Sechrist, G. and Stangor, C. (2001). Perceived Consensus Influences Intergroup Behavior and Stereotype Accessibility. *Journal of Personality and Social Psychology* 80 (4): 645–654.

Sloman, S. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, 119, 3–22.

Smith E., and DeCoster, J. (1999). Associative and Rule-Based Processing: A Connectionist Interpretation of Dual-Process Models. In S. Chaiken and Y. Trope (Eds.) *Dual-Process Theories in Social Psychology*. New York: Guilford Press. pp. 323–336.

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 (10): 1332–1345.

Storms, M., and Nisbett, R. (1970). Insomnia and the Attribution Process. *Journal of Personality and Social Psychology* 16 (2): 319–328.

Tesser, A. (1978). Self-Generated Attitude Change. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* Vol. 11. NY: Academic Press. 289–338.

Thorsteinson, T. and Withrow, S. (2009). Does Unconscious Thought Outperform Conscious Thought on Complex Decision? A Further Examination. *Judgment and Decision Making* 4 (3): 235–47.

Velleman, D. (2000). On the Aim of Belief. In *The Possibility of Practical Reason*. Oxford: Clarendon Press. pp. 244–281.

Walther, E. (2002). Guilty by Mere Association: Evaluative Conditioning and The Spreading Attitude Effect. *Journal of Personality and Social Psychology* 82 (6): 919–934.

Wilson, T., Lindsey, S., and Schooler, T. (2000). A Model of Dual Attitudes. *Psychological Review* 107 (1): 101–126.