8-2019

Implicit Bias and the Boundaries of Belief: A Single-Representational Dual-Attitude Account of Implicit Attitudes

Austin Dakota Synoground
University of Arkansas, Fayetteville

Implicit Bias and the Boundaries of Belief: A Single-Representational Dual-Attitude Account of Implicit Attitudes

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Philosophy

by

Austin Synoground
University of Arkansas
Bachelor of Arts in Philosophy, 2017

August 2019
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Eric Funkhouser, Ph.D.
Thesis Director

_____
Jack Lyons, Ph.D.
Committee Member

_____
Ed Minar, Ph.D.
Committee Member

Abstract

Since their inception, implicit attitudes have been defined as associative mental states, separate from beliefs, which are considered to be propositional in nature. Recently, several philosophers have challenged this distinction, arguing that implicit attitudes are actually unconscious beliefs. In turn, I argue that the attitudes detected by current experimental paradigms are blind to distinctions between implicit attitudes, which I define as the products of an associative learning mechanism, and unconscious beliefs, which are the products of a propositional learning mechanism. Specifically, I argue for a single-representational dual-attitude account of implicit bias.

Table of Contents

Section 1. Introduction

The term 'implicit bias' was first introduced in the psychological literature in 1995 by social psychologists Mahzarin Banaji and Tony Greenwald (Banaji & Greenwald, 1995). Originally focused on implicit social cognition, the term has come to refer to a host of unconscious processes capable of facilitating downstream effects on judgment and behavior. A recent compendium of implicit bias defines the phenomenon as the "relatively unconscious and relatively automatic features of prejudiced judgment and social behavior."[1] For example, a hiring manager with a commitment to meritocracy may nevertheless, upon inspection, have a history of hiring only those applicants with *Western* names. Interviewees with appellations outside this category (Taj, Allegra, Kahdijah) are at a decided disadvantage for employment, though this fact may be lost on both themselves and their interviewer. In this case, it might be said that the interviewer has an *implicit bias* towards foreigners; or, alternatively, that the interviewer possesses an *implicit attitude* against foreigners. While often used interchangeably, the former term is preferred when stressing agent-level actions and the latter when indicating the mental content responsible – either in whole or in part – for a biased action. Notwithstanding the most ardent dispositionalist, this subtle distinction ought to be acceptable to most philosophers, regardless of their doxastic commitments. Moving forward, I will use *implicit bias* when speaking of the actions and consequences of individual/group behavior, and *implicit attitude* when discussing the representational state of a mental item implicated in prejudiced actions.

[1] See Michael Brownstein, "Implicit Bias," in Stanford Encyclopedia of Philosophy, ed. Edward N. Zalta (https://plato.stanford.educ/archives/spr2017/entries/implicit-bias/).

With this distinction underfoot, we are free to ask: why should philosophers be interested in implicit bias? First, the harms of implicit biases are self-evident. For instance, Hoffman and colleagues found that black Americans are undertreated for pain when compared to their white counterparts (Hoffman et al., 2016). Faculty members from psychology departments around the country – which are chiefly liberal (Buss & von Hippel, 2018) – rate the same CV as stronger when it is accompanied by a typical male, rather than female, name (Steinpries, Anders, and Ritzke, 1999). While inadequate pain treatment and employment challenges are concerning in themselves, the infamous 'Weapon Bias' study reveals that subjects are much more likely to perceive someone as carrying a weapon when they are black as opposed to white (Payne et al., 2001). Even more disturbing, participants in a fast-paced virtual simulation are much more apt to 'shoot' an armed subject if he is perceived as black, and contra positively, participants are much quicker to 'not shoot' a subject perceived as white (Correll et al., 2002). The continued shootings of unarmed black men by police in America underscores the fact that there is, unfortunately, a very real connection between how a person behaves in an experimental setting and their behavior in real-world situations (Banaji and Greenwald, 2013).

In addition to these cases, where the harm to targets of implicit bias is readily apparent, there are many subtle ways in which discriminatory behavior can flourish without attracting attention. This is most clearly evinced in cases of microaggression, where the verbal and behavioral interactions between persons is mediated by certain social facts (e.g. race, gender, age, weight, and so on), but whose harm is masked by the inconspicuous manner in which it unfolds. Consider the elderly white woman who clutches her purse as she passes a black man (Sue et al., 2007), the male Supreme Court member who incessantly interrupts his female

colleague (Jacobi & Schweers, 2017), or skeptical questions from students about their racialized instructor's ability to teach English (Ramjattan, 2019). The prevalence of these subtle discriminations can continually chip away at the well-being and self-esteem of individuals while operating under the guise of conventional behavior and accepted social norms, ultimately culminating in a 'death by a thousand cuts.' No wonder, then, that victims of microaggressions are liable to lash out a perceived slight or unwitting refrain, much to the consternation of the individual whose action is by all appearances (except the target's) a minor infraction.

Secondly, social psychologists have provided ample evidence that people can sometimes behave in ways that challenge their expressed beliefs and preferences, as in the case of our 'meritocratic' interviewer. This divergence between word and action is so pronounced that even staunch egalitarians can behave in ways diametrically opposed to their avowed commitments to social equality and fairness (De Houwer et al., 2009). This disparity between the conscious-level beliefs of an agent (including hypothetical beliefs, like what they would do in situation *X*) and their actual behavior is troubling from both an epistemic and moral perspective.[2] The actions of an agent who professes to believe that all people are to be treated equally without regard to feature *F*, yet behaves in a systematically biased manner towards people with *F*, is susceptible to questions regarding their sincerity, rationality, and moral acumen.

To help make sense of these inconsistencies, psychologists have coined the terms 'implicit' and 'explicit' to refer to the functional and structural differences between attitudes. While there are many usages of the term *attitude*, the ABC model of attitudes (Affect, Behavior,

[2] For discussion of the epistemic threat of implicit biases, see: (Gendler, 2011; Saul, 2013a, 2013b; Peters, 2018; Puddifoot, 2017).

and Cognition) has gained prominence in recent years (Solomon, 2008).[3] This view states that every attitude will have an affective, behavioral, and cognitive component, which can cohere or conflict to varying degrees; this is seen as an improvement over the classic model proposed by Rosenburg and Hovland (1960), which required each portion of an attitude to scrupulously align. According to the ABC model, the affective component of an attitude exemplifies one's feelings or emotions towards an object – e.g. I am afraid of snakes. The behavioral aspect of the theory details the effects our attitudes have on our actions – e.g. I will flee if I see a snake.[4] And finally, the cognitive component of an attitude is determined by an agent's knowledge or beliefs about an object – e.g. I believe snakes are a threat.

The dominant view within psychology holds that implicit attitudes are unconscious, associative, and acquired within the context of an agent's particular learning history (Levy, 2015, p. 803). These associations are the result of repeated pairings between a representation and a particular evaluative response (good, bad, deceitful, polite, etc.). For instance, continued exposure to hateful rhetoric and sensational media can create an association between the concept IMMIGRANT and the negative evaluations (bad, dangerous, etc.) used to describe or reference migrant groups. This can manifest in a host of biased behaviors, from decreased eye contact (Dovidio et al., 1997) to an inability to find adequate housing (Ahmed and Hammarstedht, 2008) and employment (Bendick et al., 2010). Some go further and distinguish implicit cognitive attitudes, which are propositionally structured mental representations, from implicit affective

[3] For instance, in Chapter XI of *Psychological Types* (1932), Carl Jung broadly defines an attitude as 'a readiness of the psyche to act or react in a certain way'.
[4] Knowledge of what one will do when confronted with a stimulus or state of affairs is not necessary for the behavioral component of the ABC theory. What is necessary for an action or response to qualify as behavior is a relatively stable reaction to stimuli under similar conditions.

attitudes, which are stereotypical input-output relations realized in one's valuational mechanisms (Carruthers, 2018).

The most famous method for measuring unconscious bias is the implicit association test; henceforth, IAT (Greenwald et al., 1998).[5] The IAT measures response times of participants when pairing images with words. The terminology chosen for these tasks are affectively-laden and/or can be used to form a stereotypical coupling (ASIAN + INTELLIGENT). For example, in a standard race IAT, a subject is presented with an image of a white or black face and asked to sort positively and negatively valanced words with the image. Someone with an implicit bias will perform better when the task demand accords with stereotypical classifications and slower when asked to classify pairings that defy those stereotypes. The difference in response times and accuracy between stereotypical and non-stereotypical pairings is due to the effort it takes to inhibit the automatically generated response that *this concept* goes with *this word*. Someone without an implicit bias will exhibit similar speed and accuracy regardless of whether the task demand accords with social stereotypes or not.

Explicit attitudes refer to the mental states an agent is consciously aware of. These attitudes are typically viewed as propositional in nature, making them sensitive to evidence and logical relations, as well as the constituents of conscious thought. The standard approach for identifying a subject's explicit attitudes are via verbal reports. Though vulnerable to regulated

[5] Other methods include the Affect Misattribution Procedure (AMP; Payne & Lundberg 2014); the Sorting Paired Feature Task (SPFT; Bar-Anan, et al. 2009); and the Weapon Identification Task (Payne 2001). For our purposes, we will focus on IAT experiments because they have been the most heavily researched.

responses and self-deception, psychologists are careful to control for these variables when designing and administering their experiments.

Surprisingly, some research suggests that implicit attitudes are a better predictor of behavior than explicit attitudes, reinforcing the aforementioned epistemic threat (Greenwald et al., 2009). Other studies question this assertion, with one major meta-analysis reporting little correlation between implicitly biased behavior within the lab and actions without (Oswald et al., 2013). Neil Levy (2015) questions the efficacy of these findings, citing various meta-analyses currently underway with preliminary reports showing stronger correlations than what Oswald et al. found in their experimental matrix. Nevertheless, even if implicit bias turns out to be a rare occurrence and the claims of Greenwald and colleagues unwarranted, the fact remains that a single instance of implicit bias has the potential to drastically alter, or end, a person's life.

The implicit/explicit distinction remains the dominant view in psychology and has been warmly received by philosophers in favor of dual process theories of mind (DPT). Dual process theory posits two different modes of cognitive processing: 'Type 1' and 'Type 2' (e.g., Evans and Stanovich, 2013) or alternatively, 'System 1' and 'System 2' (e.g., Frankish, 2010; Kahneman, 2012; Sloman, 2014).[6] The functional differences between these two groups of processes is often represented using contrastive pairings, such as slow/fast, effortless/effortful, evolutionarily old/evolutionarily recent, and so on. While I formally endorse DPT as the best framework in which to examine implicit attitudes, it is likely that many of the sharp distinctions

[6] System, here, refers not to a single cognitive mechanism but a diverse range of cognitive processes with similar functional traits. For instance, the mechanisms which compose our auditory and visual processes are specialized networks with little, if any, overlap; nevertheless, each would be classified as a System 1 process in virtue of being automatically activated and introspectively opaque.

that such pairings suggest are more fluid in nature. Indeed, some of the studies we will discuss in this work point to sophisticated interactions between Type 1 and Type 2 processes, a fact which ought to give pause to those that wish to divide the mind into neat categories. Following Nick Byrd, I will refer to these two cognitive systems as Non-reflective and Reflective (Byrd, 2018). One motivation for adopting these monikers in an already terminologically-laden field is that should future research challenge some of the sharp distinctions that typify DPT, the proposed division can be maintained by appealing to either system's relation to higher order cognition. For instance, in response to recent criticism, several philosophers have provided persuasive evidence of a perception/cognition border (Burge, 2010; Block, 2014; Firestone and Scholl, 2014; Mandelbaum, 2018); if substantiated, such a border indicates a delineation between conscious thought (*Reflective*) and those processes that operate automatically, are grounded in modular systems, and are relatively encapsulated from top-down influences (*Non-Reflective*). A recurrent question in this work is whether the operations of a mental representation are due to architectural/structural limitations or extrinsic factors, such as interactions with other mental contents. By the end, I hope to show that an appeal to structural limitations can help us delineate between the functional capacities of implicit attitudes and doxastic states.

Moving forward, I adopt Sophie Stammers' four couplets to frame my discussion of how DPT relates to conventional distinctions between implicit and explicit attitudes (Stammers, 2017). I align these couplets under the Non-Reflective/Reflective distinction to keep in mind the relation these processes have to higher-order thought:

Non-Reflective	Reflective
(a) Unconscious	(a^1) Conscious
(b) Associatively Structured	(b^1) Propositionally Structured
(c) Automatically Activated	(c^1) Deliberately Controlled
(d) Not avowed	(d^1) Avowed

Figure 1. Four distinctive features of dual-process theory

Recall our 'meritocratic' interviewer. Her belief in the ideal that 'The best person gets the job" meets propositions $a^1 \ldots d^1$, but her behavior is inconsistent with this belief, and operates in an unconscious and systematically biased manner towards a certain group of people – those with *non-Western* names. Here, the dual-process theorist posits that the output of the interviewer's non-reflective processes, whatever they may be, are in conflict with the outputs of her reflective processes.[7] Given that the woman's bias is (a) unconscious, (b) acquired in an associative manner, (c) activated automatically and immune to top-down suppression, and (d) consciously disavowed, then the bias is implicit in nature, caused via non-reflective processes, and in contention with the woman's reflectively endorsed attitudes. Endorsing a dual-attitudinal account of mind equips us with the structural framework needed to make recurrent clashes between attitudes (and other mental representations) intelligible: there are two functionally distinct systems whose outputs conflict along several cognitive dimensions, only one of which enjoys access to conscious awareness.

The outline of this work is as follows. Section two has four aims: (1) introduce the dual-mechanism theory of implicit bias, (2) clarify the role of top-down attention in the formation of

[7] A caveat is in order. Most of our explicit and implicit attitudes align, as in the case of a vitriolic vegan possessing both: a) explicitly negative beliefs and attitudes about meat-eaters, and b) implicitly negative attitudes against meat-eaters and their ilk. The interesting cases are when the two diverge, as when a social egalitarian learns they are implicitly biased against a certain social group.

subliminal associations, (3) present a leading theory of consciousness that helps establish the functional profile of implicit/explicit attitudes, and (4) defend the distinction between implicit attitudes and unconscious beliefs. Section three assesses several of the leading implicit attitude studies through the lens of dual-mechanism theory, and argues against the idea that implicit attitudes might constitute a unique mental state distinct from associations or beliefs. Section four examines the implications that a resurgence in associative research has for discussions of implicit bias. And section five concludes this work with a suggestion for future empirical studies into the nature of implicit attitudes, as well as applications for dual-mechanism theory.

Section 2. Dual-Processes and Multiple Learning Mechanisms

There are currently four attitude models that employ the terms *implicit* and *explicit* in different ways. Because these models vary in terminology and conceptual commitments, philosophers writing on implicit attitudes ought to be clear as to which framework they adopt and the reasons motivating their endorsement. First, some have argued that the two terms track distinct mental representations (Greenwald & Banaji, 1995; Wilson et al., 2000). Alternatively, some use the distinction *solely* in reference to attitude measurements, thereby remaining agnostic towards the representational status of any underlying mental state (Fazio, 2007; Petty et al., 2009). A third approach distinguishes attitudes via the processes they feature in, and admits of single, dual, and multi-process views (De Houwer et al., 2009). And finally, the terms have been used to capture significant differences in a person's evaluative responses (Gawronski & Bodenhausen, 2011).

In addition to this dispute, there is disagreement over what mental items the association/proposition distinction is meant to track. They can refer, alternatively, to the *learning mechanism* that encodes an attitude in memory, the resulting *mental representation*, or the *processes* by which an attitude manifests in behavior (Gawronski et al., 2017). Endorsing the second psychological model and referring to associative or propositional processes would shift the discussion of implicit biases into the world of behavior – welcome territory for those wishing to address the problems of implicit bias without endorsing particular claims about the representational or functional status of attitudes. Such a maneuver would look very different from someone who championed the first attitude model and the third associative/propositional pair – a representationalist view of belief with specific cognitive commitments – whose work would primarily focus on providing empirical evidence of distinct mental representations and separate memory stores.

I endorse a dual-process version of the third attitude model and use the terms *associative* and *propositional* to refer to two distinct learning mechanisms within the brain. Specifically, I present a single-representational dual-attitude account, which holds that all attitudes are similarly structured and stored in a shared memory format, but are disposed to feature in either automatic or deliberate processes. As we shall see, the rigid distinctions imposed by traditional versions of DPT are flaunted by evidence of propositional attitudes featuring in automatic processes. Findings such as these encourage a soft version of DPT, where the processes and mechanisms which have historically been associated with one categorical attitude can interact under certain circumstances. To justify these claims, I now turn to a recent meta-analysis that assessed the

performance of the four attitude models in relation to the last thirty years of evaluative conditioning (EC) research (Corneille & Stahl, 2019).

Evaluative conditioning can be understood as the evaluation of a conditioned stimulus (CS) due to its pairing with a positive or negative unconditioned stimulus (US) (De Houwer, 2007). A good example is the picture-picture paradigm, which repeatedly pairs a subjectively neutral picture of a human face (the CS) with a subjectively liked or disliked face (the US) until the affective qualities of the latter bind to the former (Baeyens et al., 1992). Understanding the state of affairs within EC research is important, since it provides "…the strongest support for the existence of an associative attitude learning process…" (Corneille & Stahl, 2019, p. 162).

The results of their meta-analysis found that no single attitude model could address the diversity of complex and sometimes contradictory findings within EC research, however, a rendition of the third model – De Houwer's propositional approach to attitude learning (PAL) (De Houwer, 2009) – faired best. According to this theory, all instances of associative conditioning are best explained by the formation of propositional mental constructs, an idea which undercuts over a century of associative research.[8] For De Houwer, a proposition about a stimulus relation is best understood as "…a mental representation that contains information about the nature of the relation between stimuli (e.g. A *predicts* B, A *causes* B, A *co-occurs with* B) (De Houwer, 2018, p. 3). As such, there are no 'associations' in the normal sense of the word. Because propositions can record mere co-occurrences as well as more specific relational information, the need for an associative learning mechanism or other such mental process is

[8] This is an application of Mitchell and colleagues' (2009) broader claim that all instances of learning in humans is propositional in nature.

rendered obsolete. This makes PAL a single-process account of attitudes, where the process responsible for producing all attitudes in human cognition is a single propositional learning mechanism. Unfortunately, there is currently no method for determining whether or not an attitude possesses a purely associative structure, as most psychologists contend, or a propositionally structured recording of a mere contiguous relation.[9] As such, purveyors of the field are tasked with assessing how well an attitude model conforms to empirical findings in a parsimonious manner.

Another central claim of PAL is that in order for specific relational information to be recorded, an agent must be consciously aware of the subsequent stimuli, events, or concepts involved. More precisely, "…it is assumed that a relation in the world can influence behavior only after a proposition about that relation has been consciously entertained as being true." (De Houwer, 2018, p. 6). The necessity of conscious awareness in the formation of propositional structures is supported by evidence of single-instruction attitude formation (Gast & De Houwer, 2013; Smith et al., 2013) and the fact that implicit evaluation is moderated by relational information, as evinced by implicit evaluation change in participants given affective-laden descriptions of strangers (Peters & Gawronski, 2011). Moreover, there is strong evidence that propositional structures can be formed and stored in memory at dizzying speeds (De Houwer, 2014), which ought to comfort those who think awareness is too severe a limitation to account

[9] Problems of falsifiability are a common trend amongst attitudinal models. Both De Houwer's propositional approach to attitude learning (PAL) and Gawronski & Bodenhausen's associative-propositional evaluation (APE) model cannot be falsified by current psychological paradigms (Gawronski & Bodenhausen, 2018; De Houwer, 2018). This is in large part due to the difficulty of discerning whether *A co-occurs with B* is merely a contiguous relation between stimuli or the propositionally structured recording of a contiguous relation.

for the rapid construction of attitudes.[10] As it stands, PAL offers doxastic accounts of implicit bias a sound theoretical model for why implicit attitudes are unconscious beliefs: simply put, all attitudes are the result of a single-propositional mechanism (De Houwer, 2014).

Moving forward, I provide evidence against a single-process account of attitudes and address some of the concerns of Corneille and Stahl's (2019) meta-analysis. In doing so, I argue that there are two separate learning mechanisms in the brain: an associative/non-propositional learning mechanism and one that is propositional in nature. Associative mental states have traditionally been cast as the internal representation of an external contiguous relation, thereby accounting for their insensitivity to logical relations and inferential patterns. Attitudes resulting from a propositional learning mechanism, in contrast, are encoded with precise relational information over their relata, and exhibit an unparalleled ability to commingle with other mental states, feature in inferences, and operate in accordance with the laws of logic. The next section articulates one way in which associative mental links can operate beyond mere contiguity, despite upholding a sharp distinction between the inferential promiscuity of associative and propositional states.

2.1 The Role of Top-Down Attention in Associative Conditioning

PAL states that all propositional learning, including associative conditioning, requires conscious awareness (Hughes et al., 2011). For information to be stored in a propositional

[10] This is one such example of the fluidity of traditional DPT pairings. Propositional structures are usually associated with deliberate thought, and the fact that they can feature in – in fact, even be formed by – automatic processes threatens rigid distinctions between Type 1 and Type 2 processes.

format, the higher-order processes which accompany conscious thought must identify the relation that holds over stimuli and concepts.[11] This strong requirement creates a link between awareness and sophisticated mental recordings, such that subliminal conditioning of any kind ought to be physically impossible within a mental architecture solely designed to process and produce propositionally structured mental representations. Motivated by debates between associative and propositional models, Custers and Aarts (2011) set out to test whether or not predictive relations of unidirectional associations could be formed in the absence of conscious awareness. Associations can be formed in a bi-directional manner, where perceiving either of two events brings the other to mind, or in a unidirectional manner, where noticing E1 evokes E2, but not vice versa. Predictive relations refer to the stored knowledge of how two or more events relate to one another, e.g. E1 consistently precedes E2. Because unidirectional associations capture important relational information, these mental constructs are typically seen as requiring conscious attention, and would be classified as propositional according to De Houwer's inclusive notion of propositionality.

The results of their three-part study indicate that unidirectional associations can be formed in the absence of conscious awareness so long as attention is 'tuned' to process predictive relations. In experiment one, participants were split into two groups: group one was primed to process predictive relations before participating in the acquisition phase of the experiment, whereas group two directly entered the latter phase. Priming was achieved by asking participants to quickly sort two targets (a circle or triangle) whose classification could be

[11] For an explanation of how a purely propositional structure could account for the dizzying speeds of automatic processes, see De Houwer (2014). In short, while conscious awareness is necessary to kickstart many of these automatic processes, it has little to do with the implementation of those processes.

predicted by a subtle cue. As hypothesized, those in the priming condition were statistically more likely to form unidirectional associations despite not being consciously aware of the predictive cues. This suggests that top-down attentional processes track relations between stimuli and concepts, and govern the storage of relational information independently of conscious awareness.

One of the main criticisms of Corneille & Stahl's (2019) meta-analysis was the quality of evidence for subliminally acquired EC. While there are many studies supporting the idea that evaluative associations can be acquired without conscious awareness, Corneille and Stahl note that these experiments often fail to ensure that conscious awareness is properly masked or otherwise re-directed from predictive cues. If a CS-US pair is consciously apprehended, even tangentially, then this could account for any subsequent conditioning. In fact, this is precisely what PAL suggests: noticing a predictive cue would be enough to trigger storage of the relation, even if the cue was not ascertained *as a* cue by the subject or was subsequently forgotten due to top-down processes failing to store the information. This means that the vast majority of evidence for unconscious EC is dubious at best.

To avoid these criticisms, experiments two and three implemented a pre-mask to prevent any awareness of predictive cues. As in the first experiment, participants whose top-down processes were unconsciously primed to track predictive relations were much more likely to form unidirectional associations in the second phase, providing striking evidence of unconscious conditioning. This suggests that the criticisms levied by Corneille and Stahl against subliminal associative acquisition may be due to a conflation of conscious awareness with top-down attention. As many theorists have argued, these two facilities are dissociable, and may play

different roles in learning (Baars, 1997; Dehaene et al., 2006; Lamme, 2003; Koch & Tsuchiya, 2007. In fact, Custers and Aarts dispute some of the findings that Corneille and Stahl use to base their critique of subliminal EC precisely on these grounds.[12]

Together, these findings support the idea that attention-tuning can occur independently of conscious awareness and that top-down processes govern the storage of predictive relations in memory. How damaging are these findings to PAL? In a forthcoming book chapter, De Houwer acknowledges the weight of this evidence leaves the single-process propositional theorist with two options: 1) concede that a second non-propositional mechanism produces certain instances of associative learning, or 2) drop the assumption that propositional models must be entertained consciously before they can influence behavior (De Houwer, forthcoming, p. 11). Option two is an ad-hoc assumption and contradicts one of the main tenets of PAL, so barring any future findings, De Houwer grants that there are likely two separate learning mechanisms responsible for the production of attitudes.

2.2 Implications for the Class of Implicit Attitudes

The existence of two learning mechanisms marks dual-process theories as the best psychological model to evaluate implicit attitudes. But as we have seen, how we define relevant terminology can hinder our evaluation of the desired mental phenomena. By defining propositionality as the recording of any relation between stimuli or concepts, De Houwer cannot

[12] In particular, Custers and Aarts dispute the findings of Pleyers, Corneille, Luminet, and Yzerbyt (2007). Since the domain of top-down attention exceeds that of conscious awareness, future arguments against subliminal conditioning should focus on clarifying the role of top-down awareness in implicit learning.

account for instances of subliminal conditioning without abandoning the idea that propositional encoding requires conscious awareness. But evidence of subliminal unidirectional associations suggest that associatively linked mental constructs can record more than just contiguous relations, indicating that associations, as a class, can possess a range of specificity over their relata. While likely not very large, this range still establishes satisfaction conditions for associatively structured mental states.

A unidirectional association between events E1 and E2 ensures that a primitive causal connection holds between the perception of E1 and the associatively evoked E2. The associative link will respond to the appropriate stimuli presented in the format A → B, but not to the same stimuli presented as B → A. Bi-directional associations, on the other hand, are blind to the logical implications of their relata, activating the other member of an associative link when either of the two is perceived. What we have, then, is an associative link that operates over and above the Hebbian principle *fire together, wire together*. Unidirectional associations are not merely the representation of contiguous stimulus pairings, but specific renditions of an environmental regularity, one that accords with a predictive principle that furthers an organism's ability to navigate its surroundings. The interaction between bottom-up processes (which supply information) and top-down attentional processes (which allocate mental resources and governs the storage of information) allows for the specification of conceptual relations over associatively linked mental states. These links are more sophisticated than their bi-directional counterparts, yet still fall far short of the kind of inferential promiscuity characteristic of beliefs.

According to De Houwer, for a representation to be propositionally structured just is for that representation to possess relational information over its relata. Having such information means that a representation is more than a reflection of external regularities, indeed, it is sensitive to logical and semantic considerations (Mandelbaum, 2013). Specifying the relation between two or more concepts signifies that a representation is making a statement about the world, and such statements are subject to accuracy conditions; the relation A → B can be true, false, or possess some degree of predictive accuracy. Top-down attentional processes are sensitive to the success that these relations have, and store information states accordingly.

This notion of propositionality supports a large philosophical canon committed to a tight syndicate between language and thought. This idea has been thoroughly defended by Jerry Fodor, who holds that thought must be composed of representational-like vehicles (that is, propositional structures like those found in language) in order to realize the systematicity characteristic of reflective cognition (Fodor, 1975; 1983; 1987; Fodor & Pylyshyn, 1988). While this focus on 'language as thought' has been criticized in recent years (Camp, 2007), the prevailing notion is that higher-order thought – at least in humans – is achieved through language via propositions and their relations. The sense of propositionality endorsed by the likes of Fodor and Mandelbaum forges a link between the activity of a mental representation and the operations of a rule-governed mental schema, such that propositionality is predicated on the functional capacity of a representation to respond to rational concerns and conform to the laws of logic. But if one admits that, then we are left to wonder by what standard we are to adjudicate the relative propositionality of a representation. In other words, if the determinant of propositionality is the functional capacity to meet certain satisfaction conditions, then what are we to make of those

representations that display a haphazard sensitivity to such conditions? Admittedly, this is not a problem for language of thought theorists. After all, one can accept the idea that a representation has a compositional semantics, in that there is meaning to be discerned from the representation's relation with other mental contents, without endorsing the belief that said representation must scrupulously adhere to the laws of logic.[13]

But this does pose a problem for philosophers who treat evidence of propositionality as evidence of a mental state's being a belief (Smith, 2005, 2012; Egan, 2011). While this inference is undoubtedly motivated by the ability of beliefs to respond to evidence and interface with other mental states – indeed, these capacities are exactly what one would expect of a mental representation that was sensitive to rational demands and normative constraints – evidence of distinct learning mechanisms and relatively sophisticated associative mental links ought to give pause to those whose notion of doxasticity is equivalent to rule-governed mental states. The fact that the central claims of two of the most popular attitude models – APE and PAL – cannot currently be falsified reinforces the limitations of arguments that use these claims as justifications for rigid distinctions within the mind. If there are to be hard distinctions, these divisions must have sound support before we insert epistemic theories and folk psychological notions of mental phenomena into the uncertainty of attitude research. For this reason, I suggest using the associative/propositional distinction to refer to the learning mechanisms responsible for producing attitudes. This articulation allows for somewhat sophisticated interactions amongst associative states without threatening the perception of propositional states as uniquely situated

[13] See the Language of Thought Hypothesis at SEP for more: https://plato.stanford.edu/entries/language-thought/

to form the basis of higher-order thought, in virtue of their content-responsiveness and interactive capacities.

There are a few caveats that need mentioning. First, the presence of dual-learning mechanisms does not imply that there are distinct memory stores for associative and propositional attitudes. While some have argued for just such a view (Rydell & McConnell, 2006; Smith & Decoster, 2000), findings in favor of distinct memory stores have failed to be replicated. Moreover, dual-representational views had the least success at predicting the last thirty years' worth of EC findings (Corneille & Stahl, 2019).

Second, while I have only provided evidence of two such mechanisms, others have recently argued that the best explanation for the confusion within EC research, and attitudinal studies at large, is the presence of multiple learning mechanisms, each with its own operating principles and conditions (March et al., 2018). Such a view is compatible with a single-representational account of attitudes so long as one posits that all learning mechanisms feed into the same memory system. In fact, the idea that distinct learning mechanisms share a single memory system is an integral part of Gawronski and Bodenhausen's (2011) associative-propositional model (APE) of attitudes. Though conducive to my aim to portray implicit attitudes as distinct from beliefs, there is currently only evidence of two such learning mechanisms; as such, I am formally committed to the more parsimonious dual-mechanism account.

Third, the interaction between associative mental links and top-down processes may at first seem to substantiate Angela Smith's claim that implicit attitudes are reflective of an agent's

authorial stance. Afterall, a stereotypical unidirectional association, such as the associative mental link between the concept HISPANIC and the feature HARDWORKING, appears to reflect a sophisticated appraisal on behalf of an agent. But as we have seen, the relevant factor in the storage of unidirectional associations is predictive value, which occurs independently of one's conscious awareness. Smith's account seems wedded to the idea that top-down processes are equivalent with, or subservient to, conscious awareness, but a third experiment performed by Custer and Aarts challenges such an idea. When participants were given an explicit learning goal to predict targets based on the primes in the activation phase of the experiment, the resulting associative mental links were bi-directional rather than unidirectional, leading Custer and Aarts to conclude that conscious awareness can impede the functional ability of top-down processes to track and store predictive cues, even when this is the express goal of an agent. If top-down processes are to be included in the set of mental states that comprise an agent's authorial stance, then it appears that by pursuing certain goals we undermine our own agency.

Finally, I conceive of the associative/propositional distinction as tracking the division between implicit attitudes and beliefs. Specifically, I equate the products of the associative learning mechanism with implicit attitudes, and the products of the propositional learning mechanism with explicit attitudes, i.e. beliefs. This characterization forms the basis of what I call the dual-mechanism theory of implicit bias (henceforth, DMT): the idea that all instances of implicit bias are attributable to the activation of either an associative attitude (an implicit attitude) or a propositional attitude (an unconscious belief). Strictly speaking, implicit attitudes are only associative in nature, however, explicit attitudes operating below conscious awareness can issue in the same prejudiced behavior. Consequently, only a subset of implicit biases are

caused by implicit attitudes. The next section introduces a leading theory of consciousness that justifies the distinction between implicit attitudes and beliefs, and clarifies how explicit attitudes can remain barred from conscious awareness.

2.3 The Global Workspace Theory of Consciousness

The global workspace theory of consciousness (henceforth, GWS) was first proposed by Bernard Baars (1988). He argued that the mind possesses a common workspace where distinct and specialized systems can assemble, compare, and communicate information. To occupy the GWS is for an information state to be occurrently tokened in the mind of an agent. Conscious mental states are made widely available to a host of different systems, including "…those for forming memories, for forming new values, for creating affective states, and for reasoning and decision making…" (Carruthers, 2015, p. 52). Hence, representations occurrently tokened in the workspace enjoy unparalleled access to information states and diverse cognitive systems.[14]

Baars used the blackboard model to explain how exchanges of information and decision-making unfolds in a largely modularized mind. The blackboard model was originally proposed by artificial intelligence researchers investigating how domain-general processing can occur in a system largely composed of specialized processes (Nii, 1986). To make sense of the model, they proposed the following analogy: consider a room filled with specialists who can only communicate by writing on a blackboard; only one specialist may use the board at a time, and they may only share information resulting from their own expertise. When a problem or idea is

[14] For Levy (2015), it is the widespread connection with other mental states – which allows for greater integration and refined thought – that makes conscious information the determinant of moral responsibility.

introduced, each specialist with something to contribute vies for access to the board where they may share their findings. Information continues to accrue until the problem is resolved or the idea concluded.

The GWS model proposes that information only becomes conscious when three conditions are met (Robinson, 2009). First, incoming information must be represented by networks of sensory neurons, such as those that make up our perceptual systems, e.g. auditory and visual neurons. Second, for this representation to reach consciousness it must gain access to a second stage of processing, requiring the mental state to outcompete other information states. Third, bottom-up propagation of information and top-down amplification must cohere in such a way as to cause a general integration of information states across a variety of brain regions. These areas form an interactive web of cortical space that respond as needed to task demands and internal information processing. However, not all areas of this cortical space must be active to 'ignite' conscious thought – only those required for the task at hand.[15] Moreover, the space itself is designed in such a way that only one conscious representation can be sustained at any given time (Sergent et al., 2005; Sigman and Dehaene, 2005, 2008).

To see how this works, consider the intense competition that occurs between the outputs of specialized sensory processes. Relevance and limited attentional resources restrict what information is broadcast and how frequently it attains entrance to the shared workspace. Successful entrance can be attributed to bottom-up processing (as when you see a snake-like figure on the sidewalk) or amplification from top-down processes (when your current concerns

[15] I stress this point to avoid accusations that the GWS is a modern declaration of a 'Cartesian theatre'.

and goals leap to mind). Thus, not only is pertinent contextual information made readily available to one's conscious self, but the things we deem significant (fighting with a spouse, having an important deadline, wishing to eat healthier) are given priority over less germane ideas and, as a result, are frequently broadcast within the GWS.

Peter Carruthers has used the GWS to argue for a single-representational dual-attitude account of implicit bias, which holds that any functional difference between implicit and explicit attitudes is due to extrinsic relations with other mental items (Carruthers, 2018). According to this view, all instances of implicit bias are ultimately traceable to the activation of either implicit cognitive attitudes (propositional structures) or implicit affective states (evaluative structures).[16] While the two often align, studies show that the two are dissociable (Amodio & Devine, 2006; Gilbert et al., 2012) and are stored in different areas of the brain (Phelps et al., 2014). Hence, someone can have "…a stereotype (even a negative stereotype) about a social group without having a negative affective attitude toward that group, and vice versa." (Carruthers, 2018, p. 55).

A stereotypical cognitive attitude can be systematically barred from conscious awareness for two reasons: either the attitude is continually outcompeted by other mental states vying for access to the shared workspace, or the attitude is actively inhibited by top-down processes. The latter is particularly pertinent for doxastic models, since an agent with a vested interest in avoiding uncomfortable truths about themselves can unwittingly (though willingly) inoculate themselves to their own attitudes. For example, someone committed to egalitarian values may nevertheless possess some racial prejudice, but because this belief conflicts with their other

[16] This work will focus mostly on implicit cognitive attitudes, since implicit affective attitudes are equally compatible with the dual-learning mechanism view of attitudes.

deeply cherished beliefs, the underlying attitude might be suppressed to avoid cognitive dissonance. This suppression does not have to be a conscious effort on the part of the individual, but can occur via the guiding influence of top-down processes over information states supplied by bottom-up processes.

Philosophers of race are keenly interested in the ways in which we can isolate ourselves from unwelcome information, particularly as it concerns certain social facts. Consider Elizabeth Spelman's analysis of the willful ignorance of many white Americans in response to *g*: the idea that Black America's grievances are real (Spelman, 2007). According to Spelman, there are two features that characterize this epistemic state:

1) *W* does not believe that *g* is true and does not want to believe that *g* is true.
2) *W* does not believe that *g* is false but wants to believe that *g* is false.

A person in this state is neutral in regard to the veracity of proposition *g*, yet their motivational stance is biased towards *g* being false. Someone in *W*'s position avoids even thinking about the challenges of black Americans – to do otherwise would be to threaten their cherished epistemic neutrality; such strong motivations could easily kindle the kind of top-down suppression that bars some attitudes from reaching conscious awareness. Furthermore, this inability to acknowledge reality forms the basis of what Robin DiAngelo calls 'White Fragility', a state in which "…even a minimum amount of racial stress becomes intolerable, triggering a range of defensive moves" (DiAngelo, 2011, p. 54). These defensive moves can be seen in personal-level behavior, such as blocking certain persons on Facebook, but are also present at the

subliminal level, as in the kind of early-stage competition that Carruthers states is typical of information states. Someone in *W*'s position may possess a stereotypical cognitive or affective attitude that remains unconscious due to top-down suppression and heavy competition with other mental states. This inhibition is likely caused by the agent's need to perpetuate their self-image and the host of beliefs tied to the falsity – or at least neutrality – of *g*.

Carruthers' single-representational dual-attitude view is, in my opinion, the strongest doxastic account of implicit bias to date. As stated in the last section, I think much of what implicit measures track are unconscious beliefs; the GWS eloquently demonstrates how attitudes which would otherwise be 'explicit' can remain relegated to the subliminal domain in a diachronic manner, either they are outcompeted in early-stage information processing or they are suppressed by top-down processes because they conflict with conscious mental states. Despite its virtues, this view overlooks a crucial fact about attitudes – they can be associative. As discussed, associative attitudes can be the reflection of contiguous states of affairs (bi-directional structures) or the somewhat more sophisticated associative predictive relations of environmental occurrences (unidirectional structures). Carruthers' claim that all attitudes are solely differentiated in virtue of how they relate to other mental contents is undermined by the presence of an isolable associative learning mechanism. As such, there is a legitimate psychological distinction between attitudes produced by an associative learning mechanism, which operate according to little or no satisfaction conditions, and attitudes produced by a propositional learning mechanism, whose content can possess a range of highly specific relational information. The presence of dual learning mechanisms suggests that attitudes can vary not only in terms of

their relation to other mental contents – an extrinsic fact – but in virtue of their formative history as well – an intrinsic fact.

2.4 The Implicit/Explicit Distinction According to Dual-Mechanism Theory

Associations can be quite resistant to change. In fact, there are only two ways to alter an associative mental link: either via counterconditioning (which replaces the former association with a new one) or extinction (where the mental link dissolves from disuse). This points to a fundamental difference between propositional and associative attitudes: propositional attitudes can be sustained in conscious awareness and modified by an agent himself. But the processes which take associative attitudes as content merely deliver one member of an associative link to the domain of awareness – the relation itself does not feature in awareness nor can it be modified directly. To illustrate, consider the associatively evoked image of your grandmother in response to an airy perfume. While you are free to ruminate on this image, you cannot ruminate on the association itself. Even if you were made aware of this association (perhaps you took psychology 101) this kind of knowledge is one step removed from the associative link between GRANDMA and PERFUME which, at heart, is the co-activation of one stimulus response with another, and does not allow for the kind of rapid update common to beliefs. For instance, if I thought that the second president of the United States was Thomas Jefferson, and you provide sufficient evidence that it was most definitely John Adams, I have *the ability* to change that belief (whether or not I do is another matter). You, on the other hand, can no more will the association between PERFUME and GRANDMA to cease than I can will Thomas Jefferson to be the second POTUS. Any successful modification of this associative link will proceed indirectly, either via

counterconditioning or extinction, neither of which exhibits the remarkable responsiveness to counterevidence that typifies doxastic states.

In line with this observation, Grace Helton has recently proposed a moderately revisionary view of belief that makes conscious modification the mark of a doxastic state (Helton, 2018):

> **THE REVISABILITY VIEW OF BELIEF (RVB)**
>
> Necessarily, if some subject's mental state that *p* is a belief and if that subject has sufficiently strong, undefeated evidence that *not-p*, then that subject is able to revise that mental state, given her current psychological mechanisms and skills.

Beliefs are distinguished from other cognitive attitudes (entertained thoughts, pretenses, non-doxastic delusions) by their capacity to respond to counterevidence and for information to be updated at the level of conscious awareness. As such, RVB should be seen as articulating a necessary feature of belief. If a mental state is inherently incapable of entering conscious awareness, or is characteristically insensitive to relevant information, that state is not a belief. To be clear, my aim here is not to defend this particular view of belief, but to use conscious modification to pry apart the conflation of implicit attitudes with unconscious beliefs. Recall the Reflective/Non-Reflective distinction introduced in section 1:

Non-Reflective	Reflective
(a) Unconscious	(a^1) Conscious
(b) Associatively Structured	(b^1) Propositionally Structured
(c) Automatically Activated	(c^1) Deliberately Controlled
(d) Not avowed	(d^1) Avowed

Figure 2. Four distinctive features of dual-process theory (re-visited)

The Non-Reflective and Reflective columns represent the functional profile of implicit and explicit attitudes respectively. On the account I have sketched, implicit attitudes are necessarily unconscious mental states. Bi-directional associations operate in accordance with the Hebbian principle *fire together, wire together*, where two concepts can develop an associative mental link in virtue of frequent co-activity. When either member of the link is perceived, the other is evoked and is likely to gain access to conscious awareness. While either member of the associative link can achieve global broadcasting, *the link itself* cannot enter the global workspace; indeed, I am not sure what it would even mean for an association, as a mental link, to be consciously represented. Given that the GWS can only sustain one mental representation at a time, and associations are, by definition, two or more mental representations linked through habitual co-activation, it seems that global broadcasting of an associative link is a psychological impossibility given the mental architecture that humans possess.

If associations were encoded with precise relational information, then perhaps they could be modified indirectly via top-down processes when a member of the associative link was tokened in conscious awareness (and presented with sufficient counterevidence). But since associations possess little to no relational information and are only ever represented as a constituent of a mental link, it is a mystery how conscious awareness could alter the subliminally acquired and habitually reinforced reflection of a contiguous state of affairs. Unidirectional associations, being the product of interactions between bottom-up and top-down processes, are associative links encoded with predictive relational information. While more sophisticated than bi-directional associations, these mental states are also immune to conscious modification since

the associative mental link, being a co-activation between mental states, cannot enter the shared workspace, with its entry limitations.[17]

Because implicit attitudes are associations, and associations operate via automatic processes, we should expect subjects to reveal certain attitudes besides those they consciously avow when operating under time-constraints or cognitive strain. And this is precisely what implicit measures report. Though implicit and explicit attitudes often coincide, the two can diverge – a fact made possible by the diachronic and subliminal nature of associative formation. Moreover, the rapid social change that has characterized the last fifty years of American history – which has moved from debates over busing in the 1970's (a recent hot topic) to the swearing in of America's first African American president – presents the perfect context for tension between those attitudes deemed socially acceptable and those that are not. Indeed, the residue of our nation's unsavory past can be traced to the automatic responses of everyday individuals, whose repeated exposure to stereotypical images, rhetoric, and ideas, usher implicit attitudes through the backdoor of our mental life.

Explicit attitudes, on the other hand, are the products of a propositional learning mechanism, and despite the name, are not always made explicit to those who have them. When propositional attitudes are produced, top-down processes store precise information governing the relation between their relata in memory, allowing these mental states to operate in a systematic

[17] The claim that associative links can be consciously represented is best understood as a category mistake. Although agents can reflect on associatively evoked mental states, and even ponder the relation between those representations, associative links are never consciously broadcast, i.e. they do not feature in awareness with their semantic content. Propositional attitudes are inseparable from their semantic information, and since they are stored with specific relational information, are easily updated at the level of conscious awareness – a feat simply not possible for associative links, which are merely the inward representation of an external regularity.

manner with other mental contents. The reason I cannot pay you to believe the sky is green is that too many of your other beliefs state otherwise; given the inclination of beliefs to abide by normative constraints – a byproduct of their specific content and connections with other mental states – the only way to change a belief is by influencing those around it, or by presenting evidence which appeals to the content of the target belief. Moreover, given the fact that propositional attitudes have been shown to feature in both automatic and deliberate processes (De Houwer, 2014), (c_1) should be read as a capacity of explicit attitudes, not a limitation. Since propositional attitudes are not constituents of a mental link predicated on co-activity with other mental states, they and their semantic content can be accurately represented in conscious awareness. The specificity and availability of this content allow for the rapid updating of beliefs and the storage of these updates in memory via top-down processes. Finally, because inconsistent explicit attitudes result in cognitive dissonance, there is internal pressure for an agent to align their beliefs into a congruent whole, culminating in remarkably consistent – though ultimately imperfect – coalitions of diverse information states.

This last fact is plainly demonstrated by David Lewis, who articulates how even conscious beliefs can fall short of flawless integration:

> "I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel… So each sentence in an inconsistent triple was true according to my beliefs, but not everything was true according to my beliefs." (Lewis, 1982, p. 436).

The fact that Lewis failed to notice the inconsistency between three of his beliefs – beliefs used on other occasions to justify or inform behavior – points to the fact that even conscious attitudes can fail to interact with other mental states in an optimal manner. Note, however, that

this limitation is not due to anything inherent to the attitude itself, but is instead a consequence of the attitude's extrinsic relations with other mental contents. It is safe to say then, that perfect inferential promiscuity is not required for an attitude to be a belief so long as these systematic limitations can be attributed to extrinsic circumstances; otherwise, a host of mental items we normally take to be paradigmatic doxastic states, such as those held by Lewis, would not be considered beliefs.

The next section uses DMT to explain some of the most influential studies within implicit attitude research. As we will see, the presence of two separate learning mechanisms best explains why some unconscious attitudes are somewhat responsive to evidence and other mental states, while also allowing for the kind of incremental attitude change evinced by classical conditioning.

Section 3. The Dual-Mechanism Theory of Implicit Bias

Philosophers and psychologists tend to characterize implicit attitudes as a uniform kind. Examples include: low-level associations (Rydell & McConnell, 2006; Gawronski & Bodenhausen, 2011), behavioral dispositions (Schwitzgebel, 2010; 2013), patchy endorsements (Levy, 2015), character traits (Machery, 2016), unconscious beliefs (Mandelbaum, 2015; Carruthers, 2018), and even unconscious imaginings (Sullivan-Bissett, 2018). Each of these views assumes that implicit attitudes are either a homogenous class of mental items composed of a similar structure and functional profile, or a uniform set of behaviors (Holroyd & Sweetman, 2016). But reconciling a unified account of implicit attitudes with the tumultuous state of affairs within attitude research has had fair-to-middling success, prompting some to toy with the idea

that implicit attitudes may constitute a homeostatic-property cluster (Stammers, 2018). The purpose of this section is to clarify the ambiguous nature of implicit attitudes encouraged by seemingly contradictory empirical findings. In doing so, I defend the idea that veritable implicit attitudes are a uniform category composed of low-level associations, but that much of what implicit measures track are propositionally structured unconscious beliefs.

I begin by examining two studies that suggest implicit attitudes are capable of much more than what associative accounts predict. These studies form the basis of Mandelbaum (2015) and Carruthers' (2018) doxastic theories of implicit bias, and have much to offer in the way of articulating the operation of unconscious beliefs. After examining these studies through the lens of DMT, I transition to three studies that suggest implicit attitudes are only mildly responsive to other mental states and largely oblivious to the logical implications of their subject matter. Neil Levy (2015) uses these studies to advance the position that implicit attitudes are a sui generis class of mental states, somewhere between mere associations and bona fide beliefs in terms of content-responsiveness and inferential capacity (Levy, 2015). After responding to these concerns, I close by reviewing the state of associative research and its effect on implicit attitude studies.

3.1 The Case for Unconscious Belief

Our first study examines the role of cognitive balance theory in the construction of interpersonal attitudes (Gawronski et al., 2005). One of the basic suppositions in Fritz Heider's (1958) balance theory is that there tends to be a consistency in the triadic relationships that

represent our affective responses. For example, if P1 likes P2, and if P2 likes object *X*, then P1 is more apt to like *X* as well. Gawronski and colleagues wanted to test whether the predictions of balance theory would hold when applied to the operation of implicit attitudes within a social context. They began by presenting participants with an unfamiliar photo – CS1. Subjects were exposed to repeated pairings of the photo with positively or negatively valanced words, until subjects developed a corresponding bias towards CS1. Once instilled, subjects were exposed to a new photo – CS2 – and told that CS1 either liked or disliked CS2. To conclude the test, subjects were given an affective priming task that tracked their implicit reactions towards both photos.

As balance theory predicts, subjects were more likely to have positively valanced responses to CS2 when they were conditioned to like CS1. In other words, the positive affect originally directed to CS1 was extended to CS2. This bodes well for AIB, which holds that associations between concepts and things can accord with affective transference – one positive plus another positive can result in a further, additional positive. But when subjects were conditioned to respond negatively to CS1, and subsequently told that CS1 disliked CS2, subjects demonstrated a positive implicit bias towards CS2 – the exact opposite of what balance theory (and AIB for that matter) predicts. This case demonstrates how two negative affections – one directed at CS1 and the other produced by CS1 disliking CS2 – can result in a third, positive affective response. Associative accounts of implicit bias cannot explain how an automatic and classically conditioned attitude can operate in accordance with rudimentary logic, i.e. a double negative. As Mandelbaum describes it, this experiment suggests that subjects are making an inference akin to the proverbial saying *the enemy of my enemy is my friend*. The inferential capacities of unconscious processes should not be surprising in and of themselves, after all, they

form an integral part of modular theories of the mind (Fodor, 1983; 2000). Yet, given the supposition that implicit attitudes are purely associative, the accuracy with which they track the semantic contents of propositional structures provides a strong reason to think that the functionality of implicit attitudes extends beyond mere contiguous relations between concepts.

The second study reveals that mere mental ruminations can produce implicit biases in equal strength to those generated by classical conditioning. Gregg et al. (2006) devised a test to evaluate if DPT could account for any differences, should they arise, between 'concrete' versus 'abstract' learning styles and the acquisition of implicit attitudes. They define 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" and abstract learning as "hypothetically assuming that an object possesses particular characteristics" (Gregg et al., 2006, p. 4).

In experiment one of their four-part study, participants were divided into two groups: a 'concrete learning' group and an 'abstract learning' group and asked to evaluate two hypothetical tribes – the Luupites and Niffites. Participants in the concrete condition underwent 240 rounds of classical conditioning, in which positive valence words were paired with the Luupites, and negatively valanced words were paired with the Niffites. The abstract condition was merely asked to imagine the two tribes, one of which (the Luupites) was peaceful and civilized, and the other (the Niffites) barbarous and brutal. Each group was then asked to take an IAT to see if any implicit biases had developed, and if so, to evaluate their relative strength. Lo and behold, both the concrete and abstract learning groups had incurred an implicit bias towards the Niffite tribe,

and what is more, the relative strength was roughly equal across the two groups. The fact that a single instance of abstract thinking elicited the same level of implicit bias as a paradigm example of sustained associative conditioning is startling. Proponents of AIB are hard pressed to explain how a single imaginary episode can produce the same level of implicit bias as 240 rounds of classical conditioning; doxastic accounts, on the other hand, have a ready answer "all groups formed the same (strong) belief that Niffites were bad while Luupites were good" (Mandelbaum, 2015, p. 16).

According to DMT, the underlying attitude in each study was an unconscious belief, not an implicit attitude. The inferential ability evinced by attitudes in the first study should only be possible if the underlying mental state possessed a propositional structure, which, according to DMT, indicates that an attitude was encoded with specific relational information. Because this specificity is found only in attitudes produced by a propositional mechanism, doxastic theories are right to deduce the underlying mental state is an unconscious belief. Having precise relational information enables a mental state to efficaciously feature in inferences and other such computational processes. While it may be possible for an attitude with haphazard relational information, like a unidirectional association, to feature in an inference, the lack of specificity would undermine the accuracy of any process which took such an attitude as content. Therefore, given that the accuracy of the attitudes in Gawronski and colleague's (2005) study was high, it is best to interpret this study as tracking the inferential capacity of unconscious beliefs, not implicit attitudes.

The second study is a perfect example of one of the central claims of PAL: that propositional attitudes can be formed on the fly and stored in memory thanks to the guiding hand of top-down processes. When an agent is asked to consciously represent the relation between some tribe (the Niffites) and some feature (barbarous) they understand this information via the evaluation of propositional structures; the ensuing imaginary episode triggers top-down processes to store the information for future use, whose activation is subsequently detected when participants take an IAT. Traditional IAT's are designed to track disparities in a subject's response times between stereotypical and non-stereotypical information – a measurement that cannot distinguish between automatically activated associations (implicit attitudes) and automatically activated propositions (explicit attitudes). Since propositionally structured information can feature in both automatic and deliberate processes, it is probable that what this experiment demonstrates is not the spontaneous formation of implicit attitudes, but the precipitous formation and subsequent effect of unconscious beliefs on personal-level behavior.

3.2 Evidence for a Sui Generis Mental State

There are three studies that challenge my claim that implicit attitudes are purely associative mental states and that much of what implicit measures track are unconscious beliefs. The first study questions the ability of implicit attitudes to feature in inferences with any degree of accuracy, and by extension, illustrates a remarkable lack of content-responsiveness to other mental states. Rozin and colleagues (1986, 1990) tested the ability of attitudes to 'bind' to certain objects in a way that influenced personal-level behavior despite a subject's consciously held beliefs. The experiment unfolds as follows:

> "Subjects faced two empty brown 500 ml bottles. In the presence of the subject, the experimenter opened a container of "Domino" cane sugar, and poured some into each bottle, so that about ¼ of each bottle was filled. The experimenter informed subjects that she was pouring sugar into each bottle. The experimenter then presented the subject with two typed labels. One had not 'sodium cyanide', not poison written on it, with a red skull and cross bones preceded by the word 'not'. The other label had 'sucrose, table sugar' typed on it. The subject was invited to put one label on each bottle, in any way he or she chose. The experimenter then set out two different colored plastic cups, one in front of each bottle, and poured unsweetened red (tropical punch) 'Kool-Aid' from a glass pitcher into both, until they were about half full. Now, using separate, new plastic spoons for each bottle, the experimenter put a half spoonful of powder from one sugar bottle into the glass standing in front of that bottle, and repeated this with the other glass for the other sugar bottle."[18]

According to Mandelbaum, associative accounts of implicit bias cannot explain how the general apprehension of subjects in this experiment 'binds' to one bottle, as opposed to any other object or merely subsisting in general (Mandelbaum, 2013). Instead, he posits that a person forms an unconscious belief that takes the 'poison' jar as its subject. What might such a belief look like? He offers the following candidate: "THAT IS DANGEROUS CYANIDE, SO AVOID IT" as well as the following inferential pattern "THAT BOTTLE CONTAINS POISON, PEOPLE DO NOT LIKE DRINKING POISON, SO PEOPLE WILL NOT LIKE DRINKING FROM THAT BOTTLE" (Mandelbaum, 2013, p.204). But as Neil Levy points out, what is remarkable about this experiment is not that the underlying attitude transpired in an inference, but that the inference was *blind* to the semantic content of the second jar's label – *not* sodium cyanide, *not* poison (Levy, 2015). Hence, while it appears that some unconscious attitudes are content-responsive and may feature in inferences, the degree to which they accord with reason and rational constraints is far less than that of conscious beliefs.

[18] Rozin et al 1990, *op cit*.

Besides tracking logical relations, beliefs ought to be able to update accordingly when presented with new information. In our second study, Han and colleagues' (2006) tested to see whether 'extrapersonal associations', such as peer evaluations, might have an effect on implicit attitudes. First, children were taught facts about the card game Pokémon, specifically, which cards were better for someone who wanted to win the game. Next, they were exposed to a video in which two other children expressed opposing opinions about which cards were best. While subjects rejected the opinions of the children in the video (since they conflicted with their aim to win the game) a subsequent IAT revealed a change in subjects' implicit attitudes. Although the children expressed preferences for the objectively better cards, the opinions of their 'peers' was enough to modify their implicit responses in favor of the cards preferred by the children in the video. The control group, which was not shown the video, did not exhibit any discord between their implicit and explicit responses. Hence, some unconscious attitudes update when they should not, demonstrating an illicit sensitivity that one would not expect of beliefs or belief-like states, which are prone to operate under normative constraints.

Conversely, there is evidence to suggest that implicit attitudes fail to update when they should. Recall the findings of Gregg et al. (2006), which used two fictional tribes to test how learning measures effect the inculcation of implicit biases. After establishing that abstract supposition could create implicit attitudes just as effectively as prolonged associations, Gregg and colleagues tested to see if the bias could be undone or altered by new information. In experiment three, they informed participants who had already inculcated a bias that there had been a mistake and that the Luupites were actually barbarous and brutal while the Niffites were peaceful and civilized. The results found that while the self-reported preferences of subjects

updated in light of the new information, their automatic responses remained the same; in other words, subjects' implicit biases continued to guide their behavior despite their being aware that the two groups had 'mistakenly' been mixed up. Once set, the content of an implicit bias does not appear very receptive to further information, *even if* the original acquisition of said bias was caused by a single imaginary episode.

Together, these studies suggest that implicit attitudes have a broader functional profile than mere associations, but fall far short of the inferential promiscuity and content-responsiveness indicative of doxastic states. In what follows, I argue that the underlying attitudes in the above three studies are actually unconscious beliefs, and defend the view that implicit attitudes are purely associative in nature.

3.3 Against a Sui Generis Account

Critiques of doxastic theories threaten the functional account of inhibited explicit attitudes and their corresponding effects on behavior I portrayed in section 2.4. Therefore, it is necessary to show in what way the above studies are compatible with the tenets of DMT.

The findings of Rozin et al. (1986;1990) challenge associative and doxastic accounts by presenting a mental state with lackluster content-responsiveness and inferential promiscuity – which is beyond the operative abilities of associative mental states and below that of beliefs – prompting Levy to claim that implicit attitudes are 'patchy' endorsements of sorts. But DMT can explain these instances without postulating an additional mental entity. Recall the claim of

propositional theorists (De Houwer et al., 2009; Michael et al., 2009) that propositions can encode a variety of relational information, including that of mere co-occurrences; in fact, this is why they contend there are no associations, in the normal sense of the word, whatsoever. If the specific relational information of a propositional attitude encodes a co-occurrence, then the operation of that attitude will be indistinguishable from that of a pure association, despite that attitude having the *capacity* to update at the level of conscious awareness, given its being the product of a propositional learning mechanism. Since the participants of these studies were consciously aware of the propositional information presented to them (indeed, they played a part in setting up the experiment) it is possible they formed a propositional attitude encapsulating an associative relation between the bottle (object) and the affect-laden label (poison). Propositions, as the representational-vehicles of systematic thought, have no trouble featuring in inferences; nevertheless, a proposition whose relation was that of a co-occurrence would suffer in terms of accuracy in the same way that any process which took a representation with limited relational information as content would be prone to imprecise calculations. Hence, an insensitivity to the logical implications of the propositionally represented 'not' is precisely what one would expect of a propositional attitude with the semantic content of a bi-directional association.

The second study presents a case where an unconscious attitude updates in response to non-relevant information (opinion) and conflicts with the subject's conscious goals (to win the card game). This receptivity should not be possible for an associative mental state, which can only change via counter-conditioning or extinction, and challenges doxastic accounts on the grounds that beliefs are reliably content-responsive, responding to other information-states as reason warrants. For Levy, this is evidence of a unique class of mental items with a degree of

inferential promiscuity and content-responsiveness somewhere between associations and bona fide beliefs. But the GWS, which Levy is intimately acquainted with, allows for another interpretation.[19] Recall that our current concerns and goals have a differential effect on which information states achieve global-broadcasting. While a representation may have the ability to enter conscious awareness, it can remain confined to the subliminal domain due to early-stage competition between information states or from top-down suppression. As Carruthers notes, top-down processes often monitor access to the shared workspace in accordance with social norms; if a word or action could cause lasting harm to one's reputation (as a sexist or homophobic remark might) then top-down processes are likely to suppress that representation in favor of other, more anodyne mental states. So, too, can a mental state be barred from conscious awareness if it provokes cognitive dissonance, as evinced by cases of 'willful ignorance' and 'white fragility'. Similarly, subjects in the Han and colleagues' (2006) study could have formed the unconscious belief that the objectively worse cards were in some way desirable or useful in response to the opinions of their 'peers'. Because this attitude is in contention with the express beliefs of participants, it would normally be suppressed in day-to-day affairs; however, the time constraints imposed by traditional IAT's allow such attitudes to manifest in the automatic responses of participants, thereby bypassing the inhibition of any top-down forces that might be in effect.

Finally, the third study presents evidence that once formed, unconscious attitudes can be surprisingly resistant to counter-evidence, even if the attitude is the recent product of a single imaginary episode. Although the explicit attitudes of participants updated in response to the news

[19] Neil Levy (2014) has used the GWS theory of consciousness to support a view of moral responsibility that relies on the consciousness condition, which states that we only exert the requisite self-control over the moral significance of our actions when we are consciously aware of their moral character.

that the two fictitious tribes had been mixed-up, their implicit attitudes remained the same. How can DMT, which postulates that the underlying attitude is either an associative or propositional mental state, accommodate these findings? The answer lies within our affective mechanisms. As Carruthers notes, if we assume that top-down processes can quickly store appraisals of novel objects in one's valuational mechanisms, and that once set, these appraisals can only be altered through slow incremental change, then the findings of Gregg and colleagues (2006) can be accommodated by a doxastic – or dual-mechanism – view. Indeed, some studies suggest that preferences resulting from imagination can last for over three years (Sharot et al., 2012). Because the affective status of objects we encounter are unlikely to change quickly – a poisonous food will remain poisonous, lions will continue to be a threat, and so on – it makes sense that our valuational systems would be receptive to new information but resistant to wanton change.

To close, each of the studies cited by Levy to suggest that implicit attitudes are a unique mental state, separate from associations and beliefs, can be addressed by DMT. However, the gist of my arguments against the 'patchy' endorsement theorist is compatible with doxastic accounts, namely those of Mandelbaum (2015) and Carruthers (2018). As such, it is only proper to explain why implicit attitudes ought to be equated with associative mental states when so much of what implicit measures track can be elucidated by appealing to unconscious beliefs.

Section 4. Implicit Attitudes and the Resurgence of Associative Research

The success of propositional models has encouraged skepticism towards associative attitude research (Lovibond & Shanks, 2002; Newell & Shanks, 2014), prompting a renewed

effort to understand the relationship between conscious awareness and attitude formation. De Houwer's (2009) propositional account of attitude learning (PAL) has been quite effective at casting doubt on associative theories of attitudes, and has gained increasing support in virtue of its parsimonious endorsement of a single-process architecture and a solitary learning mechanism. Nevertheless, the findings of Custers and Aarts (2011) provide striking evidence of an isolable associative learning mechanism, leading De Houwer to conclude that there are most likely two distinct learning mechanisms responsible for the production of attitudes within the human mind. While I suggested that doxastic accounts of implicit bias have much to gain from incorporating the empirical evidence of PAL, any theory which ignores evidence of associative attitudes risks selectively endorsing evidence and enforcing an unwarranted distinction between human beings and the rest of the animal kingdom.

To clarify, in his discussion on animal cognition, De Houwer (forthcoming) notes that associative processes probably emerged fairly early in evolutionary history, with propositional processes developing much later. Once an organism developed a propositional learning mechanism, the need for a similar associative process would be moot, leading to the eventual rewiring of associative mental processes for other cognitive purposes. While De Houwer frames this fact in an inclusive manner, noting that many non-human animals are likely to possess propositional learning mechanisms and simple propositionally-structured thoughts (a fact I agree with), this nevertheless casts homo sapiens as a sui generis species, unique among the animal kingdom in our possession of a purely propositional mental format. Instead, it is much more likely that we, like many other animals, can form associations in response to our environment that are not propositionally structured. After all, classical conditioning has demonstrated that

animals can form associative links composed of a conditioned stimulus (CS) and a conditioned response (CR) (Pavlov, 1897). In fact, classical conditioning has been successfully forged in such diverse animal species as: honey bees, (Bitterman et al., 1983), marine mollusks (Hawkins & Byrne, 2015), snails (Takigami et al., 2015) and fish (Barretto et al., 2018). The mental architecture of these animals varies in many ways, but each species is nevertheless capable of forming associations strong enough to influence their behavioral patterns in response to contextual regularities. Moreover, although the field of animal cognition has yet to conclude whether animals can possess beliefs, the fact remains that a myriad of species can form associative links between salient stimuli, meaning it is conceptually possible for an associative link to exist irrespective of a doxastic attitude or a mental architecture capable of sustaining belief-like representations.[20] Therefore, to conclude that human cognition is solely composed of propositionally structured mental representations, and that these representations can account for all instances of associative behavior, is a daring presumption indeed.

Nevertheless, there are valid critiques to some of the designs that support associative attitudes. One of the strongest critics of implicit learning and the presence of distinct learning mechanisms comes from Shanks and Stjohn (1994), who identify four criteria that must be met to adequately demonstrate unconscious learning of supraliminal cues: 1) the *sensitivity criterion* requires appropriate sensitivity of the measures of awareness (to avoid conscious contamination), 2) the *information criterion* suggests that the measure of awareness and the experimental task

[20] Proponents of the 'language as thought' thesis affirm that doxastic states require a representational basis grounded in a propositional structure; consequently, animals with such a structure can properly be said to possess beliefs (Fodor, 1975; Cheney & Seyfarth, 2007). Others claim that animals possess beliefs by extending the representational format of doxastic states to include non-propositional structures, such as imagistic thought (Camp, 2009; Rescorla, 2009).

should probe the same information (to avoid conflating variables), 3) the *immediacy criterion* states that testing should immediately follow the experimental task (to enhance target detection), and 4) the *relevance criterion* advises that the measure of awareness should ignore irrelevant information (to avoid tracking unrelated phenomena). Alamia and colleagues (2016) tested to see whether subliminal associative learning could occur under these stringent requirements. In their experiment, participants were asked to report the motion direction of a colored patch of dots. They were not told, however, that of the three colors the dots could take, two of them were associated with motion direction. Hence, there was a predictive relation between the color of a dot and the direction in which it was oriented. Besides adhering to the four criteria, the study also asked participants a series of questions to see if any of the participants might have become aware of the relation during the experiment. Despite this maneuver, it is always possible that participants tangentially noticed the relation but failed to recall it when queried, meaning their top-down resources could have stored the information for future use. To avoid such a case, the final portion of the experiment informed participants of the predictive relation and then had them take an additional test. If participants had become aware of the relation, there would have been little discrepancy in their first and second scores; but if they were not aware of the relation, then one would expect their performance to improve the second time around.

The results found that most participants remained unaware of the association between color and motion direction, yet still acquired the predictive relation between the two stimuli. For the majority of participants, scores increased on the final test, suggesting that their increased performance before learning the rule was the result of a subliminal association predicated on predictive accuracy. Because this study conforms to the four criteria, it is not immune to the

same objections that plague traditional associative studies, such as conscious contamination and conflation of observed phenomena. Indeed, as Alamia and colleagues put it "We believe that our study provides the first demonstration of unconscious learning of simple associations…" (Alamia et al., 2016).

Section 5. Conclusion

Using the most stringent procedures, recent research has established the presence of a distinct associative learning mechanism (Custers & Aarts, 2011), that associations can form subliminally (Alamia et al., 2016; 2018) and that dual-process models which posit multiple learning mechanisms can best explain instances like the Perruchet effect, where conscious expectations and associative mental links dissociate in regard to performance ability (Destrebecqz, 2018). Together, these findings support associative views of implicit bias which postulate that: 1) implicit attitudes are low-level associations between mental states, 2) that associations can form outside of conscious awareness and in opposition to the expressed beliefs of an agent, and 3) that associations can only change via counter-conditioning or extinction.

The purpose of this work has been to defend the claim that implicit attitudes are purely associative in nature and to illustrate how much of what implicit measures track are actually unconscious beliefs. The findings of Gregg and colleagues (2006) provide an excellent illustration of how two such attitudes might arise: implicit attitudes are associative structures produced in response to contextual regularities (such as 240 rounds of classical conditioning), while explicit attitudes are the result of a propositional mechanism which operates in conjunction

with conscious awareness (such as imaginary episodes). Critics of the dual-mechanism approach may note that I have not cited a specific study showing an associatively produced implicit attitude distinct from a non-broadcasted explicit attitude. This is no accident. As the psychologists at the forefront of the attitude debate note (Gawronski & Bodenhausen, 2018; De Houwer, forthcoming) there is currently no definitive method of discerning an associatively linked attitude from a propositional structure encoding a mere co-occurrence. Hence, evidence for one or the other will have to rely on indirect cues, such as how the attitude responds to relevant counter-evidence and other mental states. As DMT predicts, true-blue implicit attitudes ought to be utterly insensitive to logical relations and mental contents beyond their associative link. To verify these claims, one would need to produce a stereotypical attitude via classical or evaluative conditioning in the absence of awareness. Since actual stereotypes would be recognized and presumably stored somewhere in an agent's memory, the experiment would have to create fictitious social groups, akin to Gregg and colleagues' (2006) study; however, it would be necessary to consciously mask the relation between a concept (Social Group A) and some feature (Good/Bad). In other words, one would have to occupy conscious awareness in an attempt to keep an agent from noticing the predictive relation between some social group and a particular feature, which is not a hard thing to do, as evinced by Custer and Aarts (2011) and Alamia and colleagues (2016).

To close, there are two ways we can characterize implicit attitudes. First, we can consider any underlying mental state which issues in a biased or prejudiced action as a member of the class of implicit attitudes. Such a construal remains agnostic towards specific claims regarding the structure and functional profile that an individual attitude has, which, given the widespread

disagreement in the field, might be a good tactic if one's primary objective is to address the problems of implicitly biased behavior. A second approach, and the one that I favor, is that we distinguish implicit from explicit attitudes in virtue of their formative history, which has a differential impact on their respective functional profiles. Being the product of an associative learning mechanism endows an attitude with features *a*…*d*, while attitudes generated from a propositional mechanism have characteristics $a_1 \ldots d_1$. This distinction provides new perspectives on old findings (heuristic value), effectively preserving the traditional depiction of implicit attitudes as purely associative mental states, while also generating new predictions (predictive value), by articulating the factors under which an explicit attitude can remain unconscious while influencing agent-level behavior. Although ultimately a terminological dispute, efforts to eradicate implicit social biases can only gain from an accurate understanding of the mental states which produce them. Prejudiced actions resulting from an implicit attitude are best combatted by counter-conditioning of the underlying associative link, or extinguishing it altogether. Unconscious beliefs, on the other hand, can be consciously addressed only if the obstacles preventing their broadcasting are removed. Sometimes these obstacles are self-imposed, as when an individual actively avoids reflecting on certain issues, or it can occur unwittingly, via the subliminal activity of one's top-down processes. Specifying the interaction between top-down processes and conscious thought presents a challenge for contemporary discussions of moral agency and moral responsibility – topics which dual-mechanism theory can help illuminate.

References

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2008). Are lesbians discriminated against in the rental housing market? Evidence from a correspondence testing experiment. *Journal of Housing Economics*, *17*(3), 234-238.

Alamia, A., Orban de Xivry, J., San Anton, E., Olivier, E. Cleermans, & A., Zenon, A. (2016). Unconscious associative learning with conscious cues. *Neuroscience of Consciousness*, *2016*(1). doi: 10.1093/nc/niw016.

Alamia, A., Solopchuk, O., & Zenon, A. (2018). Strong conscious cues suppress preferential gaze allocation to unconscious cues. *Frontiers in Human Science*, *12*(427). doi: 10.3389/fnhum.2018.00427

Amodio, D. & Devine, P. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652-661.

Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*(4), 292-309.

Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour research and therapy*, *30*(2), 133-142.

Banaji, M. & Greenwald, A. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Banaji, M. R., & Greenwald, A. G. (2013). Blindspot: Hidden biases of good people. New York, NY, US: Delacorte Press.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology, 56*(5), 329-343.

Barretto, A., Maia, C., Alves, N., & Giaquinto, P. (2018). Water jet: a simple method for classical conditioning in fish. *Acta Ethologica*, *21*(24), 1-5.

Bendick Jr, M., Rodriguez, R. E., & Jayaraman, S. (2010). Employment discrimination in upscale restaurants: Evidence from matched pair testing. *The Social Science Journal*, *47*(4), 802-818.

Bitterman, M., Menzel, R., Fietz, A., Schäfer, S. (1983). Classical conditioning of proboscis extension in honeybees (apis mellifera). *Journal of Comparative Psychology*, *97*(2), 107-119.

Block, N. (2014). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research*, *89*, 560-572.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press.

Buss, D. M., & von Hippel, W. (2018). Psychological barriers to evolutionary psychology: Ideological bias and coalitional adaptations. *Archives of Scientific Psychology, 6*(1), 148-158.

Byrd, N. (forthcoming). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese*. doi: 10.1007/s11229-019-02128-6

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, 21(1), 145–182.

Camp, Elisabeth (2009). A language of baboon thought? In Robert W. Lurz (ed.), *The Philosophy of Animal Minds*. Cambridge University Press. pp. 108--127.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford University Press UK.

Carruthers, P. (2018). Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures? *Review of Philosophy and Psychology*, 9(1), 51-72.

Cheney, D., Seyfarth, R. (2007). *Baboon Metaphysics*. Chicago: University of Chicago Press.

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161-189.

Correll, J., Park, B., Judd, C., & 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.

Custers, R., & Aarts, H. (2011). Learning of predictive relations between events depends on attention, not on awareness. *Consciousness and cognition*, *20*(2), 368-378.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences*, *10*(5), 204-211.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish journal of psychology*, *10*(2), 230-241.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1-20.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347-368.

Houwer, J. D. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342-353.

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*, e28046.

De Houwer, J. (in press). Why a propositional single-process model of associative learning deserves to be defended. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), Dual processes in social psychology. NY: Guilford.

Destrebecqz, A., Vande Velde, M., & San Anton, E. (2018). Saving the Peruchet effect: A role for the strength of the association in associative learning. *Quarterly Journal for Experimental Psychology*, *72*(6), 1379-1386.

DiAngelo, R. (2011). White fragility. *The International Journal of Critical Pedagogy*, *3*(3), 54-70.

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

Egan, A. (2011). Comments on Gendler's, "the epistemic costs of implicit bias". *Philosophical Studies*, 156(1), 65-79.

Evans, J. & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241.

Fazio, R. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603-637.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The el greco fallacy in perception research. *Psychological Science*, *25*, 38–46.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass*, 5(10), 914-926.

Fritz, H. (1958). *The psychology of interpersonal relations*. Psychology Press.

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

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

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3-71.

Fodor, Jerry A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312-325.

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. & Bodenhausen, G. (2011). The associative-propositional evaluation model. *Advances in Experimental Social Psychology*, *44*, 59-127.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300-312.

Gendler, T. (2008). Alief and belief. *Journal of Philosophy,* 105(10), 634-663.

Gendler, T. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, 156(1), 33-63.

Gilbert, S., Swencionis, J., & Amodio, D. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*, 3600-3611.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Greenwald, A., McGhee, D., & 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. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17-41.

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

Han, H., Olson, M., & Fazio, R. (2006). The influence of experimentally-created extrapersonal associations on the implicit association test. *Journal of Experimental Social Psychology*, *42*, 259-272.

Hawkins, R., & Byrne, J. (2015). Associative learning in invertebrates. Cold Spring Harbor Perspectives in Biology, 7(5). doi: 10.1101/cshperspect.a021709

Helton, G. (2018). If you can't change what you believe, you don't believe it. *Noûs*.

Hoffman, K., Trawalter, S., Axt, J., & Oliver, M. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *PNAS*, *113*, 4296-4301.

Holroyd, J. & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), Implicit bias and philosophy Metaphysics and Epistemology (Vol. 1, pp. 80-103), Oxford: Oxford University Press.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, *61*(3), 465-496.

Jacobi, T. & Schweers, D. (2017). Justice interrupted: The effect of gender, ideology, and seniority at supreme court oral arguments. *Virginia Law Review*, *103*(7), 1379-1485.

Jung, C. G. (1932). *Psychological types* (p. 425). London: Pantheon Books.

Kahneman, D. (2012). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, *11*(1), 16-22.

Lamme, V. A. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural networks*, *17*(5-6), 861-872.

Levy, Neil (2014). *Consciousness and Moral Responsibility*. Oxford University Press.

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

Lewis, David (1982). Logic for equivocators. *Noûs*, *16*(3), 431-441.

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28*(1), 3-26.

Mandelbaum, Eric (2013). Against alief. *Philosophical Studies* 165 (1):197-211.

Mandelbaum, Eric (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*, 50(3), 629-658.

Mandelbaum, E. (2018). Seeing and conceptualizing: Modularity and the shallow contents of perception. *Philosophy and Phenomenological Research*, *97*, 267-283.

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein and J. Saul (Eds.). Implicit Bias and Philosophy: Metaphysics and Epistemology (pp. 104-129). Oxford: Oxford University Press.

March, D. S., Olson, M. A., & Fazio, R. H. (2018). The implicit misattribution model of evaluative conditioning. *Social Psychological Bulletin*, *13*, e27574.

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

R Newell, Ben & R Shanks, David. (2014). Unconscious influences on decision making: A critical review – ADDENDUM. *The Behavioral and Brain Sciences*. 37. 1-19. 10.1017/S0140525X12003214.

Nii, H. P. (1986). The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, *7*.

Oswald, F., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. (2013). Predicting ethnic and racial discrimination: meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171-192.

Pavlov, I. (1897/1902). *The work of the digestive glands*. London: Griffin.

Payne, B. K. (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, K. & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, *8*(12) 672-686.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557-569.

Petty, R. E., Fazio, R. H., & Brinol, P. (Eds.) (2009). Attitudes: Insights from the new implicit measures. New York: Psychology. Press.

Phelps, E., Lempert, K., & Sokol-Hessner, P. (2014). Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience*, *37*, 263-287.

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis) liking: item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 130.

Puddifoot, Katherine (2017). Dissolving the epistemic/ethical dilemma over implicit bias. *Philosophical Explorations* 20 (sup1):73-93.

Ramjattan, V. (2019). Racist nativist microaggressions and the professional resistance of racialized English language teachers in Toronto. *Race Ethnicity and Education*, *22*(3), 374-390.

Rescorla, Michael (2009). Cognitive maps and the language of thought. *British Journal for the Philosophy of Science*, 60(2), 377-407.

Robinson, R. (2009). Exploring the "global workspace" of consciousness. *PLOS Biology*, *7*(3). doi: 10.1371/journal.pbio.1000066

Rosenberg, M. J., & Hovland, C. I. (Eds.). (1966). Attitude organization and change: An analysis of consistency among attitude components. Oxford, England: Yale U. Press.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, *50*(4), 703-712.

Rozin, P., Markwith, M., & Ross, B. (1990). The sympathetic magical law of similarity, nominal realism and neglect of negatives in response to negative labels. *Psychological Science, 1*(6), 383-384.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995-1008.

Saul, J. (2013). Scepticism and Implicit Bias. *Disputatio* 5 (37):243-263.

Saul, J. (2013). Unconscious influences and women in philosophy. *Women in Philosophy: What Needs to Change?,* F. Jenkins & K. Hutchinson (eds.), Oxford: Oxford University Press.

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

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. *New Essays on Belief*, 75-99.

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391-1400.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*(3), 367-447.

Sharot, T., Fleming, S. M., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting?. *Psychological science*, *23*(10), 1123–1129. doi:10.1177/0956797612438733

Sigman, M. & Dehaene, S. (2005). Parsing a cognitive task: A characterization of the mind's bottleneck. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.0030037

Sigman, M. & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience*, *28*(30), 7585-7598.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115*(2), 236-271.

Smith, A. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics*, 122(3), 575-589.

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by manipulations of source credibility. Personality and Social Psychology Bulletin, 39, 193-205.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and social psychology review*, *4*(2), 108-131.

Sloman, S. A. (2014). Two systems of reasoning, an update. In Sherman, J., Gawronski, B., & Trope, Y. (Eds.). Dual process theories of the social mind. Guilford Press.

Spelman, Elizabeth V. (2007). Managing ignorance. In Shannon Sullivan Nancy Tuana (ed.), *Race and Epistemologies of Ignorance*. pp. 119--31.

Solomon, M. (2008). *Consumer behavior buying, having, and being* (8th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Stammers, S. (2018). A patchier picture still: Biases, beliefs and overlap on the inferential continuum. *Philosophia*, *45*, 1829-1850.

Steinpreis, R., Anders, K., & Ritzke, D. (1999). The impact of gender of the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*, 509-528.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62*(4), 271-286.

Sullivan-Bissett, E. (forthcoming). Biased by our imaginings. *Mind and Language*.

Takigami, S., Sunada, H., Lukowiak, K. (2015). An automated learning apparatus for classical conditioning in lymnaea stagnalis. *Journal of Neuroscience Methods*, 259. doi: 10.1016/j.jneumeth.2015.10.008

Uwe, Peters. (2018). Implicit bias, ideological bias, and epistemic risks in philosophy. *Mind and Language*. doi: 10.1111/mila.12194

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*(1), 101.