ATTITUDES:

A MEMORY SYSTEMS PLURALIST PERSPECTIVE

by

Michael Scott Sechman

B.A., California State University, Northridge 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Philosophy

Institute of Cognitive Science

2021

Committee Members:
Robert Rupert
June Gruber
Leaf Van Boven
McKell Carston
Iskra Fileva

Sechman, Michael Scott (Ph.D., Philosophy and Cognitive Science)

Attitudes: A Memory Systems Pluralist Perspective

Thesis directed by Professor Robert Rupert

This dissertation develops and defends the memory systems pluralist (MSP) theory of attitudes. This holds that there are a relatively large number of species of attitudes (hence 'pluralism'), each of which is subserved by a distinct causal mechanism best conceived of as a memory system (hence 'memory systems'). On the MSP theory, attitudes are mental states that cause a special class of behaviors known as evaluative responses. Though this perspective is compatible with memory systems models on which all attitudes are of the same representational format, I will argue that data from both cognitive neuroscience of memory and social cognition research best supports a version of MSP according to which some types of attitude are associations, and some types of attitude are propositional structures. For these reasons, I suspect that the most successful MSP models will be those that license pluralism both in regard to attitudes *qua* class of psychological entities and in regard to their representational format. Moreover, I will tentatively conclude, *contra* social psychological orthodoxy, that the best available evidence does not warrant the view that implicit attitudes are categorically distinct from explicit attitudes. The central argument for this version of MSP theory takes the form of an inference to best explanation. In particular, I argue that the MSP theory has the best prospects for resolving two sets of anomalies that have long plagued social psychological theorizing on attitudes. One set of anomalies, which I call the core anomalies, suggests that measures of implicit attitudes have low predictive validity, weakly correlate with other measures of (putatively) the same phenomenon, and suffer from unacceptably low levels of test-retest reliability. The other set of anomalies, which I call the format anomalies, suggest that implicit attitudes paradoxically possess the defining features characteristic of both associations and propositionally structured representations.

DEDICATION

In Memory of Sean Hudson, Ph.D. (1989-2015)

ACKNOWLEDGEMENTS

I must begin by thanking my advisor, Rob Rupert. Thank you for the countless hours you spent reading and engaging with my work over the past several years. Thank you for helping to further encourage and cultivate my passion for various topics at the intersection of philosophy and cognitive science. And thank you for all of the support, professional and personal, that you have provided over my lengthy tenure. I wouldn't have arrived at this point without you. Thanks Rob!

I also owe special thanks to Steve Stich. Steve was responsible for my conversion to the "dark side" of philosophy, which is how he referred to as empirically informed philosophy. While writing this dissertation, I often had the feeling as though I was trying (and failing) to write *Fragmentation of Reason* but about attitudes. The best ideas in this dissertation are ones I lifted from Steve. I count my time spent with Steve as amongst the most gratifying experiences of my life. Cheers, Steve!

I am also deeply indebted to my committee members. Thanks all!

Dan Kaufman! Call me!

Special thanks are due to Jonathan Wright for the friendship, encouragement, support, and for letting me talk endlessly about my work. I have and continue to learn a great deal from him. He, in many respects, is the philosopher, the person, and in every respect the dancer that I could only ever aspire to be. Thanks man.

Tim Black was my undergraduate advisor at California State University, Northridge. Without Tim's mentorship, I would never have even considered applying to a graduate program in philosophy. Tim, you gave me both the skills and the confidence needed to do something as truly insane and horrifying as pursuing a doctorate in philosophy. Thank you again, for your excellent mentorship as well as your friendship. While we're on the topic, I'd like to thank the rest of CSUN's philosophy faculty who were haunting the Sierra Tower during my time there, especially Weimin Sun, Takashi Yagisawa, Bonnie Paller, Ron McIntyre, Adam Swenson, Rob Gressis and Julie Tannenbaum.

I owe a tremendous debt to Karen Sites, the graduate program assistant in philosophy, and the labor of all the workers in the philosophy department without whom we could not do whatever exactly it is that we do.

I quite literally would not be here to write this were it not for the love and support of my closest friends. Jade and Alex, I love you always and forever. I cannot tell you how proud I am to be a member of the cartridge family. Sergio, you will always be family. Mike, you are an inspiration. Hector, thanks for being the big brother I've always wanted. I'm also especially grateful to Anthony, Alexis, Damien, Derek, Gabe, Justin, Misho, Mony, Zack, Zach, and, of course, Sean. Thank you all for helping to sustain me at various points over the years!

Last, but not least, I want to thank my mom, dad, and sister. Mom, you deserve so much more than what I could ever hope to give. I love you more than you could ever know.

CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figures

PREFACE

From the field's inception in the early 20th century, attitudes have occupied the theoretical center stage in social psychology. Gordon Allport (1935), universally recognized as one of the founding figures of scientific social psychology, articulated and affirmed the received view of the day that attitudes are so central to the field that social psychology can be conceived of as the scientific study of attitudes. To this day, spurred on in part by the introduction of the implicit/explicit attitude distinction in the 1980s', the focus of much of social psychological research and theorizing continues to prioritize understanding the attitude construct(s) and its effects on behavior.

This sustained, longstanding interest in attitudes is not difficult to explain. Explanations, across any number of domains, frequently appeal to attitudes under various guises. Consider folk psychological explanation: Jonathan voted for Bernie because Jonathan *preferred* Bernie's platform to Hillary's; Alex bought the analog modular synthesizer because he *liked* it; Jade attended the Dandy Warhol's concert because she considers them the greatest band of all time. The assumption that attitudes predict and explain behavior is on full display in various marketing and advertising practices. In 2020 alone, political campaigns shelled out the extraordinary sum of $8.5 billion for political advertising across TV, radio, and digital media, most of which, it can be safely assumed, were aimed at changing potential voters' attitudes toward candidates, platforms, ballot initiatives, and so on, for the purpose of influencing voting behavior. Businesses produce and develop ads with the aim of changing consumers' attitudes toward their products for the purpose of influencing consumer choices. And attitudes at various points have either been thought to be one component of class, racial, gender, ethnic, and ableist prejudices (Allport 1935) or the wholly constitutive of these prejudices. When it comes to the various "-isms" at the individual level, it is widely assumed that people discriminate against members from marginalized social groups because they harbor noxious, individual or culturally enforced, attitudes toward these groups and their members.

This dissertation is primarily about the nature and sources of attitudes. It sets out to answer the deceptively simple question, what exactly are social psychologists measuring when they report their various findings on attitudes? Are attitudes mere dispositions to engage in evaluative behavior or are they mental states or representations that drive evaluative responses? Do attitudes constitute a unified, homogenous kind? Or are attitudes a heterogenous kind? On the assumption that implicit attitudes are mental states, what of their representational format? Are they associations? Are they propositional structures? What of the distinction between implicit attitudes and explicit attitudes? Does the distinction carve nature at its joints? Is the distinction merely nominal?

As social psychology has made surprisingly little progress despite a century's worth of sustained scientific interest, there is a sense in which the project at the heart of this dissertation is an extraordinarily ambitious one. And insofar as I set out to provide what I think are the most defensible answers to these questions yet to be proposed, this is true. And yet, from a different perspective, the goals of this dissertation are relatively modest as I attempt to answer these questions not by placing my empirical bets on any specific attitude model but by offering and defending a theoretical framework of attitudes that differs dramatically from the frameworks that one typically encounters in the social psychological literature. I leave it as a task for future research to determine which specific model attitudes best satisfies the constraints that must be met, or so I argue, of any plausible model.

I shall defend what I call the "memory systems pluralist" (MSP) theory of attitudes. This holds that there are a relatively large number of species of attitudes (hence the pluralism), each of which is subserved by a distinct causal mechanism best conceived of as a memory system (hence the memory systems). On the MSP theory, attitudes are mental states on which evaluative dispositions depend. And while, strictly speaking, this perspective is compatible with models on which all attitudes are of the same representational format, I will argue that data from both cognitive neuroscience of memory and social cognition research best supports a version of MSP according to which some types of attitude are associations, and some types of attitude are propositional structures. In this sense, I suspect that the most successful MSP models will be those that license pluralism both in regard to attitudes *qua* class of psychological entities and in regard to their representational format. Moreover, I will tentatively conclude, *contra* social psychological orthodoxy, that the best available evidence does not warrant the view that implicit attitudes are categorically distinct from explicit attitudes.

Insofar as I develop a theoretical perspective of attitudes that draws heavily from memory research in cognitive neuroscience, one goal of this dissertation is to integrate and provide a theoretical focus for a great deal of disparate work in cognitive science. While an interdisciplinary approach to attitude research is by no means unprecedented, social psychologists do not as a rule deeply engage with the findings from surrounding disciplines in cognitive science and *vice versa*. So, a related goal of this dissertation is to show that this lack of sustained interdisciplinary engagement has systematically thwarted scientific progress with respect to the scientific study of attitudes. By explicitly conceiving of attitudes as species of mnemonic structure, subserved by the various memory systems uncovered over decades of memory research in cognitive neuroscience, I hope to establish that this level of interdisciplinary engagement is not merely instrumental in making scientific progress but essential to it: social psychologists cannot hope to make serious progress in understanding the causes of intelligent social behavior but by adopting conceiving of their project as a thoroughgoingly interdisciplinary one.

Another goal of this dissertation is to inform, and in some cases challenge, philosophical theories that assume the truth of various empirical claims about the mind and the causal determinants of behavior. For example, Rob Rupert (2011) has recently proposed a view, supported by MSP theory, that the mind is massively representational— that is, that the mind contains multiple representational vehicles that possess the same representational content. As Rupert has urged, *ceteris paribus*, the quantity of redundant representations is an important factor in addressing a number of important questions including, for instance, consciousness. Moreover, the theory of mind and cognitive architecture that falls out of the view defended in this dissertation does not sit comfortably with standard folk psychological assumptions about the types of mental states that act as the casual determinants of intelligent behavior— *viz.* that intelligent behavior is best explained by belief-desire pairs. Some of the potential implications for projects near and dear to philosophers will be addressed in this dissertation's concluding chapter (Chapter 7). With that said, I make no sustained attempt to argue for the relevance of empirical research to traditional philosophical questions. I take as an unargued for assumption the Quinean position that philosophy of mind is continuous with cognitive science.

While this dissertation seeks to illuminate the nature of attitudes *simpliciter*, I spend relatively little time on the topic of explicit attitudes. Indeed, the vast majority of this dissertation attempts to address longstanding puzzles concerning the nature of implicit attitudes. This emphasis is, in large part, driven by the strategic considerations presented in Chapter 1.

As the aim of this dissertation is to develop a theoretical account of attitudes that *best* accounts for the full range of empirical data, the central argument for the MSP theoretical perspective takes the form of an inference to the best explanation. Successful inferences to the best explanation involve both (a) evaluations of how well available theories accommodate existing evidence and (b) evaluations of how successfully each theory coheres with our other well-confirmed theories in cognitive science. I endeavor to show that MSP outperforms its rivals on both counts spanning many different topics. Lastly, because the central argument of this dissertation is an inference to the best explanation, all of its conclusions are tentative and can be upended by future empirical findings. In keeping with my naturalist scruples, though, I consider this a feature not a bug.

For the benefit of readers whose background is in philosophy, I need to make two further remarks. First, I've already spoken at length about attitudes and it should by now be clear that the way in which I use the term 'attitude', unless explicitly stated otherwise, diverges from its standard usage in philosophy. In philosophy, the term 'attitude' is typically used to refer to *propositional attitudes*. For those whose backgrounds is in psychology, propositional attitudes are stances that one might take toward propositions (*e.g.*, I *believe* that whales are mammals, I *hope* that Sanders wins the Democratic primary, I *judge* that 2+2 = 4, I *fear* that I won't complete the project by the deadline). I leave it as an open question which of the various species of attitude count as propositional attitudes; according to MSP theory some may straightforwardly admit of such a gloss and some won't.

Second, I need to emphasize that this dissertation does not fit the mold of much of contemporary analytic philosophy in that it neither contains very much that could be recognized as conceptual analysis, nor does it make very many substantive claims that are intended to be knowable *a priori*. In light of these remarks, many philosophers might be concerned that this dissertation is largely devoid of philosophical content. Be this as it may, I am inclined to frame this dissertation as an exercise in theoretical psychology (*cf.* Carruthers 2011). As Carruthers aptly points out, this is a sort activity that Hume, and many others, would and do recognize as a kind of philosophy. And this, on my view, is the kind of philosophy that anyone who is serious about addressing questions in the philosophy of mind ought to engage in.

Disappointingly for some readers, this dissertation devotes little time to the thorny normative questions that seem to drive so much interest, popular and academic, in social cognition research. I do not spend much time on the question of whether one is a racist or sexist if one harbors problematic implicit biases toward members of some marginalized group. I won't be probing into the relationship between one's attitudes and the self. There is little discussion about how attitude research should inform debates about agency. Note well that when I say that I devote little space to exploring each issue, I do mean to suggest that these and like issues will be categorically ignored.

Many readers might expect from a work on attitudes some helpful, practical advice on how to change or eliminate socially problematic attitudes. Though I do examine a great deal of research on attitude formation and change, I address this topic only briefly in the concluding chapter. Those hoping to encounter in these pages useful advice on how to develop an effective implicit bias training seminar for one's department, university, or Fortune 500 company, are bound to be disappointed.

Finally, by way of initial orientation, I must stress an additional background assumption. The first is that the mind is real. By this, I do not merely mean that assertions about the mind have truth values— a claim that nearly everyone but the most hardcore eliminativists about the mind would accept. Rather, in keeping with many other naturalistically minded philosophers, I assume that the "mind has an existence and substantive character that goes well beyond, and is independent of, our best common-sense interpretative practices. (Carruthers 2011: xiv)" An upshot of this view, as already suggested, is that knowledge of the mind cannot be acquired merely by reflecting on these practices. It also means that those social psychologists who adopt an instrumentalist outlook on scientific theorizing may be puzzled by some discussions contained herein.

CHAPTER 1
INTRODUCTION

The goal of this opening chapter is to provide an initial outline of the theory of attitudes (*i.e.*, likes, dislikes, preferences, etc.) to be proposed and defended in this dissertation. The main claims and commitments of the theory will be laid out and contrasted with those of its main rivals. After making a few comments on the strategic importance of implicit attitudes *vis-à-vis* the central aim of this dissertation, I conclude with a chapter-by-chapter guide to the overall structure of the dissertation, indicating how the theory in question will be compared with its competitors.

1 A Theoretical Overview of Memory Systems Pluralism

In broad brushstrokes, the memory systems pluralist (MSP) perspective maintains that the mind-brain contains multiple functionally distinct memory systems, each of which is differentially involved in the acquisition, retrieval, and behavioral expression of different species of attitude (*cf.* Amodio 2018). That the term 'attitude' here is to be understood broadly, to cover all manner of likes, dislikes, preferences, appraisals, *etc.,* the tokens of which, under the right circumstances, are apt to causally produce or influence a special class of behaviors called *evaluative responses* (*cf.* De Houwer, Gawronski & Barnes-Holmes 2013). It is a *pluralist* theory of attitudes in that it denies that attitudes are a unified, homogenous class of mental states; indeed, pluralism is committed to the existence of many more than two species of attitude.

The view assumes that for each attitude object, it is very likely that the mind-brain contains multiple functionally distinct attitudes toward it (*cf.* Wilson, Lindsey & Schooler's (2000) *dual-attitude* theory). Additionally, it assumes that attitudes differ from one another in terms of (a) the processes involved in their learning and extinction, (b) the rates at which their learning and extinction occur,

(c) their functional properties (*i.e.,* with respect to the kinds of behaviors that they cause and the endogenous and exogenous features of the context that cause them to activate), (d) their representational format (*i.e.*, some are associative structures while others are propositional structures), and (d) the neural systems that subserve their storage and the various processes that operate over them. These assumptions entail the claim articulated above that attitudes are a wildly heterogenous class of psychological entities. But the heterogeneity is not endless; there are recurrent patterns that found within species of attitude and this order is explicable in terms of the underlying causal mechanisms that produce them. For these reasons, MSP theory carries a number of deeply important implications for both the empirical investigation of attitudes and for how we are to best interpret existing data. We spell out some of MSP's more salient predictions in the next section.

It may be helpful to see that the core theoretical commitments of the MSP perspective can be presented as a conjunction of four theses:

> (a) there exist multiple memory systems (or, if one prefers, learning systems) underlying the acquisition, activation (or retrieval), updating, and behavioral expression of attitudes;
>
> (b) these memory systems, while distinguishable by their functional properties and by the neural regions that subserve them, typically interact in complex ways in the production of intelligent social behavior;
>
> (c) for each attitude object, an individual may have many more than one (sometimes competing, sometimes congruent) attitude toward it; and,
>
> (d) for each attitude token, its type is determined by appeal to its underlying memory system.

As we will see, with the exception of the taxonomical principle expressed by (d), the different component claims of MSP theory are supported by differing sets of data and each will be in focus at different points in the dissertation. Taken together, these four theses entail pluralism about attitudes in general and, more specifically, implicit attitudes.

The five hypotheses presented below either fall directly out of MSP theory or are elaborations on different aspects of the core commitments. Accordingly, these additional hypotheses may be regarded as part of the hard core of MSP theory; a serious blow to the MSP theory would be struck should one of these hypotheses be shown to be false. (Note well that the following list is by no means meant to be exhaustive.)

(1) The MSP perspective assumes that a single learning episode may result in the formation of multiple attitudes that co-exist across the mind-brain's various memory systems.

(2) Different memory systems learn at different rates: some memory systems may encode an attitude after only a single learning episode, while other memory systems may require multiple learning episodes.

(3) The MSP theory assumes that the retrieval of an attitude from one memory system need not imply that an attitude toward the same attitudinal object has been retrieved for use from any other memory system.

(4) For any arbitrarily selected pair of attitudes toward the same attitude object stored in different memory systems, A_1 and A_2, it is possible to strengthen or extinguish A_1 without thereby strengthening or changing A_2.

(5) Attitudes stored in different memory systems may be expressed in distinct, but overlapping, sets of behavioral responses.

To help illustrate the view, consider an example plucked from Elizabeth Phelps and Stefan Hoffman's (2019) discussion of the challenges of applying recent advances in memory systems research (which they provocatively call "memory-editing") in clinical contexts:

> One challenge of applying memory-editing research to the clinic is that a memory for a single event can be expressed in several ways, each of which is linked to a distinct neural representation. The depiction of memory editing in science fiction mainly highlights efforts to alter the conscious recollection of life events, which is known as episodic memory. However, a traumatic event— a car accident, for example— produces multiple forms of memory expression. The victim will probably consciously recollect details such as where and how the accident happened. In addition, exposure to an accident cue (such as seeing the street corner where it occurred) may evoke momentary freezing and physiological arousal, or learned defensive responses. The person may also habitually avoid that corner. Finally, reminders of the accident may evoke negative subjective feelings. Although these different forms of memory for the same event (that is, episodic details, defensive responses, habitual actions and subjective feelings) may interact, each involves a

> distinct neural system for storage and expression. For this reason, targeting one type of memory representation for editing may or may not alter other forms of memory for the same event. This specificity can have advantages and disadvantages. For example, it might be advantageous to retain accurate conscious memory for the details of an event, but edit the associated negative feelings or defensive responses. Conversely, it could be problematic to retain defensive responses, habitual actions and negative feelings that are linked to a traumatic event that can only be consciously recalled to a limited extent. (Phelps & Hofmann 2019: 43-44)

Though Phelps and Hofmann's interests differ from ours, their discussion illustrates the assumptions, *mutatis mutandis*, just presented.

The MSP theory is broad enough to admit of a wide array of variations with respect to (1) the number of memory systems it posits (and, therefore, the number of types of attitude it posits), (2) the manner and precise natures of the types of interactions between memory systems, (3) claims about the evolutionary history or evolutionary function of these memory systems, just to name a few.

At several points over this dissertation, I will claim, for example, that there are many more than three memory systems, these memory systems stand in a variety of inhibitory, cooperative, and compensatory relations with respect to others, and that at least some of the memory systems appear to have the kinds of fitness enhancing evolutionary functions that may have led to their being selected for (for a review, see Ferbinteanu 2018). But while I am inclined to make these empirical bets at the present juncture, these claims appear as parts of arguments in support of MSP, rather than falling within the theory's scope. Consequently, the theoretical core of the MSP perspective could turn out to be true even if many of these more controversial positions are false (*e.g.*, if there are fewer memory systems than I am inclined to posit, or if some of the memory systems emerged not in response to specific evolutionary pressures but as a by-product of some other selected for change in the brain). Thus, even if my general conclusions about how we ought to conceive of attitudes are on the right track, there is a great deal of worthwhile debate to be had over which specific MSP model best accounts for the available data. As such, I make no claim here that the version of MSP that I propose in this dissertation is even a close approximation to the final word on the subject.

Amongst its theoretical virtues, the MSP theory offers us a nuanced and much needed alternative to what David Melnikoff and John Bargh (2018) call *dual-process typology* (DPT; see *Table 1*.1). Dating back to cognitive psychological research in the late-1970s (Schneider & Shiffrin 1977), there has been a prevailing tendency amongst cognitive scientists to posit two fundamental categories of psychological process or system: one that is intentional, controllable, unconscious, and resource inefficient (variously labeled Type 2, System 2, the explicit system, controlled processes, reflective processing, *etc.*), and another that is unintentional, uncontrollable, unconscious, and resource efficient (Type 1, System 1, the implicit system, automatic processes, reflexive processes, *etc.*). Melnikoff and Bargh note that dual-process typology continues to increase in popularity and influence despite thirty years of forceful and sustained theoretical objections and a lack of empirical support.

Table 1 Standard Dual-Process typological feature list *adapted from Evans & Frankish (2009)*

System 1/Type 1/Reflexive	**System 2/Type 2/Reflective**
Evolutionarily Old	Evolutionarily Recent
Unconscious, Preconscious	Conscious
Shared with animals	Uniquely (distinctively) human
Implicit knowledge	Explicit Knowledge
Automatic	Controlled
Fast	Slow
Parallel	Sequential
High Capacity	Low Capacity
Intuitive	Reflective
Associative	Rule-based
Independent of general intelligence	Linked to general intelligence

Melnikoff and Bargh are justifiably frustrated by the extent to which cognitive science seems to be clinging to an unsatisfactory theoretical framework given that there seems to be a relatively straightforward solution to the various problems that arise for DPT. This straightforward solution, as Melnikoff and Bargh put it, is "for researchers to rigorously explore each feature of a given process one by one, without making assumptions or drawing conclusions about other features that are not being studied" (*ibid*: 290). If the theoretical value of the DPT framework were merely heuristic, then one could recognize the force of Melnikoff and Bargh's proposal. But it's not; DPT frameworks aid in guiding theory not simply by hypothesizing that various properties tend to cluster together, but by focusing our attention on the various causal mechanisms thought to support these property clusters. Viewed from this perspective, Melnikoff and Bargh's advice— though warranted–– does not provide us with a true alternative to DPT theorizing. As such, we should not expect researchers to abandon DPT *en masse* (given its research–regimenting power) without having an alternative on offer. On my view, the MSP perspective does provide us with such an alternative and Melnikoff and Bargh's advice ought to be implemented in the context of MSP theorizing (see Chapter 5).[1]

Some might be inclined to treat the MSP theory as an eliminativist theory of attitudes. Consider Holroyd and colleagues' (2017) characterization of implicit attitude eliminativism:

> In this section, we introduce an alternative view that, as far as we know, is yet to be argued for[2]. This view is eliminativism about implicit bias. It holds that there is no such psychological kind and therefore no account that attempts to characterize implicit bias as a particular mental state or psychological kind will succeed. On this view, there is no unified phenomenon, with any distinctive set of characteristics, that underpins the behavioral responses found on indirect measures such as [the Implicit Association Test, the Evaluative Priming Task, the Shooter-Detection Task, etc.].

[1] Let me be clear. By saying that MSP theory offers us an alternative to DPT, I do not mean to suggest that MSP theory is inconsistent with the dual-process thesis that cognitive processes can be divided into two: those that possess system 1 properties and those that possess system 2 properties. For all we know, by following Melnikoff and Bargh's advice against the backdrop of the MSP theoretical framework, cognitive science might indeed uncover that there are two types of kinds of memory system; memory systems that tend to subserve system 1 properties and memory systems that have system 2 properties (see Samuels 2009). But the standard DPT approach recommends that we take as our theoretical starting point the assumption that cognitive processes divide into two kinds and to then to use the fruits of one's empirical research in making inferences about the mechanisms that subserve each type of process. The approach that I recommend, by contrast, suggests that the DPT approach has things backwards. Instead, we should take as a starting point the existence of multiple memory systems, each of which subserves a variety of processes, and then systematically investigate the properties possessed by each of these processes whilst making no additional assumptions that, say, a rule-based process cannot operate quickly or automatically.

[2] It is worth nothing that eliminativism about attitudes has in fact been argued for in the literature (see discussion of Schwarz' (2007) view below).

> Rather a cluster of different mental states and processes may produce these responses; and these mental states and processes may also be involved in the production of responses on other measures, such as self-report measures. (Holroyd, Stafford, & Scaife 2017: 13)

If this is all that is meant by 'eliminativism,' then the MSP theory is an eliminativist theory of attitudes: I do deny that attitudes, implicit attitudes, and explicit attitudes are natural kinds. And *if* all psychological or mental kinds are natural kinds, then MSP theory is therefore committed to the denial of the claim that attitudes (*etc.*) are psychological kinds.

Having said that, we should resist reading too much into this kind of eliminativism. MSP theory does not imply that we always speak falsely when we describe a mental state token as an attitude or that such attributions are not truth evaluable. For instance, jade is not a natural kind (because two distinct types of mineral answer to the term 'jade') and yet some minerals are instances of jade and not others. Similarly, even though attitudes do not form a natural kind, because multiple distinct types of mental state are called 'attitudes,' some token mental states are attitudes and others are not. Furthermore, the MSP theory is entirely silent on whether attitudes (implicit or explicit) form some other metaphysically or theoretically important kind (*e.g.*, a *social kind*, a *normative kind*, a *discursive kind*, *etc.*). Indeed, all such issues fall well outside the scope of both MSP theory and this dissertation at large.

At this juncture, it is worth marking a distinction between *ontological eliminativism,* concerning which entities, kinds, or properties *really* exist, and what we might call *scientific practice eliminativism*, concerning whether a particular term ought to be struck, or eliminated, from scientific discourse (*cf.* Griffiths 2004). These types of eliminativism are independent. One can be an ontological eliminativist about attitudes without being a scientific practice eliminativist (*e.g.,* one might both claim that no token mental state *really* has the property *being an attitude* and claim that the term 'attitude' is worth preserving on pragmatic or instrumental grounds) and *vice versa* (*e.g.*, one might both claim that some token mental states really do possess the property *being an attitude* and that, for whatever reason, the term should be struck from our scientific theories on the grounds that its preservation thwarts scientific progress).

Though these positions are often conflated, by attending closely to the considerations that eliminativists offer up in support of their position, it quickly becomes evident that each position is defended on different grounds. For instance, Norbert Schwarz (2007) stakes out a position that is both ontological and scientific practice eliminativist with respect to attitudes. According to Schwarz, what social psychologists have called 'attitudes' are actually fleeting, on-the-fly constructions or judgements that are responsive to situation-specific demands. If he's right (and given a very specific characterization of attitudes according to which attitudes are situationally robust, temporally enduring dispositions to behave in attitude relevant ways), then no state possesses the property *being an attitude* (see *ibid*: 649, 651). Strictly speaking, these considerations are silent on the matter of whether we should discourage the use of the term 'attitude' in the context of scientific theorizing. Schwarz' commitment to scientific practice eliminativism is revealed by his claim that attitude-talk inevitably primes us to think of evaluative judgements as being underwritten by situationally robust, temporally enduring features of persons and, so, to continue to use the term 'attitude' would serve only to systematically thwart scientific progress by leading researchers to ask and, consequently, investigate the wrong questions (*ibid*: 651).

Returning now to the main thread, I have already addressed the question concerning whether, and in what respect, the MSP theory is an ontological eliminativist theory of attitudes. But does MSP theory warrant scientific practice eliminativism? If one assumes that scientific theories should only traffic in natural kind terms, then MSP theory does commit us to scientific practice eliminativism for the reasons already given. But there may be good reason to resist this assumption. It may be that attitude-talk earns its theoretical keep if the attitude kind is, say, an *investigative kind*, where an investigative kind is one that fosters scientific interest and/or engagement (Griffiths 2004). However, if attitude-talk serves primarily to systematically thwart scientific progress by virtue of its being hopelessly ambiguous, then, whatever scientific engagement it fosters, we may nevertheless want to eliminate attitude-talk from scientific discourse (*cf. ibid*: 903). Ultimately, the question of whether attitudes should be treated as a valuable investigative kind is an empirical question. Alternatively, it may be that attitude-talk ought to be preserved on the grounds that it plays an important, if not indispensable, role in our moral theories and practices (*cf.* Doris 2000). For example, though Murphy and Stich (1998) recognize that mental illness is not a natural kind, they suggest that mental illness is an important normative kind on the grounds that is helps us identify ways in which psychological processes go wrong by causing problems for those who have it or those around them. At the end of Chapter 3, I consider the possibility that implicit bias can be treated similarly. Whether attitudes, on the MSP theory, are an important normative kind turns on a suite of issues that fall outside the scope of this dissertation. For these reasons, I will refrain from taking a stand on whether MSP theory commits us to scientific practice eliminativism.

Finally, to what extent is the MSP theory a novel attitude framework? Those familiar with the attitude literature may be forgiven for thinking that the MSP theory is just Amodio and colleagues' *memory systems model of social-cognition* (MSM-SC; see Amodio 2019, Amodio & Berg 2018, Amodio & Devine 2006) under a different name. I must confess that I, myself, sometimes feel as though this is the case. There are several reasons that animate this suspicion. For instance, MSM-SC explains various social cognitive phenomena in terms of a suite of independent but interacting memory systems (see *Figure 1* and *Table 2*). Moreover, this model has already shown to be fruitful in identifying various (what Amodio calls) forms of implicit bias (Amodio & Devine 2006, 2009;

Amodio & Ratner 2011), the neural correlates of these biases (Amodio & Devine 2006, 2009), and the ways in which various memory systems interact to drive various forms of evaluative response (see Amodio 2019 for a review; also, Hackel *et al.* 2015). It's also worth noting that many of Amodio's critiques of the orthodox accounts of attitudes are echoed at various points in the dissertation.

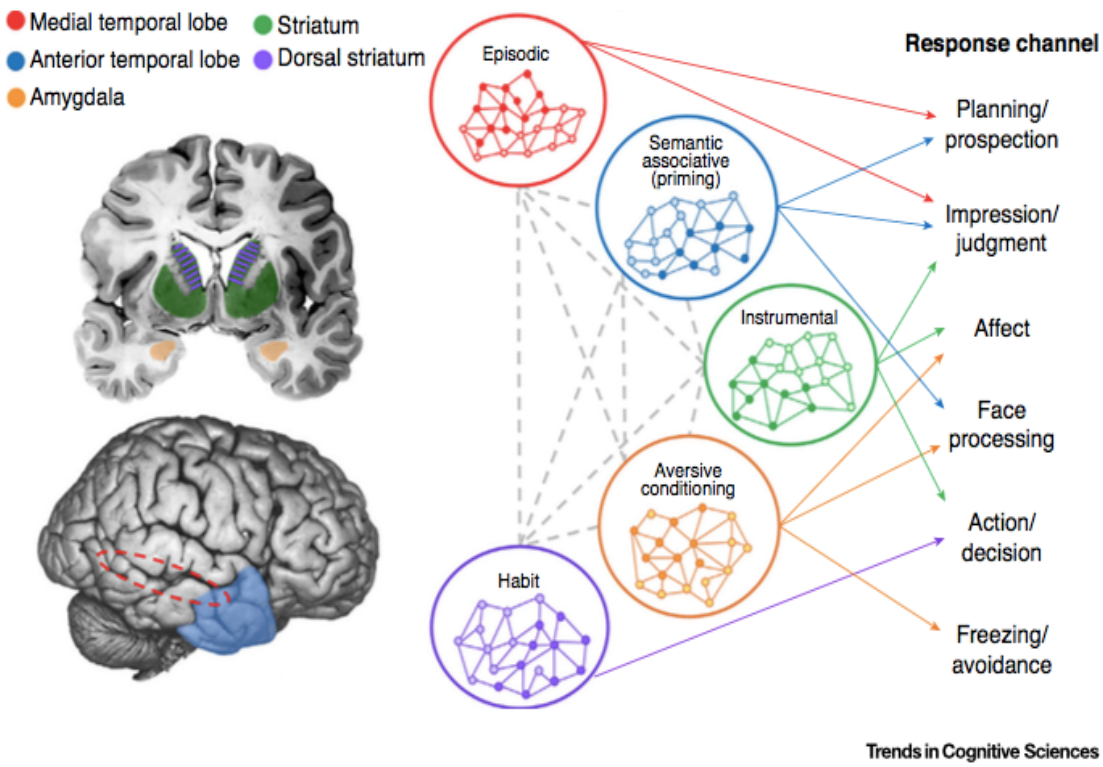

Figure 1. Amodio's Multiple Systems Model of Social Cognition (MSM-SC). *This figure depicts five memory systems, their putative neurological substrates, and their expected roles with respect to influencing the kinds of judgments and behavior most relevant to social cognition research. Some memory systems have been omitted for presentational purposes. On the MMS-SC, the successful prediction/explanation of some social phenomenon requires understanding the joint contributions of multiple-memory systems. The dotted-lines connecting the various memory systems indicate hypothesized between-system interaction, while the solid arrows represent hypothesized causal influence on behavior. Adapted from Amodio 2019*

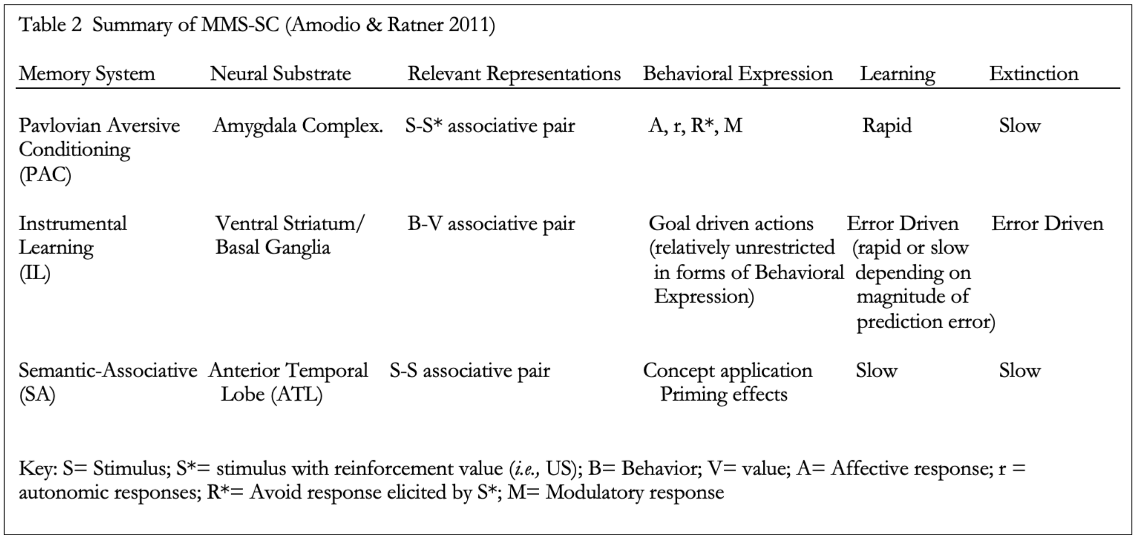

Table 2 Summary of MMS-SC (Amodio & Ratner 2011)

Memory System	Neural Substrate	Relevant Representations	Behavioral Expression	Learning	Extinction
Pavlovian Aversive Conditioning (PAC)	Amygdala Complex.	S-S* associative pair	A, r, R*, M	Rapid	Slow
Instrumental Learning (IL)	Ventral Striatum/ Basal Ganglia	B-V associative pair	Goal driven actions (relatively unrestricted in forms of Behavioral Expression)	Error Driven (rapid or slow depending on magnitude of prediction error)	Error Driven
Semantic-Associative (SA)	Anterior Temporal Lobe (ATL)	S-S associative pair	Concept application Priming effects	Slow	Slow

Key: S= Stimulus; S*= stimulus with reinforcement value (*i.e.,* US); B= Behavior; V= value; A= Affective response; r = autonomic responses; R*= Avoid response elicited by S*; M= Modulatory response

But the MSP theory differs from the MSM-SC in two important respects. First, it is possible that the MSM-SC theory may ultimately turn out to be false and MSP theory true. It is difficult to conceive of how the converse could hold. MSP theory is a theoretical perspective admitting of multiple variations. To the extent that MSM-SC is related to the MSP theory, it is by virtue of being an instance of MSP theory.

This brings us to the second point. It is far from obvious that MSM-SC is committed to attitude pluralism. For instance, Amodio suggests that the MSM-SC model vindicates the tripartite model of attitudes, which treats attitudes as dispositions with affective, behavioral, and cognitive components (see §3 below for a slightly move involved discussion of this view):

> The model of attitudes suggested here— involving separate, interacting representations of conceptual, affective, behavioral associations – may sound familiar. Indeed, the classic tripartite theory in social psychology posits that an attitude reflects some combination of beliefs, affective responses, and behavioral tendencies regarding the attitude object. However, empirical support for these distinctions was hampered by methodological limitations, and the attitude concept

> has become increasingly associated with affect, relative to cognition, and rarely behavior. (Amodio 2019: Box 3)

To be fair, Amodio goes on to say that "the memory systems model described here parallels this classic view, but with some critical updates and differences" (*ibid.*). Nevertheless, if the MSM-SC model is committed to the view that attitudes are dispositions built from distinct components, subserved by the various memory systems, then this would place the MSM-SC directly at odds with the MSP theory, if not empirically, then conceptually (see §3 below).

But let us now suppose, for the sake of argument, that I have merely reinvented the wheel here: Amodio and colleagues really are primarily interested in developing an MSP theoretical account of attitudes. If so, then the value of this work is to be found both in the application of the MSP theoretical framework to questions that have not hitherto been addressed by other MSP theorists and in my defense of it from a range of objections that MSP theorists have not yet responded to. With that said, let's now move on to some of the main predictions of the MSP theory.

2 Salient Predictions of MSP Theory

The MSP theory makes a variety of predictions that collectively serve to set it apart from nearly every other theory of implicit attitudes on the market.[3] Of these various predictions, the three described below are warranted directly by the MSP theory, whereas the others that take center stage

[3] One potential set of exceptions are the trait accounts to be discussed in Chapters 2 and 3, according to which attitudes are traits, or broad-track dispositions of a particular sort, the strength of which are determined by a heterogenous set of psychologically real states and processes. The set of mental states and processes that is assumed to be sufficient for the possession of a trait to a particular degree is sometimes called the "psychological basis" of a trait. If it is assumed that different measures of attitude tap different components of the psychological basis of a given attitude *qua* trait, then these theories can be presented in such a way as to be empirically equivalent to the MSP theory, while disagreeing with the latter about how attitudes are best described. Thus, debate between MSP theory and the trait theory turns on issues that are more conceptual than empirical.

at various points of the dissertation are motivated by considerations over how one should further develop the MSP theory of attitudes so as to best account for the widest range of data made available by both attitude research and surrounding disciplines.

2.1 Low Convergent Validity and High Discriminative Validity Between Measures of Implicit Attitude

The standard view is that differing measures of implicit attitudes tap the same implicit attitude. If this were true, then we should expect any pair of measures to exhibit high *convergent validity* and low *discriminant validity*. Convergent validity is the extent to which two measures that are assumed to be measures of the same construct are actually measures of the same construct. As Krosnick, Judd, and Wittenbrink (2005: 30) state, "To argue for convergent validity, a researcher must show large correlations between different items that are all believed to measure the construct of interest." Discriminant validity is the extent to which two measures that are assumed to be measures of different constructs actually are measures of different constructs. Simplifying a bit, to argue for discriminant validity, a researcher would want to show low correlations between measures that are assumed to be measures of different phenomena.

The MSP theory, however, predicts that differing measures of implicit attitudes frequently tap different attitude constructs. As such, the MSP theory predicts that many measures of implicit attitude, but by no means all, will exhibit low convergent validity and high discriminative validity. In other words, the MSP theory predicts that we will observe low correlations between the various methods that are assumed to measure implicit attitudes. With respect to those measures that do correlate strongly with one another, the MSP theory should have a plausible, independent story to tell about why this would be. Having said this, for reasons to be discussed in Chapter 3, we should not automatically infer that two methods of measurement tap different types of attitude on from correlations alone. Thus, the MSP theory predicts that low correlations between different methods of measurement will be observed when we have good reason to suspect that each method typically recruits different memory systems. I describe the evidence for this claim in Chapters 3 and 6.

2.2 Different Rates of Learning and Updating

The MSP theory predicts that the same or different methods of measurement will reveal differential rates of learning or updating depending on which memory systems are engaged during the experiment's learning phase and which memory systems are tapped during the measurement phase. More concretely, when attempting to measure implicit attitudes, the MSP theory claims that we will sometimes observe rapid learning or updating in some experimental contexts and slow or gradual learning or updating in others. The evidence for these predictions is discussed at length in Chapters 4 and 6.

In a similar vein, because there is no guarantee that the memory systems that are most active during the learning phase of any given experiment are the same memory systems that are recruited during measurement, the MSP theory predicts that there will be occasions on which performance on a particular measurement task administered before the learning phase will not differ significantly from performance after the learning phase despite our having independent reason for suspecting that attitude-relevant learning has occurred. If we are to accurately track the effects of the learning phase on attitudes, this suggests that additional care is needed to ensure that the memory systems engaged during the learning phase correspond, well enough, to those recruited during measurement.

These predictions contrast with those of most competing theoretical perspectives on the market, in that such theories predict that rates of attitudinal learning and/or updating are uniform across experimental contexts. Generally speaking, theories of implicit attitude according to which implicit attitudes are associations predict that implicit attitudes are both acquired and updated gradually in response to new information (Rydell & McConnell 2006, Strack & Deutsch 2004, Wilson, Lindsey & Schooler 2000). Propositional theories of implicit attitude, by contrast, predict both rapid acquisition of attitudes in response to new information and rapid updating under the appropriate circumstances (De Houwer 2014).

2.3 Low External Validity of Attitude Measures

The third prediction requires a little bit of set up. A prevailing assumption in the attitude literature is that implicit attitudes and explicit attitudes sometimes compete for control over behavior. This view assumes that for any token evaluative response, *b*, *b* is such that it can either be the effect of the occurrence of a token implicit attitude or a token explicit attitude. Call this the *domain generality* assumption. For instance, that the behavior in question is a *vote for Bernie Sanders*.[4] A vote for Bernie Sanders, on this view, could among many other things reflect either the tokening of a pro-Sanders implicit attitude or a pro-Sanders explicit attitude. This applies to performance on various attitude measurement tasks as well; in principle, variants of the Implicit Association Test (IAT) can be developed such that IAT performance reflects the activation of explicit attitudes, by, for example, giving participants sufficient time on each trial to make classificatory judgements that are congruent with their explicit attitudes toward the stimuli.

[4] To be sure, some models of attitudes, typically those of the dualist stripe (see §3) break from this assumption. On such views, there is some class of behaviors, B_I, such that behaviors in that can be caused by implicit but not explicit attitudes. These models might also assume that there is some other class of behaviors, B_E, such that behaviors in that class can be caused by explicit but not implicit attitudes. Of course, this is still consistent with the view that there is a third class of behaviors, B_{I+E}, such that members of that class could be caused by either implicit or explicit attitudes. But as this last set of behaviors ranges over far more behavioral types than would be expected on MSP theory, the same concerns raised against the prevailing assumption identified in the main text can be raised, *mutatis mutandis*, for the models of attitudes discussed here.

To be sure, all such views also assume that there are conditions under which implicit attitudes are more likely than explicit attitudes to drive evaluative responses. For instance, Fazio's *Motivation-and-Opportunity-as-Determinants-of-Evaluation* (MODE; Fazio 1992) model assumes that implicit attitudes are likely to drive behavior when motivation to control behavior is low and/or one has no opportunity to control behavior; otherwise, explicit attitudes are likely to be in the driver's seat. But these views are nevertheless consistent with the assumption that any token behavioral outcome, *b*, is such that it can be caused by the activation of either an implicit attitude or an explicit attitude: whether *b* is the effect of implicit attitude activation or explicit attitude activation depends on the specific conditions under which attitude activation or retrieval occurs.

Given the preceding discussion, the standard view predicts that if the conditions under which an attitude is measured mirror those conditions under which a corresponding evaluative behavior is measured, then, *ceteris paribus*, measurement outcome should predict to a reasonable degree the target behavior. To pick an arbitrary example, if Smith's attitude toward Sanders is negative when motivation to control one's responses is absent or the conditions of measurement preclude opportunity to control one's responses and these conditions obtain when Smith is in the voting booth, then the standard view predicts, *ceteris paribus*, that Smith would not pull the lever for Sanders. Similarly, the same measurement outcome should reasonably predict how close Smith decides to sit to Sanders at the inauguration under the assumption that the conditions of measurement obtain when Smith has to pick his seat. This point generalizes for any type of Sanders relevant evaluative response that may occur under those or functionally similar conditions that obtain during attitude measurement (assuming, of course, that the content of the underlying attitude has not undergone substantial change).

The MSP theory rejects the domain generality assumption (see Figure 1 and Table 2) in favor the *domain-specificity assumption*. The domain-specificity assumption places constraints, sometimes severe, on which types of evaluative behavior we should expect to observe given the occurrence of a token attitude of a specific type. Positively stated, MSP theory assumes that different types of attitude drive different types of evaluative response. For example, the species of attitude the tokens of which causally contribute to Smith's performance on a measure of Sanders attitudes that involves speeded categorization judgements (*e.g.*, IAT performance) may be entirely unrelated to the type of attitude the tokens of which causally contribute to approach/avoidance behavior (*e.g.*, decisions of whether to sit near to or far from Sanders at the inauguration). Thus, to consistently generate accurate behavioral predictions on the basis of attitude measurement requires, among other things, that we ensure that the mechanisms involved in attitude measurement are likewise involved in the manifestation of the target behavior.[5]

If this is right and standard methodological practice does not reflect the domain specificity assumption, then we should generally expect to observe modest to weak correlations between, say, implicit racial attitudes and discriminatory behavior. Generally speaking, this is what we find (Forscher *et al.* 2019). Moreover, as that standard methodological practice is just as likely influence criteria governing the inclusion of studies for meta-analytic purposes, the MSP theory likewise expects meta-analyses to reveal low to modest correlations between implicit attitudes and behavior. This is due to the fact that standard methodological practice simply neglects the possibility that the outcome of an attitude measure and a putatively corresponding behavior may be the products of

[5] The consistency hedge is important. It may turn out that the effects of two independent causal mechanisms are such that the obtaining of one effect, E_1, is positively correlated with the obtaining of the other, E_2, under a specific set of conditions, C_1, but not under other conditions, C_2. If so, the presence of E_2 would predict E_1 under C_1 but not C_2. For instance, even if the mechanisms that generate speeded categorization judgements (E_1) are independent of those that generate approach/avoidance behaviors (E_2), we may nevertheless expect that the tendency to classify a particular stimulus as positive under severe time constraints is correlated with a tendency to approach, rather than avoid, the stimulus under some conditions and not others. But one will be unable to *consistently* make accurate predictions if one incorrectly assumes that the relationship between E_1 and E_2 remains constant across C_1 and C_2.

different underlying mechanisms even when the conditions under which attitude measurement and the evaluative behavior occurs are held fixed.

Two upshots of this discussion are worth attending to. First, it suggests that attitude measures may suffer from low levels of external validity not because attitude measures fail to provide us with meaningful information about attitudes (*a la* Oswald et al. 2013, Oswald, Mitchell, Blanton, Jacard, & Tetlock 2015) or because attitudes, implicit or otherwise, are capable only of exerting small effects on behavior (Greenwald, Banaji & Nosek 2015), but because researchers have not based their behavioral predictions on the specific functional profiles of the types of attitude that the standard suite of attitude measures likely recruit. This implies that that attitude measures can be made to better predict behavior with the aid of memory systems research.

Second, it is worth noting that existing research shows that implicit measures are reasonably good predictors of some specific types of evaluative behavior. For instance, we have strong evidence that some measures of implicit political attitudes are good predictors of voting and policy preferences (Friese *et al.* 2012). Moreover, while implicit measures of attitudes toward Obama and explicit measures of attitudes toward Obama each weakly predicted Obama versus McCain voting preferences, predictive power was found to have increased substantially when subjects' implicit measure performance was found to be congruent with subjects' explicit measure performance (*ibid.*). The MSP theory, as with any other, must provide a principled account of both the predictive failures and successes of past attitude research. The MSP theory suggests that order is to be found amongst among the predictive successes and the predictive failures of attitude research, and that this order will be explicable in terms of underlying memory systems and their preferred channels of behavioral expression— and not, as others have suggested, merely in terms of content domains (see Holroyd & Sweetman 2016, Payne, Vuletich & Lundberg 2017).

3 Rival Perspectives

In the attitude literature, one can identify two broad theoretical frameworks that stand in opposition to the MSP theory. On the one hand, *attitude monists* maintain that attitudes are a unified, homogenous kind. Amongst monists, disagreement abounds as to how best to cast the distinction between implicit and explicit attitudes, whether attitudes are stable mental states or behavioral dispositions, and over the extent to which attitude research is continuous with the research interests of those in surrounding fields of cognitive science. On the other hand, *attitude dualists* maintain that attitudes divide into two fundamental kinds, the implicit and the explicit, each of which is treated as a homogenous, unified kind. As attitude dualism appears to have fallen out of favor in recent years, largely due to (misplaced) concerns over parsimony (see Chapter 3), there are comparatively fewer areas of active disagreement amongst attitude dualists. The areas of disagreement that are salient in the literature generally concern issues related to cognitive architecture.

In addition to attitude dualism and attitude monism, there are at least two influential theoretical perspectives that don't fit comfortably within either of the two aforementioned camps. In order to justify their neglect from here on out, it will be useful to say a few words about them here.

3.1 Two Varieties of Attitude Monism

One can identify broad two strands of monism in the literature. On the one hand, there is what Amodio and Ratner (2011) call *the single-system models* of attitudes. These models stand united in their assumption that all attitudes, implicit or otherwise, are encoded, updated, and retrieved from a single long-term memory system. An upshot of single-system monism is that individuals have only a single attitude toward any given attitudinal object. In order to account for the well-documented phenomenon whereby implicit attitudes dissociate from their explicit counterparts, most single-system monist theories (but by no means all) embrace the existence of two types of process that operate over this underlying memory system. One and the same attitude structure stored in long term memory could lead to different evaluations in different contexts depending on how that structure is processed (see Fazio 1990, Gawronski & Bodenhausen 2011, Strack and Deutsch 2004).[6]

Because these views posit the existence of only a single memory system operated on by two types of process, they therefore make contrasting predictions to the MSP theory regarding dissociations of different types of attitude in response to localized anatomical damage, forms of learning that differentially affect different memory systems, and the extent to which different measures of putatively the same construct should correlate with each other. Moreover, none of the single system monist accounts allow for the possibility that, for instance, implicit attitudes may be stored in different representational formats (*e.g.,* associations and propositions).

The other variety of monism requires less discussion. This view conceives of attitudes as evaluative dispositions. The classical *tripartite model of attitudes* conceives of attitudes as dispositions with cognitive-affective-behavioral components (Rosenberg & Hovland 1966). A more recent formulation of this sort of view, articulated and defended by philosopher Edouard Machery (2016), conceives of attitudes as traits, or broad-track behavioral dispositions, which are analogous in various respects to personality traits.[7] While Machery seems to take himself to be articulating a novel

[6] Those sufficiently familiar with the literature may balk at my claim that Strack and Deutsch's (2004) Reflective-Impulsive Model (RIM) is a single-system monist theory. After all, Strack and Deutsch explicitly state that they conceive of their view as a dual-systems account of attitudes. It must be emphasized, however, that the two systems they posit are *processing systems*, not *memory systems*. They posit that a single long-term memory store is operated on by two processing systems. In their own words: "The impulsive system can be thought of as long-term memory, whereas the reflective system has the properties of a temporary storage in that the amount of information that can be represented at any given time is limited, and the representation will fade if it is not rehearsed. […] How such representations [of the reflective system] generated? We assume that the elements of the proposition that is, one or more concepts and the relation that is applied to them— are retrieved from the impulsive system. (*ibid*: 225)" Thus, on our taxonomy, Strack and Deutsch's RIM model is a member of the monist single-system models of attitude.

[7] Importantly, some dispositionalist theories of attitude are monist with respect to dispositions but ontological eliminativist with respect to implicit/explicit attitudes (see Machery 2016; Chapter 2).

perspective of attitudes, this trait-conception has been in currency for quite some time (see Schwarz 2006, 2007; Payne, Vuletich & Lundberg 2017). Importantly, all dispositional models of attitude reject the view that attitudes are best conceived of as mental states or representations. It is also worth stressing a point made in an earlier footnote that dispositional accounts may be empirically indistinguishable from the MSP perspective. As such, disputes between monist dispositional views and the MSP theory may end up turning on conceptual rather than empirical considerations (see fn3).

3.2 Attitude Dualism

Dualist accounts of attitudes assume that individuals may sometimes possess at most two attitudes toward the same attitude object (Wilson, Lindsay & Schooler 2000). It will prove helpful to consider three of the more salient assumptions of this family of views as limned by Wilson and colleagues (*ibid*: 104):

> 1. Explicit attitudes (A_E) and implicit attitudes (A_1) toward the same attitude object can coexist in memory.
>
> [2]. Even when the explicit attitude has been retrieved from memory, A_1 influences implicit responses, namely uncontrollable responses (*e.g.,* some nonverbal behaviors) or responses that people do not view as an expression of their attitude and thus do not attempt to control.

> [3]. Explicit attitudes change relatively easily, whereas implicit attitudes, like old habits, change more slowly. Attitude change techniques often change explicit but not implicit attitudes.

Importantly, while Wilson *et al.* assume that explicit and implicit attitudes toward the same object can coexist in memory, they are silent on whether each type of attitude is stored in a single memory system or are separately stored in different memory systems.

Building on Wilson and colleagues' assumptions, dualist theories of attitudes generally posit the existence of one explicit (/declarative/reflective) memory system and one implicit (/nondeclarative/impulsive) memory system (see Rydell & McConnell 2006, Smith & DeCoster 2000). Each memory system contains different components that govern the acquisition, storage, updating, and subsequent retrieval of attitudes. Moreover, because one can have an explicit, propositional attitude toward attitudinal object X that differs in content and valence from one's implicit, associative attitude toward X, and attitudes are typed differently depending on which memory system they belong to, some dualist accounts of attitude (but by no means all), strictly speaking, endorse the core theoretical commitments of MSP theory. This should not be surprising as, with the notable exception of Wilson *et al.*'s dual attitude account, every other dualist account on the market is directly motivated by the same area of research in the cognitive neuroscience of memory that motivates MSP (see Smith & DeCoster 2000, Rydell & McConnell 2006).

It is also noteworthy that many dualists are open to the possibility that each individual may have many more than two attitudes toward the same object. On this point, Wilson *et al.* state:

> Once we entertain the possibility that people can have more than one attitude stored in memory, why not allow for more than two? Our main argument is that attitude

> models should allow for the fact that people can have both implicit and explicit attitudes toward the same object. It is possible, in principle, for people to have more than two attitudes; people might have one implicit attitude, for example, that is automatically activated, plus two or more newly constructed attitudes that were formed in specific contexts, stored in memory, and retrieved in those contexts. Similarly, people might have two or more implicit attitudes that are activated automatically, depending on context (Wittenbrink, Judd, & Park 1998). Given the empirical challenge of showing that people have three or more attitudes stored in memory at the same time, we have limited ourselves to the case of dual attitudes, with the acknowledgement that this may be a special case of multiple attitudes. (Wilson, Lindsey, & Schooler 2000: 121)

Indeed, the task of showing that people have three or more attitudes stored in memory at the same time *is* a challenge, as the length and complexity of this dissertation attests to. It is nevertheless a challenge that must be undertaken if we are to (a) make substantial progress in attitude research and related fields, and (b) fully integrate this active area of research with the field of cognitive science as a whole.

Due to this theoretical overlap between the dualist and the MSP perspective, disputes between the two perspectives turn primarily on the question of *how many* memory systems one must posit to best account for widest range of empirical results. Dualists posit two memory systems, each of which subserves a species of attitude. MSP theorists generally posit many more than two memory systems. Over the course of this dissertation, I hope to show that MSP theories outperform their dualist rivals not only in terms of predictive/explanatory power, but also in terms of coherence with surrounding disciplines.

3.3 Miscellaneous Perspectives

There are two other *potentially* competing perspectives worth acknowledging as they have proven influential albeit at different times and to differing degrees. These are the *summary representation view* (Krosnick, Judd & Wittenbrink 2005) and the *bias of the crowds* model (Payne, Vuletich & Lundberg 2017). Neither of these views fall squarely in either of the two camps just described. Moreover, the extent to which these views are inconsistent with the MSP theory on empirical matters is unclear. These points are worth elaborating on if for no other reason than to justify my setting them aside for the remainder of this dissertation.

3.3.1 The Summary Representation View

Here's how Krosnick, Judd and Wittenbrink (2005: 24) articulate summary representation view:

> However, we see great theoretical and practical value in resisting [the dual attitude view presented above] and prefer still to hypothesize that a single attitude exists in a person's mind: the net evaluation associated with the object. The observable report of the attitude, representing the integration of evaluative implications at a given point of time, may vary as a function of the specific context in which the integration takes place, but the underlying ingredients from which that report is built (and which constitute the attitude in our formulation) are relatively stable over time.

The net evaluation of an attitude object and the ingredients on which the net evaluation is based may sometimes come apart. The authors illustrate this by way of example:

> For instance, perhaps when you think about your neighbor, you think about the fact that his yard is messy, that he accumulates rusting cars in his driveway, and that he has a couple of dogs that are nuisances. Each of these attributes that you associate with your neighbor tend to have negative evaluative overtones: you generally don't like messy yards, resulting cars, and nuisance dogs. But, somehow, you never integrated these evaluative implications into a net evaluation of your neighbor. In this case, when there is no summary evaluation of the object (*i.e.,* the neighbor), can we really speak of an attitude? We believe that we can, although the latent evaluation is doubly latent. Not only is it not observable by someone who wishes to measure it, but it also never exists as a discrete stored association. Rather, it becomes crystallized only under circumstances that demand a summary evaluation, such as when an overall attitude is demanded by a behavioral encounter (*e.g.,* when you are asked, "So, do you like your neighbor?"). (Krosnick, Judd & Wittenbrink 2005: 23)

These two passages leave us with the following picture. The representations that manifest behavior on the range of indirect measurement tasks are not the attitudes, but the raw materials from which an individual's attitude is fabricated. This process of fabrication (or integration, or crystallization) occurs under circumstances that demand a summary evaluation. Generally speaking, these are circumstances under which subjects are asked to *report* their attitudes. Thus, it wouldn't be unreasonable to infer that the authors prefer to reserve the term 'attitude' to describe that which is manifested during direct measurement tasks— that is, attitudes are what others call 'explicit attitudes.'

In order for us to determine the precise nature of the relationship that the summary representative view bears to the MSP theory, several further questions must be addressed. For instance, does the attitude formed at a given time supervene on the various attitudinal ingredients such that we ought to expect no changes in attitude overtime without corresponding changes to these ingredients? Or are there circumstances such that the attitude formed at t_1, A_1, comes to be stored in long-term memory such that A_1 can be retrieved later on even though the ingredients on which A_1 was formed have undergone substantial change? If the latter is the case, then it is, in principle, possible for one to harbor multiple distinct attitudes toward any given attitudinal object. Additionally, does the attitude formation process always reflect the integration of *all* underlying ingredients or only a proper subset of those ingredients? If the attitude formation process only reflects the integration of a subset of those ingredients, then, again, this view allows formation of multiple, perhaps incongruent, attitudes toward the same object depending on the situation.

The fact that these further questions have not been satisfactorily addressed in the literature renders the relationship between the MSP theory and the summary representation view opaque. Insofar as the MSP theory does not privilege any particular type of representation as *the attitude*, the MSP theory clearly conflicts with the summary representation view on a conceptual level. And, again, such conflicts matter. But, as far as empirical considerations are concerned, it may turn out that the MSP theory can accommodate the summary representation view if it is either the case that summary representations can be stored in a long-term memory system (*e.g.*, semantic memory) for later retrieval and use, or if the process that produces summary representations takes as its inputs the outputs of different memory systems as a function of context. Until these additional questions are addressed, we are left in the dark as to whether the primary point of contention between the MSP theory and the summary representation view is over the appropriate use of the term 'attitude.'

3.3.2 The Bias-of-the-crowds Model

A recent proposal, which has garnered a great deal of attention over the past few years, is Payne, Vuletich and Lundberg's (2017) *bias-of-the-crowds model*. In essence, this view treats implicit bias (which, for our purposes, may be conceived of as a subset of implicit attitudes that drive discriminatory behaviors) as a feature not of persons or individuals but of environments or situations:

> It has long been recognized that a person's context and culture contribute to their implicit biases (*e.g.,* Banaji, 2001). But this formulation still assumes that implicit bias resides as an attitude or belief with some degree of permanence in the minds of attitudes and that the context is simply one kind of input that helps shape those attitudes. […] [B]ut we believe it is more accurate to consider implicit bias as a social phenomenon that passes through the minds of individuals but exists with greater stability in the situations they inhabit. (*ibid*: 236)

This switch in emphasis from what they call a "person-based analysis to a situation-based view" is motivated by findings that implicit bias effects are (1) large and stable when aggregated but small and unstable at an individual level, (2) diachronically stable at an aggregate level but unstable at an individual level, and (3) measures of implicit bias are strongly associated with various race, gender, class disparities when examined at aggregate levels such as nations, states, and metropolitan areas but only weakly correlated with discriminatory outcomes at the level of individuals. If, as the authors maintain, that only the bias-of-the crowds model resolves these puzzles, then we ought to jettison the individual-based approach in favor of the situationist approach.

A thorough discussion of the merits of the argument that they advance in favor of their perspective and the relationship between the bias-of-the-crowds model and the MSP theory would each require its own chapter. Having said that, due to its rapidly growing influence on the field of attitude research, I would be remiss were I not to briefly address both issues.

If Payne *et al.*'s central argument in favor of the bias-of-the-crowds model holds, this would set their perspective at odds with the MSP theory. After all, they do claim that their perspective better accounts for fullest range of relevant findings when compared to the alternatives. However, there are two reasons to think that it doesn't. First, Payne *et al.* assume that attitudes are stable, broad-track dispositions to engage in attitude-congruent evaluative responses (analogous, in this respect, to how personality traits are classically conceived of in social-psychology (Mischel 1968)). Indeed, most of their criticisms of individual-based attitude research lose their force should this assumption be abandoned (for reasons given below). Not only is this conception of attitudes highly controversial, but it is also a conception that the MSP theory rejects on independent grounds.

Relatedly, many of the empirical findings that animate their argument against the individual-based perspective are either findings that the MSP theory straightforwardly predicts or are such that they are consistent with MSP theory. For instance, against the individual-based approach, they cite evidence that (1) "performance on implicit bias tests is malleable in responses to various manipulations in context" (*ibid*: 235), (2) interventions tend not to have long-lasting effects, and that (3) measures of implicit bias suffer from low external validity (*ibid*: 235-236). While such findings may undermine theories according to which attitudes are conceived of as cross-situationally stable, and temporally enduring broad-track dispositions, these findings in no way challenge the MSP theory. In fact, many of these findings support it (see Chapter 3).

Setting aside the argument that Payne *et al.* marshal on behalf of their model, how ought we to think of the relationship between the bias-of-the-crowds perspective and the MSP perspective? Are they competing perspectives of the same phenomena? Are they complementary perspectives, addressing different sets of theoretical and empirical questions? Do they merely operate at different levels of explanation? Though I am not at this stage prepared to take a strong stance on the latter two questions, it would be a mistake at this juncture to think of these two perspectives as being at odds in any substantive way.

To see why, let's begin by noting that the two perspective differ in explanatory scope. Whereas MSP theory is a perspective on attitudes *simpliciter*, Payne *et al.* stress that their perspective is far narrower in scope: "The arguments articulated here apply specifically to the concept of implicit biases toward social groups. They do not question the notion that people hold attitudes in general" (Payne et al. 2017: 237). This suggests that any conflict between the two perspectives, should it arise, concerns how best to understand implicit attitudes toward various social groups. So, in principle, it is possible for MSP theory to be the right theory of attitudes *simpliciter*, but, for whatever reason fail to account for implicit biases in particular.

So, are the views necessarily at odds over their treatment of implicit bias toward social groups? No; Payne *et al.* neither deny that implicit bias toward social groups could exist in principle nor do they deny that this phenomenon can be accurately captured by implicit measures at the individual level:

> Our situationist view of implicit bias does not deny that attitudes exist. If people store evaluative knowledge in memory and that knowledge can be chronically

> accessible, then we would interpret that stability as evidence of dispositions that deserve to be called attitudes. Our model is instead a response to empirical findings that fit poorly with the traditional view where implicit intergroup bias is concerned. Other kinds of implicit evaluations may operate much more as personal dispositions and, therefore, attitudes. For example, Gawronski and colleagues (2017) found much higher test-retest reliability for implicit evaluations of political candidates than for racial groups. Other research has found that individual differences in political attitudes are substantially associated with individual behaviors such as voting and policy preferences (Hawkins & Nosek, 2012; Lundberg & Payne, 2014). These findings suggest that some kind of implicit evaluation reflect stable aspects of the person. (Payne, Vuletich, & Lundberg 2017: 237)

Stated differently, individuals might have stable implicit biases toward social groups, where social groups are typed by political affiliation. But once we allow that individuals can have implicit biases toward social groups (typed on the basis of political affiliation), we should also be interested in having a theoretical account of this phenomenon. This project falls outside the scope of the bias-of-the-crowds perspective, but squarely within the scope of the MSP theory. Unless one can produce a principled reason for why someone's harboring unfavorable attitudes that manifest differential responses toward members of different groups on the basis on political ideology (*e.g.*, Marxists versus liberals) should not count as implicit bias toward a social group, it's far from obvious that the two perspectives offer competing accounts of the same phenomenon.

Indeed, the view that the *bias-of-the-crowds* model offers a competing explanation of individual-level theories of implicit bias is rendered all the more implausible in light of the example that they provide that serves to illustrate their central claim that "implicit bias is a psychological marker of systemic prejudice in the environment" (*ibid*: 239):

> Suppose that Town A has relatively high levels of systemic racism. Housing patterns and schools are highly segregated, and they are correlated with large disparities in incomes. Because of economic inequalities and segregation, crime is concentrated in the mostly poor and mostly minority areas. A person walking through Town A will see mostly White teachers, doctors, and bankers, and mostly non-White service workers operating cash registers and emptying the trash. When police pull over motorists, or criminal suspects described in the local news, they are disproportionately Black or Hispanic. Town B, by contrast, has low levels of systemic racism. Residents know all the same stereotypes as everyone else. But strolling thrown the town, residents are unlikely to see those stereotypes confirmed by inequalities in living conditions and social roles on a daily basis. Because of the difference in reminders of inequality, the average accessibility of stereotypical links will be different in the two towns. Implicit bias will be higher than Town A and Town B. (*ibid.*)

Salient here is the fact that Payne and colleagues *assume* the orthodox, single-system monist account of bias in their explanation of the presence of different levels of implicit bias— that is, they assume that individual-level bias can be cashed out in terms of the chronic accessibility, mediated by situational circumstance, of stereotypes and the like. So, individual-level explanations of how attitudes are acquired and how they exert influence of behavior are still operative even on this model. Indeed, there is no reason why we couldn't swap out orthodox assumptions about the mechanisms that subserve implicit bias for MSP theoretic assumptions. But this is puzzling, to say the least, if the bias-of-the-crowds model were a *competing*, as opposed to complementary, account of the very phenomena that individual-based theories attempt to capture. This suggests that whatever the relation between individual-based projects and situation-based projects, the two perspectives are not at odds.

The above discussion suggests that the two perspectives may indeed be complementary. Since the MSP theory offers an alternative to this overly simplistic picture, the MSP theory has important implications for how an individual's being exposed to the inequalities of the sort prevalent in Town A might continue to exert strong, problematic effects on behavior should the individual relocate to Town B. For instance, on the MSP theory certain types of attitude— *e.g.*, those that do not generally manifest the types of behavior found on the standard stock of indirect measures— are highly resilient to extinction (see Amodio & Ratner 2011). As such, if Smith has acquired such attitudes in Town A and relocates to Town B, it is likely that Smith may continue to exhibit various forms of discriminatory behavior toward minorities in Town B *despite* Smith's no longer harboring the stereotypes *qua* semantic-associations that Smith had formed in Town A. This would be a phenomenon that neither the single-system monist nor this version of the bias-of-the-crowds model could explain, since this type of attitude is neither a semantic association nor, *ex hypothesi*, tracking the kind of situational feature highlighted by Payne *et al.*'s account. All of this serves to highlight the importance of more traditional individual-level theorizing even if, at the end of the day, the macro-level phenomena highlighted by Payne *et al.*'s perspective are not wholly reducible to their individual-level bases (*cf.* Wright, Levine, & Sober 1992).

4 The Strategic Importance of *Implicit* Attitude Monism

Though the MSP theory is indeed a theoretical perspective on attitudes *simpliciter*, it may not have escaped the reader's attention that most of the discussion hitherto concerns *implicit* attitudes in particular. Indeed, I do not return to the nature of explicit attitudes and that construct's relation to implicit attitudes until the concluding chapter. But having described MSP theory's two chief rivals, we are now adequately positioned to see the strategic value of emphasizing discussions of the nature of implicit attitudes over explicit attitudes in motivating the MSP theory.

Single system monists generally claim that all implicit attitudes are associations stored in long-term memory and are operated on by associative processes. (Even those single system monists who maintain that implicit attitudes are propositions, operated on by propositional processes, nevertheless treat implicit attitudes as a homogenous kind (De Houwer 2016).) But dualists do not generally reject this assumption. Instead, they argue that the best models of attitude are those that posit two types of attitude, implicit and explicit. Thus, generally speaking, single system monism and attitude dualism, whatever their other differences, are united in their assumption that *implicit attitudes* are a homogenous, unified mental kind. Call this view, implicit attitude monism (IA-Monism). So, if one could show IA-Monism to be false in a way that supports pluralism about implicit attitudes, then we one would have therefore be warranted in concluding that the MSP theory better accounts for the relevant phenomena over its two main rivals.

5 Chapter-By-Chapter Overview

The following two chapters continue to lay the foundation for the rest of the dissertation. Chapter 2 considers Edouard Machery's argument for the trait view of attitudes, which entails a form of eliminativism about implicit attitudes in particular. Machery motivates the trait view primarily by showing that only the trait view has the theoretical resources needed to account for three sets of anomalous findings, the existence of which have led many to declare that the field of attitude research is in crisis (see Bartlett 2017, Nordell 2017, Singal 2017, Machery 2017). Though Machery takes himself to have shown that the trait view outperforms every account according to which attitudes are mental states, I argue that Machery's argument, at best, succeeds in establishing that the trait theory outperforms theories of attitude that are committed to IA-Monism. At no point does he take seriously the possibility that a pluralist theory attitudes could similarly resolve these anomalies.

Chapter 3 further explains and elaborates the MSP theory of attitudes and shows how it receives direct and indirect support from our best theories of memory in cognitive neuroscience. Moreover, in the process of motivating the MSP theory as a viable alternative to both various forms of IA-monism and Machery's trait view, I consider and respond to various *prima facie* objections to the MSP theory. Once the MSP theory is sufficiently motivated, I then show that we ought to accept the MSP theory over Machery's trait view, or any other dispositional theory of attitudes.

The next three chapters are dedicated to the exposition and tentative resolution of what I dub the *format problem*. Chapter 4 describes three sets of influential findings that, taken together, suggest that implicit attitudes behave *both* as if they were propositionally structured representations *and* as if they were associative structures. I survey three IA-Monist accounts of attitude, *viz.* (i) those on which all implicit attitudes are associations, (ii) those on which all implicit attitudes are propositional structures, and (iii) those on which all implicit attitudes are *sui generis*. I then argue that each account fails to account for the full range of findings that generate the format problem. I conclude that the MSP theory has the best prospects for resolving the format problem and outline the MSP theoretical approach to its resolution that I favor. Simply put, this resolution assumes that implicit measures sometimes tap both attitudes *qua* associations and attitudes *qua* propositional structures.

Chapter 5 argues that propositionally structured representations can influence performance across an array of measurement tasks in ways thought proprietary of associative structures and mechanisms. Importantly, the memory systems that mediate performance on attitude measures co-exist and sometimes compete with memory systems that traffic in associative structures. Thus, we are warranted in treating some types of implicit attitudes as associations and some as propositions. Chapter 6 aims to continue to lay the theoretical groundwork for a full resolution of the format problem introduced in Chapter 4 by developing a format pluralist account of one of the three controversies that animates it.

The concluding chapter, Chapter 7, begins with a summary of the dissertation's main argument, the upshot of which is that there are many types of attitude and that each type is grounded in a functionally distinct, modular, memory system. I then proceed to tie up a number of theoretical loose ends strewn across the dissertation, amongst these is the question of how we ought to conceive of explicit attitudes and their relation to implicit attitudes on the MSP theory.

CHAPTER 2

MACHERY'S ELIMINATIVIST CHALLENGE

Few psychological constructs have garnered as much attention over the past 30 years as the implicit attitude. Reasons for this abound. Sincere egalitarians with respect to race, gender identity, sexual orientation, *etc*., nevertheless often behave in ways that run contrary to these deeply held convictions (Dovidio & Gaertner 1998, Gaertner & Dovidio 1986). If people neither endorse racial stereotypes nor desire to behave as the racist behaves, then what exactly is causally responsible for these behaviors? If implicit attitudes are what drive these behaviors, then developing an adequate theory of implicit attitudes is of great theoretical and practical importance.

And while the study of implicit attitudes has long had its share of critics (Blanton & Jaccard 2006, Tetlock & Mitchell 2009), skepticism over the explanatory value of implicit attitude research seems to have reached a fever pitch both in popular and academic circles over the last few years (for critiques in the popular press, see Bartlett 2017, MacDonald 2017; for academic critiques, see Forscher, Mitamura, Dix, Cox, & Devine 2017, Machery 2016, 2017a, 2017b). This new wave of skepticism is fueled primarily by the perception that social cognition researchers have long ignored several key anomalies. Amongst the most frequently discussed include findings that implicit measures (a) weakly correlate with other implicit measures (Fazio & Olson 2003, Nosek, Greenwald & Banaji 2007a), (b) vary in their test-retest reliability along seemingly trivial dimensions (Cunningham, Preacher & Banaji 2001, Gawronski, Morrison, Phills, & Galdi 2017), and (c) are poor predictors of behavior (Cameron, Brown-Iannuzzi, Payne 2012, Forscher, Mitamura, Dix, Cox, & Devine 2017, Greenwald, Poehlman, Uhlmann, & Banaji 2009). Let's refer to these as the *core anomalies*. Taken together, the core anomalies seem to imply that social cognition researchers have spent the better part of the last three decades using unreliable tools to measure a causally inefficacious *who-knows-what* (Machery, 2017b).

The most forceful articulation of this skeptical position is found in Machery's (2016) "De-Freuding Implicit Attitudes." Machery argues for the position that the *Freudian picture* of attitudes—the picture on which implicit attitudes are mental states—is deeply misguided on the grounds that any social cognitive model that presupposes it is incapable of successfully resolving the core anomalies. According to Machery, the trait view offers the best available explanation of the core anomalies. Moreover, since traits are dispositions and it makes no sense to speak of implicit or explicit dispositions, it follows that there exist neither implicit attitudes nor explicit attitudes. In other words, to claim that attitudes are implicit or explicit is to commit a category mistake. Thus, to endorse Machery's trait view is to commit oneself to a particularly strong form eliminativism about implicit attitudes.

This chapter and the next, when taken together, builds a case both against the prevailing orthodoxy about implicit attitudes and Machery's trait view. This case has both negative and positive dimensions. The aims of the present chapter are largely negative: while Machery is largely correct to claim that the orthodox theories of implicit attitudes are incapable of rising to the challenge of accounting for the core anomalies, the central thesis of this chapter is Machery draws the wrong conclusion about the viability of the so-called Freudian picture of attitudes *simpliciter*. Far from establishing that implicit attitudes are not best construed as mental states, his critique only warrants the weaker position that no monist theory of implicit attitudes will prove to be empirically adequate. In the next chapter, I will develop a pluralist alternative that both (a) is capable of resolving the core anomalies and (b) outperforms the trait view across a number of theoretical dimensions.

In outline, this chapter proceeds as follows. As the introductory chapter makes clear, there is little agreement within the field of attitude research as to what exactly attitudes amount to. This lack of consensus is exacerbated, not alleviated, by the introduction of the implicit attitude/explicit attitude distinction. The first aim of this chapter is to articulate a conception of attitude that makes as few substantive assumptions about the nature of attitudes as possible while still allowing us to evaluate the various theoretical frameworks that have been proposed. Armed with such a conception, I go on to spell out Machery's eliminativist challenge. Along the way, I also describe and defuse some of the recent dialectical moves that proponents of the standard views of attitude make in attempting to head off the eliminativist challenge. Once Machery's challenge has been sufficiently motivated, the final aim of this chapter is to establish that the eliminativist challenge presupposes monism about implicit attitudes, which is a substantive empirical claim that we have independent grounds for rejecting. This sets the stage for the MSP theoretical response to the challenge developed in Chapter 3.

1 How We Should Think About Attitudes

Machery's central negative thesis is that implicit attitudes are not a type of mental state. To many, the negative thesis may sound like a contradiction in terms. What could attitudes *simpliciter* and implicit attitudes in particular be if not types of mental representation? Because controversies over the properties of implicit attitudes extend to those very properties that have been historically used to identify implicit attitudes, we need some way of homing in on the target phenomenon that is both (a) sufficiently topic-neutral so as to not unfairly stack the deck in favor of one substantive account of attitudes over the other, but (b) not so topic-neutral that we lose sight altogether of the very phenomenon central to our inquiry.[8]

[8] The need for such an account becomes more pronounced as soon as it is also recognized that even within the so-called Freudian camp, it is similarly controversial whether (i) implicit attitudes are propositionally structured or associatively structured, (ii) implicit attitudes are always, or even typically, inaccessible to consciousness, or (iii) implicit attitudes are

A quick example will help to illustrate the concern that I have in mind. One venerable account of implicit attitudes is one on which implicit attitudes are semantic associations (*e.g.*, BLACK–CRIMINAL, MALE-SCIENTIST) stored in long term memory and are activated automatically and below the conscious threshold. While this account does delineate a key phenomenon of theoretical interest, our taking it for granted would rule out the trait theory of attitudes by terminological fiat. Accordingly, this account runs afoul of the topic-neutrality desideratum articulated above. Indeed, I am not the first to raise such concerns about the topic-neutrality of such accounts with respect to homing in on the precise nature of implicit attitudes. De Houwer, Gawronski, and Barnes-Holmes (2014) articulate analogous concerns in their influential article in which they develop their meta-theoretical framework of attitudes.

In light of these concerns, I propose a working conception of attitudes in general. We can then use this conception to aid us in distinguishing between implicit attitudes and explicit attitudes. With that said, an attitude is the sort of thing, tokens of which are non-trivially involved in the causal production of a range of so-called *evaluative responses*. Evaluative responses that are of special interest to social psychologists includes, but are by no means limited, to each of the following: a smile, the dilation of one's pupils, a feeling of bodily tension, a choice of teammate just before a game of basketball, a hiring decision, a decision about how likely it is that an unknown other is armed and dangerous, a decision as to which charity to donate to, a swipe-left on a dating app, or the relative speed with which one can identify a positive word after thinking of an attitudinal object. As stated in Chapter 1, these are responses to objects, events, properties, *etc*., that, in most ordinary contexts, reflect how an individual evaluates or appraises it.

acquired and/or expressed outside volitional control. Each of these properties has, at one point or another, been used to identify implicit attitudes and to distinguish them from their explicit counterparts. So, not only will we need a conception of attitudes that is topic-neutral with respect to the issue of whether implicit attitudes are representational structures or dispositions but one that is similarly topic neutral with respect to their "essential" properties.

With this general conception of attitudes in hand, we're now in a position to distinguish between implicit attitudes and explicit attitudes. Roughly, implicit attitudes are the type of entity that causally produce or otherwise explain the range of evaluative responses that are elicited via indirect measurement procedures. Here are two representative examples. The most popular variant of the First-Person Shooter (FPS) task measures the relative difference in error rates and reaction times to shooting unarmed white targets versus unarmed black targets in a videogame simulation. The Black-White variant of the Implicit Association Task (IAT), measures the speed with which a subject can successfully categorize valenced words and Black and White faces on congruent blocks relative to incongruent blocks. The list goes on. By contrast, explicit attitudes are the type of entity that causally produce or otherwise explain the range of evaluative responses elicited or measured by the various paradigmatic explicit measures of attitudes. All such measures involve the self-report. For instance, the Thermometer Task is one on which subjects are to indicate how warmly or coolly they feel toward some social group by indicating their levels of warmth on a digital representation of a thermometer. Aside from the thermometer task, subjects might be asked to fill out a questionnaire that inquires about various thoughts, feelings, and perceptions about one or several social groups. So, according to this working conception of the implicit/explicit distinction, the way in which attitudes are measured are prioritized (for purposes of explication) over the properties of the thing being measured. In this sense, we might think of the conceptions offered here as akin to a species of operational definition. One of the goals of a fully mature science of attitudes is to discard this working conception as scientific knowledge accumulates.[9]

[9] Another distinction frequently drawn in the literature is one between *implicit evaluations* and *implicit stereotypes*. Implicit evaluations are typically conceived of as mental representations that causally mediate positive or negative feelings toward a social object, whereas implicit stereotypes causally mediate attributions of traits, properties, or characteristics to a social object on the basis of perceived group membership (Amodio & Devine 2006, 2009). Having said this, I leave it as an open question as to whether this distinction is real or merely nominal (see Madva & Brownstein 2018).

2 The Eliminativist Challenge

According to Machery, attitudes are best thought of as traits, or multi-track dispositions (*cf.* Schwitzgebel 2002). On this view, to attribute to Don the attitude *dislikes people of color* is to attribute to Don the disposition to avoid people of color, experience negative feelings when thinking about people of color, support policies that tend to affect people of color negatively, and so on. Call the thoughts, feelings, etc. that ground Don's attitude the *psychological bases* of his attitude (Machery 2016: 112-3; *cf.* Carruthers 2013). But because attitudes are traits and traits are not the appropriate objects of introspection, to attribute implicitness or explicitness to an attitude is to make a category mistake. If this is right, then it is hard to imagine a scenario in which purging psychology of implicit/explicit attitude-talk would not thereby improve the quality of psychological research. Thus, if so, then endorsing the trait view is tantamount to endorsing implicit attitude eliminativism.[10]

Machery arrives at the conclusion that attitudes are traits by way of abductive inference. Specifically, he argues that his trait theory of attitudes is better positioned to explain the core anomalies than any view on which implicit attitudes are a distinct kind of mental state (the Freudian picture) (*ibid*: 115-117). Henceforth, I'll refer to this argument as the *eliminativist argument*. If the reasoning holds, then we have good reason to prefer the trait view over any Freudian view. And since the trait view entails eliminativism, we've got good reason to be eliminativists. With the big picture squarely in view, it's now time to get down to the details.

[10] It is important to stress that even though Machery's trait view warrants eliminativism toward implicit and explicit attitudes, it doesn't follow that Machery is an eliminativist with respect to *attitude* (in the narrower psychological sense) *simpliciter*. Indeed, Machery (2016) is inclined to think that (a) a science of attitude is possible (*ibid*: 121), and (b) that traits or dispositions can enter into genuine type-causal relations even if they never enter into token-causal relations (*ibid*:111).

3 The Core Anomalies

3.1 Weak Correlations Across Indirect Measures

The following rule of measurement strikes me as a reasonable one: For any two non-identical measurement procedures *x* and *y*, if *x* and *y* are measures of the same phenomenon, then, *ceteris paribus,* the outcome of *x* will correlate reasonably well with the outcome of *y*. So, I can measure the length of a hallway in steps or I can measure it using a tape measure. Suppose that I apply these two procedures to measuring the lengths of five different hallways. If steps measure the same phenomenon as a tape measure, then the number of steps should correlate reasonably well with the number of meters as indicated by the tape measure. There are, of course, a whole host of caveats here, most of which are taken care of by the *ceteris-paribus* clause, some of which will be discussed here.

It turns out that the various indirect measures of attitude tend to run afoul of this rule. Several studies from the previous decade have produced evidence that indirect measures are either only weakly correlated with one another or are uncorrelated with one another (*e.g.*, Bosson, Swan, Pennebaker 2000, Olson & Fazio 2003, Payne, Govorun & Arbukle 2008, Sherman 2008; for discussion see Nosek *et al.* 2007). For instance, Bosson, Swann, and Pennebaker (2000) found that seven implicit measures of self-esteem were only weakly correlated— the correlations range from –.014 to 0.23). Sherman (2003), moreover, was unable to find any statistically significant correlations between various IATs and evaluative priming. As Machery points out, this is not what the Freudian picture would predict: If all these measures were really tapping the same latent construct, then investigation should reveal stronger correlations between measures (*cf.* Bar-Anan & Nosek 2014: 683). Unlike the Freudian view, the trait view is in a position to explain why indirect measures do not correlate well: The psychological bases of an attitude are heterogenous, and different measures tap different components of these psychological bases.

Brownstein and colleagues (2017) offer two quick responses to Machery's claim that weak correlations across indirect measures cannot be accommodated by standard theories of implicit attitudes. Their first response is that the strength of inter-measure correlations has often been shown to be a function of the content of what is measured. For instance, Bar-Anan and Nosek (2014: 683) "found moderate to strong relationships among seven indirect measures in at least two topics (politics and race) and poor relationships between the measures in one topic (self-esteem)." All this shows, according to Brownstein and colleagues (2017), is that "the weakest correlations [were] found when the most complex targets [were] targeted." If this is the correct interpretation, then this pattern of results ought to be expected.

However, it is not at all obvious that self-esteem is more complex *qua* psychological construct than is race *qua* psychological construct, partly because the authors say nothing about how the term 'complexity' is to be understood in this context. Here are two possibilities. The first is that probing self-esteem generally involves the probing of several qualitatively distinct self-esteem relevant constructs, whereas probing racial attitudes generally involves the probing of comparatively fewer racial-attitude relevant constructs. While racial implicit attitudes are traditionally thought of as associations between race and valance, self-esteem implicit attitudes are sometimes treated as consisting of *evaluative* attitudes toward oneself (where evaluative attitudes, in this context, are associations between oneself and a trait) and *affective* attitudes (where affective attitudes, here, are associations between oneself and a valance) (see Greenwald & Farnham 2000). But this difference is contrived, as there is no reason why we should not also allow that probing one's attitudes toward race also generally involves probing one's conceptions of racial groups (*e.g.*, stereotypes). A second possibility— though, not one that has been articulated in the literature— is that people's attitudes toward themselves are, generally, more ambivalent and so more complex than are people's attitudes toward racial groups. This hypothesis is likewise problematic. We know that people's implicit evaluations of racial groups are highly context dependent such that subtle differences in context are sometimes correlated with large differences in evaluation (see §1.2.2). So, we have good evidence that people's attitudes toward are racial groups are ambivalent, which further implies that we have no reason to think that racial implicit attitudes admit of less ambivalence than implicit attitudes of self-esteem.

Brownstein and colleagues' second response is as follows. Everyone should admit that no widely used indirect measure is *process pure*, which means that there is no one-to-one correspondence between measures and the states or processes that they tap. For instance, in implicit social cognition research, one's performance on any arbitrarily selected indirect measure is assumed to involve the contribution of several distinct cognitive processes, some of which are paradigmatically explicit and some of which are paradigmatically implicit. Moreover, there is good reason to expect that performance on some measures will be more heavily influenced by paradigmatic explicit processes than implicit ones. To illustrate this, consider the structure of the Implicit Association Test (IAT; Greenwald, McGhee, Schwartz 1998) and the Affective Misattribution Procedure (AMP; Payne, Cheng, Govorun, Stewart 2005). On stereotype incongruent blocks of the IAT, subjects must override their initial inclination to use the response key associated with negative words when categorizing a Black face in order to produce the accurate response as quickly as possible. But the AMP works quite differently. In the AMP, subjects are shown pictures of neutral stimuli (*e.g.*, Chinese characters) that are preceded by pictures of members of the target category (*e.g.*, Black faces). Subjects are then told to ignore the primes and classify only the neutral stimuli as pleasant or unpleasant. Whereas the dependent variable of the IAT is reaction time (RT), the dependent variable of the AMP is a judgment of the relevant object as being pleasant or unpleasant. So, even if both measures are designed to assess whether the subjects implicitly associate blackness with negativity, they likely do so by way of tapping different cognitive processes. As such, we ought to expect the performance on IAT weakly correlate with performance in the AMP even if they are both measures of the same underlying implicit attitude.

The problem here is that Brownstein and colleagues show only that it is possible for two measures to correlate weakly even if they are influenced by the same state. However, they do not provide us with any reason to think that such measures are generally influenced by the same kind of state as opposed to a range of different states and processes. And it is here where Machery's position has the advantage. On Machery's (2016: 119) view, an attitude's psychological basis "encompasses good old-fashioned mental states and processes such as emotions, self-control, and so on." If these measures are sensitive to the interactions of these more familiar states and processes, then we're left with no reason to posit implicit attitudes. So, without reason to think that these various indirect measures are sensitive to a single kind of mental state—*viz.* the implicit attitude—Ockham's' razor would seem to favor eliminativism.

3.2 Unexplained Variability in Test-Retest Reliability

A now common criticism of implicit social cognition research is that its measurement tools are simply too unreliable to be of much use (Bartlett 2017, Machery 2017a, 2017b). While this is a bit dramatic, current estimates of reliability tend to be lower than is desirable. A recent meta-analysis of studies that report test-retest correlations places the reliability of the IAT somewhere in the vicinity of .42 (Gawronski, Morrison, Phills, & Galdi 2017). Other indirect measures have fared slightly worse. The same study estimates that the test-retest reliability of the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart 2005) is .38 (Gawronski *et al.* 2017). This has raised concerns as low test-retest reliability implies that a score obtained via indirect measurement at t_1 is likely to be a poor predictor of the same person's score at t_2 (Cunningham, Preacher, Banaji 2001). For instance, the IAT's score of .42 implies that less than 20% of the variability in implicit bias can be explained by one's levels of implicit bias in the weeks prior (Payne, Vuletich, Lundberg 2017: 234).

But perhaps low test-retest reliability of indirect measures is not so worrisome as Machery makes it out to be. Low test-retest reliability would only be a major concern for implicit social cognition research were it also the case that most indirect measures also suffered from low or unacceptable levels of *inter-item consistency*. But since indirect measures exhibit acceptable levels of inter-item consistency (Gawronski, Morrison, Phills, Galdi 2017), low test-retest reliability is probably attributable to changes in whatever is being measured. Consider the following example. Suppose that you are taking an AMP that is designed to measure your feelings toward men kissing. This version of the AMP has high inter-item consistency if your classification of the neutral stimulus as, say, pleasant after being exposed to the men-kissing prime at t_1 is strongly correlated with your classification behavior across all other relevant trials. The strength of inter-item consistency within a given procedure, then, is evidence of the degree to which a measure is tracking something systematically. So, a test with low test-retest reliability but high inter-item consistency strongly suggests that differences in outcome over time are at least sometimes attributable not to measurement error but to a change in whatever it is that is being measured (*e.g.*, your implicit attitudes about men kissing) (Gawronski et al. 2017, Payne, Vuletich, Lundberg 2017).

The hypothesis that changes in what's being measured over time are responsible for the low estimates of test-retest reliability (as opposed to poor psychometric attributes of the measure) gains additional support from the finding that one can experimentally manipulate test-retest reliability *via* the manipulation of *prima facie* irrelevant contextual features. For instance, in a series of studies conducted by Gschwender, Hoffman and Schmitt (2008), implicit attitudes towards Germans and Turks were measured using an IAT over a two-week period. In one condition, the stimulus items (*e.g.*, the words and faces to be categorized) appeared against a screen-filling background of an image of a garden. In another condition, the same items appeared against a screen filling background of an image of a mosque. The temporal stability (*i.e.*, test-retest reliability) of the attitudes in the garden condition was quite low (.29) whereas the temporal stability of the attitudes in the mosque condition was a comparatively strong .72 (*ibid*: 76). The lesson here appears to be that the test-retest reliability of any given measure is partly a function of context. As such, holding fixed the relevant contextual variables on any measurement occasion should boost a measure's test-retest reliability.

So, it appears as though test-retest reliability fluctuates according to seemingly irrelevant situational factors. While this might, *prima facie*, be construed as evidence that substandard test-retest reliability can be explained away as an artifact of past failures to control for the right situational variables, too much unexplained cross-situational variance presents a problem for the standard narrative if it is unable to predict/explain such situational volatility in a principled, non-*ad hoc* manner. And, indeed, there is a tremendous amount of context-dependency to account for, much of it unexpected (see Dasgupta & Greenwald 2001, Schaller, Park & Mueller 2003, Rudman & Lee 2002; for surveys, see: Blair 2002, Machery 2016).

So, why think that the trait picture can better account for the context dependency of implicit attitudes than the so-called Freudian picture? While Machery (2016: 119) concedes that the Freudian picture explains contextual-dependence by appealing to the fact that the strength of mental state activation is often a function of context, he asserts that the trait picture is better equipped than the Freudian picture to make sense of the "nature of this context-dependence."

> The Freudian picture postulates a new kind of mental state—implicit biases—and as a result it is unclear why some contextual factors rather than others (why darkness?) would influence their activation. By contrast, the trait picture hypothesizes that attitudes depend on psychological bases that encompass good old-fashioned mental states and processes, such as emotions, self-control, and so on, and that indirect measures tap into some of these components… [B]ecause darkness heightens stress in humans… it should modulate implicit association test scores if this test measures, or is somehow influenced by, stress. (*ibid.*)

In other words, the outcomes of indirect measures seem to co-vary with situational variables in ways that implicate good old-fashioned mental states and processes. And if these are the states and processes that indirect measures are tapping, then one should be able generate reasonable predictions as to how the manipulation of some situational feature will influence the outcome of an indirect measure. By contrast, while the Freudian view can accommodate context-dependence, it has had trouble generating theoretically grounded predictions concerning how and/or when changes in context should influence performance on indirect measures.

One might be tempted to reply that most models of implicit social cognition can accommodate a great deal of the relevant context dependency by way of *concept accessibility* (see: Gawronski 2017, Payne *et al.* 2017). Very roughly, the idea here is that certain contexts serve to make some concepts or representations in memory more accessible. So, being in a dark room may serve to make one's *danger* concept more accessible and therefore, *via* spreading of activation principles, any other concepts that it is strongly associated with (Schaller, Park, Mueller 2003). As such, we would expect being in a dark room to facilitate one's responses to danger-related stimuli relative to being in a well-lit room. Thus, there is nothing surprising or unexplainable, from the perspective of the standard narrative, about the kind of context dependency that Machery is concerned with.

But while this notion of accessibility is perfectly legitimate, its employment in this context is problematic. With no plausible principle of situation individuation in hand, theories that explain context-dependence in terms of concept accessibility do so at the risk of becoming vacuous. After all, one can always appeal to differences in situational context to explain away recalcitrant findings, as any two measurement occasions will differ with respect to at least some of their features. By the same token, since any two measurement occasions are likely to have at least some properties in common, it is all too easy to explain enhanced temporal stability by appeal to sameness of situation. In other words, without a plausible theory about how situation tokens ought to be typed and/or a theory about which types are most relevant given one's research interests, any scientific theory that gives the situation a starring role is likely to be of dubious predictive/explanatory value (see Machery 2017a for a similar argument). Consequently, models that rely primarily on concept accessibility to account for context-dependence are often limited to *post-hoc* rationalizations. Worse still, theorists who depend on concept accessibility to explain context-dependence have given us no reason to think that a plausible theory of situation individuation is forthcoming. At best, then, such explanations are no more theoretically illuminating than Machery's similarly unhelpful claim that such context-dependence can be explained by appealing to "good old-fashioned" mental states.

3.3 Predictive Validity of Implicit Measures

In the very first paragraph of this chapter, I explained that a primary source of value of implicit bias research is that it promises to explain not only why people engage in various forms of discrimination, but also how it can be true of people who sincerely endorse egalitarian values. It is for this reason that the findings reported in several widely-discussed meta-analyses (Greenwald, Poehlman, Uhlmann, & Banaji 2009, Oswald, Mitchell, Blanton, Jaccard & Tetlock 2013, Forscher, Lai, *et al.* 2019) have caused some long-time supporters of implicit social cognition research to lose confidence in the field's central constructs and methods. Each of these meta-analyses found that performance on an indirect measure is not highly predictive of discriminatory behavior. Correlations between indirect measure outcomes and discriminatory behavior range from .14 to .28.

In a recent and highly influential meta-analysis, Forscher and colleagues (2019) report that attempts in the literature to change behavior by changing the contents of implicit attitudes largely fail. The idea is that if Smith implicitly associates members of a marginalized group with incompetence, then this implicit attitude might cause Smith to spend less time looking over a member of that marginalized group's application in a fast-paced environment even if Smith were to disavow this stereotype sincerely and publicly. To test this, one might first change the content of Smith's association—for example, by "training" the subject to associate this group with competence by having the subject assent to lots of statements that express counter-stereotypes (*e.g.*, Kawakami, Dovidio, Moll, Hermsen, Russin 2000)— and then have Smith take another indirect measure to see whether this change takes hold. Assuming it does, one might then give Smith another round of applications to look over in order to see whether this change in implicit attitude brings about a corresponding change in behavior. What Forscher and colleagues (*ibid.*) have shown is that even when the content of an implicit attitude changes, it more often than not fails to accompany a corresponding shift in discriminatory behavior (*e.g.*, Smith no longer implicitly associates members of the marginalized group with incompetence, but continues to spend less time reviewing a stigmatized group member's application than Smith does with the applications from members of non-marginalized groups).

To be sure, Machery (2016) does admit that the Freudian view is able to accommodate the fact that implicit measures fail to perfectly predict behavior. After all, even proponents of the conventional Freudian views can, and do, admit that every indirect measure is noisy and that implicit attitudes are only one of many causal determinants of discriminatory forms of behavior. Nevertheless, Machery claims that poor predictive validity of indirect measures coheres better with the trait view than the Freudian view:

> On the Freudian picture of attitudes, indirect measures of, for example, the attitude toward black people tap a single mental state, namely, the implicit attitude toward black people. By contrast, on the trait picture, a particular indirect measure taps into one of the many components that determine behavior, such as emotions, associations between concepts, and so on. So, where the Freudian picture posits a single determinant of behavior (*i.e.,* the implicit attitude), the trait picture posits many (*i.e.,* the components of the psychological basis of an attitude). An indirect measure should be worse at predicting behavior on the trait picture than on the Freudian picture (in fact it should be a poor predictor) because it only measures one of the many components of the psychological basis of an attitude... (*ibid*: 120)

But, again, maybe the low-predictive validity of indirect measures is not so troubling as it may first appear. Gawronski (2018) points out that the meta-analyses most responsible for fueling skepticism toward implicit social cognition research report *zero-order correlations* between indirect measure performance and discriminatory behavior.[11] But since no serious theory of implicit attitudes

[11] A *zero-order correlation* is a correlation between X and Y such that no further variable, Z, is fixed. By contrast, a *partial correlation* is one that tracks the relation that X holds to Y given that some further variable Z is fixed.

predicts strong zero-order correlations between implicit attitudes and discriminatory behavior, the presence of low zero-order correlations fails to threaten the empirical adequacy of any of these views (Gawronski 2018). Indeed, Gawronski (2018) predicts that once the standard theoretically derived moderators are taken into consideration, future meta-analyses would reveal stronger correlations. For example, consider the widely held view that implicit bias is particularly likely to influence behavior when one is either exhausted or under severe cognitive load (Fazio, 1990). In order to determine whether past empirical research has supported this hypothesis, one should only include studies that use, say, IAT results to predict behavior while controlling for cognitive load. Past meta-analyses have generally not controlled for these moderators. Consequently, their findings are of little use when evaluating specific models of implicit attitudes.

The problem is that predictive power remains low even after the standard moderators (*e.g.*, controllability, awareness, correspondence) are accounted for (Kurdi, Seitchik, Axt, Carroll, Karapetyan, Kaushik, Tomezsko, Greenwald & Banaji 2019).[12] The researchers found that implicit attitudes are about as strongly correlated with spontaneous, uncontrollable behaviors as they are with more deliberate and controlled behaviors (*ibid*: 14). But this finding sits uncomfortably (to put it mildly) with the dominant narrative of implicit bias on which implicit attitudes are likely to exert more influence on spontaneous uncontrollable behaviors than on deliberate and controllable behaviors (Fazio 1990). In an attempt to explain this surprising finding, the authors suggest that social cognition researchers

> have a harder time generating good intuitions about the mechanisms of implicit cognition. One can see this in current theorizing that stems from dual-process

[12] *Correspondence*, here, is a term of art. Cognition and behavior *correspond*, in the relevant sense, to the extent that the cognition and behaviors being measured are effects of the same causal process (see Azjen & Fishbein 1977)

> accounts and posits that indirect measures capture unconscious cognition and should predict automatic behaviors, whereas direct measures capture conscious cognition and should predict controllable behaviors. This thinking may need to be relinquished (Kurdi *et al.* 2018: 14).

If Kurdi and colleagues' diagnosis is accurate, then there is good reason to suspect that Machery's trait theory will be able to better account for these findings than dual-process theories of implicit attitude. And since nearly every extant theory of implicit attitudes is a dual-process theory of implicit attitudes, it follows that there's good reason to suspect that Machery's trait theory will better accommodate the relevant data than nearly any extant theory of implicit attitudes.

3.4 Taking Stock

Implicit social cognition researchers have long assumed that implicit attitudes are mental states. Machery argues that a theory on which attitudes are traits better explains the core anomalies than theories on which implicit attitudes are mental states. Because Machery's trait view entails eliminativism about implicit attitudes, this is bad news for any theory that posits implicit attitudes. Worse still, his argument appears to have withstood the first wave of criticism.

4 Limitations of Machery's Argument

Machery's eliminativist argument is not as solid as it may appear in light of the above discussion. As Machery, himself admits, the target of his critique is not the Freudian view *simpliciter*, but a specific (albeit widely held) version of the Freudian view on which implicit attitudes are a single, unified kind of mental state (*i.e.*, *monism*). Consequently, even if the above discussion provides us with reason to prefer the trait view over monism, it does not thereby follow that Machery has given us reason to prefer the trait view over *pluralism*. But even if Machery does not have pluralism in his crosshairs, it may nevertheless be the case that Machery's objections to monism also undermine pluralism *via* shared commitments. In this section, I argue that this isn't the case: Machery's arguments against monism land precisely because he is successfully able to exploit monist-specific theoretical assumptions about the nature of implicit attitudes.

4.1 The Eliminativist Argument Exploits Monist-Specific Assumptions

There's good reason to think that Machery treats 'Freudian view' as a generic label for any monist theory of implicit attitudes. That he deploys the label in this manner is on full display in the following passage:

> Some social psychologists deny that there is a single mental state that is people's implicit attitude. Rather, on their view, behavior is caused by many different kinds of mental state, some of which are conscious while others are not, some of which are associative while others are propositional, some of which are controllable while others are not, and so on. This view is not the target of the present chapter. In many respects, it is similar to the trait view of attitudes developed in Section 2. For the target of this chapter, each (either implicit or explicit) attitude is a unitary mental state. (Machery 2016: 107n5)

While I think his decision to ignore pluralist theories is ultimately a strategic error, it is nevertheless a reasonable one to make for the following reasons: First, despite Machery's claim to the contrary, no social psychologist has articulated or defended such a view in print—at least, not by the time that Machery's article had been published. Because no such pluralist model had been developed, there was no specific proposal to criticize. Second, monism was and continues to be the received view of implicit attitudes. As such, it makes a great deal of sense to motivate the trait view by positioning it against the dominant theories of implicit attitude. Third, if pluralism is sufficiently similar to the trait view, it may make little sense to argue against it. With respect to this third reason, Machery overestimates the similarity between pluralism and the trait view (as will be demonstrated in §4).

Having said this, the fact that Machery's central target is monism is not by itself sufficient to establish that his argument does not tell against any potential pluralist position. What remains to be seen is that his argument in fact exploits monist-specific theoretical assumptions about the nature of implicit attitudes. If this can be shown, then this creates an opening for a pluralist Freudian picture of implicit attitudes.

In order to make plausible the claim that Machery's argument exploits monist-specific theoretical assumption, we need to take a closer look at monism. Doing so will also help us to better locate pluralism in theoretical space. Monism falls out of what I call *the standard model of implicit social cognition* (hereafter, SM). We find two general commitments at the theoretical core of SM:

(SSP) *Single-System perspective*: Implicit attitudes are informational states stored in a single long-term memory system.

> (AP) *Associationist perspective:* Long term memory is an associative network (consisting of many interconnected nodes) that operates in accordance with associative principles.

These two theses are so widely held that it is far easier to point out the few models that reject them than it is to list all of the models that subscribe to them. For instance, at the time of writing this, SSP has been explicitly incorporated into every theoretical model (influential or not) of implicit social cognition with the sole exception being Amodio and colleagues' multiple-memory systems model.[13] As for AP, only two extant theories unequivocally reject it (De Houwer 2014, Mandelbaum 2014) in favor of a propositional perspective.[14]

One can see these commitments at work in Strack and Deutsch's (2004) influential Reflective-Impulsive Model (RIM) of social cognition. For our purposes, this model stands out for two reasons: The model's commitments are spelled out with admirable clarity, and it is highly representative of the most frequently encountered models in the literature.

RIM posits the existence of two cognitive systems that operate in parallel: the Reflective System and the Impulsive System (see **Figure 2**). The Impulsive System is posited to explain phenomena characteristic of implicit cognition while the Reflective System is posited to explain phenomena characteristic of explicit cognition. The Impulsive system is assumed to operate on an associative memory store and this store is treated as containing many elements which are variously related by *associative links* (*ibid*: 223). Each element represents a concept in semantic memory, and the

[13] To forestall any potential confusion, a commitment to SSP does not preclude one from hypothesizing a distinct memory system that stores *explicit* mental states.

[14]Gawronski and Bodenhausen's (2018; see also Gawronski, Brannon and Bodenhausen 2017) most recent elaboration of their Associative-Propositional Evaluation (APE) model of social cognition might be read as rejecting AP despite explicitly claiming that all information is stored associatively in a single associative network. This is because they view the association/proposition debate as merely a terminological dispute.

activation of one element results in the *spreading of activation* to other associatively linked elements where the activation strength is proportional to the strength of the link. Here's Strack and Deutsch (2004) on some of the other properties of this system.

> In general, links are created or strengthened if stimuli are presented or activated in close temporal or spatial proximity. The resulting links reflect correlations between aspects of the environment and cognitive, affective, or motor reactions, without representing the causes of such multimodal correlations… In essence, we assume the associative store of the impulsive system works like a simple memory system (see Johnson & Hirst 1993), which slowly forms enduring, nonpropositional representations of the typical properties of the environment over many learning trials (see McClelland et al. 1995, Smith & DeCoster 2000). The impulsive system has low flexibility but is fast and needs no attentional resources. (Strack & Deutsch 2004: 224)

Implicit attitude monism (IA-Monism) seems to fall out of this model, given its assumption that attitudes are just the informational states stored in this associative system that are encoded, updated, and extinguished according to classical associationist principles.[15]

[15] A word about terminology. Broadly, there are two theories of memory storage assumed by various models of implicit social cognition. There is the classical *file-cabinet* conception on which individual memories are like discrete sheets of paper that can be stored in a file cabinet only to be independently accessed at a later time through search and retrieval processes. On this view, the memories like the sheets in the file cabinet are there even when they are not being accessed. And then there is the *connectionist conception* on which information is stored across the weights of a network of nodes and representational states are patterns of activation over a set of interconnected nodes. So, in contrast to the file-cabinet view, there is a sense in which information is not there when the nodes are not in an active state. Unless explicitly noted otherwise, I'll use the term 'storage' permissively so as to remain neutral between these interpretations. This permissive use is appropriate even in describing the RIM model as the RIM model is not committed to a connectionist architecture despite the appropriation of connectionist concepts and terminology.

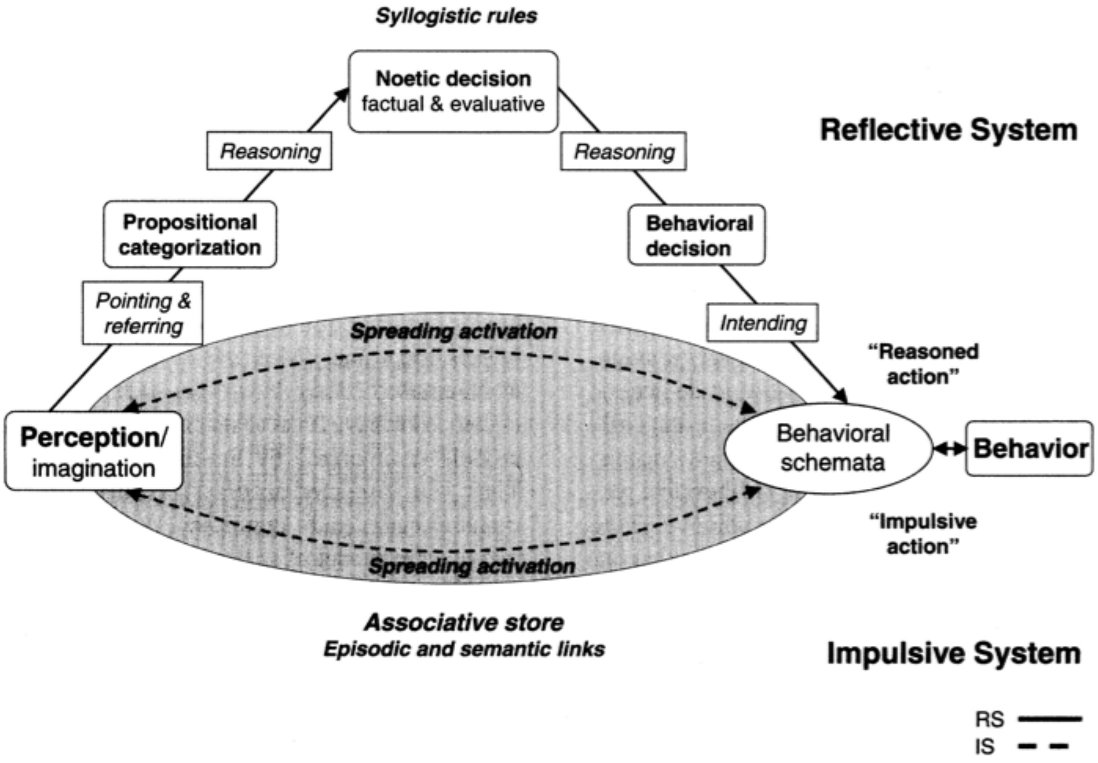

Figure 2. Strack & Deutsch's (2004) Reflective-Impulsive Model (RIM) of social cognition. Solid lines indicate reflective processes while broken lines indicate impulsive processes. RIM is a highly representative instance of the *standard model* of implicit social cognition in that the model contains a single associative memory system that contributes to the production of implicit attitude-relevant behavior. Adapted from Strack & Deutsch (2004).

It's not difficult to see how Machery's argument gains traction by exploiting the assumptions of SM. A tremendous amount of data has been produced by implicit social cognition research, and much of this data is surprisingly messy. Each of the core anomalies serves to drive this point home by highlighting a different aspect of this mess. Machery's trait view has the upper hand here precisely because his view distributes the explanatory labor across the heterogeneous set of states and processes that constitute the psychological bases of attitudes. In other words, his view is well-positioned to account for the messiness of the phenomena because the psychological bases of traits are themselves messy. By contrast, SM and its specific variants promise to impose order on the phenomena by appealing to only two kinds of cognitive process and a single implicit memory store. And despite the fact that SM has been the dominant perspective for over three decades, it does not look as though it has come any closer to fulfilling this promise. The goal of what follows is to develop a pluralistic account of implicit attitudes that has all of the trait view's virtues and none of its vices.

4.2 A Potential Response

One might complain that the verdict that Machery says nothing that impugns the viability of pluralism has been arrived at too hastily. After all, Machery (2016) does appear to have pluralism in mind when he discusses potential responses to claim that Freudian views are unable to account for the weak relationships that obtain between indirect measures:

> […] a proponent of the Freudian picture may reply that there are in fact several distinct implicit attitudes corresponding to the different indirect measures. This response would be a stark departure from the usual description of implicit attitudes. It is also bad scientific practice to postulate a theoretical entity for every measure. (Machery 2016: 117)

It looks as though Machery is rejecting a version of pluralism. If so, then maybe his arguments do undermine both monist and pluralist versions of the Freudian view.

But there are two problems with this response. The first is that the position that Machery attributes to the hypothetical proponent of the Freudian picture is ambiguous between two readings. On one reading, Machery is attributing to the Freudian the view that for each indirect measure (*e.g.*,

M_1, M_2, M_3… M_n), it taps a different token implicit attitude (*e.g.*, c_1, c_2, c_3…c_n) belonging to the same implicit attitude type C. The token reading implies *token-pluralism* about implicit attitudes, which can be treated as the view that for each implicit attitude with content C, there are many token implicit attitudes (individuated by their non-semantic properties) that have C as their content within a single mind-brain. While I do endorse token-pluralism (see Rupert 2011 for a defense for a general view of this sort), it's an altogether different view than the one I aim to develop and defend here. As such, if Machery *does* have in mind the token-reading, then he neglects the sort of view that I'm interested in developing here as an alternative to monism.

On the second reading, Machery attributes to the Freudian the view that each measure taps a different type of implicit attitude (call this the *type-reading*). This objection, as I am understanding it, asserts a one-to-one correspondence between measurement tasks and types of implicit attitudes. Consequently, in order for each indirect measure to tap a different kind of implicit attitude, it must be possible that two token implicit attitudes share the same representational content yet differ in kind by virtue of their having different, say, functional properties (call this view, *type-pluralism*). While the type-pluralist may sometimes want to claim that different indirect measures tap different types of implicit attitude in order to account for weak correlations across measures, this is not the only kind of explanation available to her. For instance, while it may be that each indirect measure taps a different kind of implicit attitude, it may also be the case that different indirect measures exhibit greater sensitivity to certain species of implicit attitude relative to others. According to the latter hypothesis, while M_1 and M_2 both tap the same set of implicit attitudes, say, IA_1, IA_2, and IA_3, M_1 assigns greater weight to the contributions of IA_1 and IA_2, while M_2 assigns greater weight to IA_2 and IA_3 (see §4.1 for a more involved discussion). Of course, the pluralist will have to provide a principled story for how this could be the case and, ultimately, provide some evidence that this is sometimes the case (if this is to be her preferred strategy). At this stage, my point is merely that the pluralism does not necessitate the introduction of a new attitude for every measure in order to account for weak correlations, much less a one-to-one correspondence between species of implicit attitudes and measurement tasks. So, either Machery is responding to a version of pluralism that is irrelevant to the matter at hand or Machery is critiquing a caricature of the pluralist position. Either way, nothing Machery says in the above passage makes trouble for the sort of pluralist position that I develop next chapter.[16]

[16] Henceforth, I'll refer to type-pluralism simply as pluralism.

CHAPTER 3

THE MSP THEORETICAL RESPONSE TO MACHERY'S CHALLENGE

In the previous chapter, I articulated Machery's eliminativist challenge and defended it against monist criticism. The chapter ended with an argument that Machery's neglect of pluralism is significant. Far from establishing that implicit attitudes are not best conceived of as mental states, Machery only manages to establish that his trait view outperforms monist theories of implicit attitudes with respect to accounting for the core anomalies. While this result is significant, it does not warrant us in accepting the trait view.

The question before us now is whether Machery's trait view similarly outperforms pluralist accounts of implicit attitudes. And this is where Machery's failure to take seriously implicit attitude pluralism undermines his central argument. In what follows, I argue that there is strong, independent theoretical and empirical support for implicit attitude pluralism. Pluralism receives much of its support from interdisciplinary cognitive neuroscientific memory research. Indeed, the same field of memory research from which implicit social cognition researchers borrow the implicit/explicit distinction has fruitfully presupposes the existence of several memory systems that I, and others (see Amodio 2019), claim best explains many of the regularities at the center of social psychological explanatory project. Memory systems are individuated by their operating principles, the kind of information they store, the representational format of the stored information, their anatomical and functional connections with other memory systems, their channels of behavioral expression, and their underlying neural circuitry (Schacter & Tulving 1994, Ferbinteanu 2018). The systems, while distinct, interact in a variety of complex ways, and, in the wild, intelligent behavior is best understood as a product of these interactions. On the pluralist theory that I propose here, indirect measures typically stand in a one-to-many relation with respect to memory systems and these systems jointly contribute to the kinds of behavior that are of special interest to implicit social cognition research. Moreover, because the different memory structures encoded by these memory systems are different kinds of mental state, this licenses the conclusion that there are different kinds of implicit attitude. I call this view *Memory Systems Pluralism* (MSP). I contend that MSP theory easily outperforms Machery's eliminativist trait view in accounting for the core anomalies. And because Machery's trait view out-performs extant monist accounts of implicit attitudes, the same is true of MSP theory.

In outline, this chapter proceeds as follows. I begin by developing the contours of MSP in greater detail while also anticipating potential objections to the view (§1). I then develop the case against Machery's trait view from the MSP perspective and argue that this critique of Machery's trait view generalizes to other dispositionalist accounts of attitude (§2). The chapter ends with a discussion of some of MSP's limitations and some promising avenues of future research revealed by this perspective (§3).

1 Memory Systems Pluralism Revisited

The central aim of this section is to add some flesh to the bones the MSP theory, which was introduced in Chapter 1. In service of this goal, I outline Amodio and colleagues' (Amodio 2018, Amodio & Berg 2018, Amodio & Ratner 2011) multiple memory systems model of social cognition (hereafter, MMS-SC). This model has four important properties. First, Amodio's MMS-SC enjoys a great deal of initial plausibility owing to the fact that it is the product of a straightforward application of a highly confirmed model of memory in an adjacent field. Second, it is comparatively conservative, as it preserves the bulk of the insights already gleaned from social cognition research.

Third, the MMS-SC supports implicit attitude pluralism. Finally, MSP outperforms Machery's trait theory across a number of dimensions (to be discussed in §4).

1.1 The Initial Plausibility of The MSP Approach to Social Cognition

There are, perhaps, several routes that lead to implicit attitude pluralism, but one of the more direct routes involves the rejection of SSP. The reasoning here is that if implicit attitudes are stored in multiple memory systems and these memory systems share relatively few properties in common, then there is a strong case to be made that there are several kinds of implicit attitude. While the rejection of SSP certainly is uncommon in the social cognition literature, it is far from unprecedented: Amodio and colleagues (Amodio 2018, Amodio & Berg 2018, Amodio & Devine 2006, Amodio & Ratner 2011) have been urging social cognition researchers to supplant SSP with the MMS perspective for over 15 years. In what follows, I'll argue that MMS-SC strongly supports pluralism.

Before I describe MMS-SC in detail, a brief description of multiple memory systems (MMS) research in general is warranted. Squire and Zola-Morgan (1991) introduced a particularly influential MMS model (henceforth referred to as *the classical model*) with the goal of integrating several areas of research, spanning multiple decades, involving memory impairments in human and non-human animals (see **Figure 3** for a summary; Squire and Zola 1996). The classical model, as with every other model developed from the MSP perspective, treats memory not as being unitary but as being modular (Ferbinteanu 2019). The claim that memory is modular, in this context, amounts to the claim that the brain stores information based on the independent and parallel activity of a number of different memory systems "each with distinct properties, dynamics, and neural bases" (*ibid*: 61). As this is assumption is core to the MSP perspective, many of the current debates in memory research concern how best to update the classical model in light of rapidly accumulating data across various fields without sacrificing the commitment to this kind of modularity (*ibid*: 66). While it is possible that the final theory of memory would be one on which "there is only one memory system, which preserves all experiences and is used in all tasks" (Whittlesea & Price 2001: 244), the day currently belongs to the MMS perspective (Squire 2004, Ferbinteanu 2018).

The recognition that our best theories of memory presuppose the MSP perspective allows us to anticipate a likely misconception about memory systems models of social cognition—*viz.* memory systems models of social cognition are unparsimonious on the grounds that they are ontologically profligate (see Von Dessel, Gawronski & De Houwer 2019). But if our best theories of memory already commit us to the existence of multiple memory systems, it is not as though the number of memory systems somehow increases simply by explaining social-cognition relevant effects in terms of the operations of these systems; *mutatis mutandis*, for implicit attitudes if MMS-SC licenses pluralism (see §3.3.1 for an elaboration of this point).

The precise details of our commitments ultimately depend on the details of the best supported MSP model. As mentioned above, this is a topic of lively (and at times, extraordinarily technical) debate (Ferbinteanu 2018, Henke 2010). While resolution of these debates and general advancements in the field may force us to revise the MMS-SC accordingly, the general shape of MMS-SC is unlikely to be rejected, as there is little controversy surrounding the existence of the memory systems that are central to MMS-SC's explanations of the relevant social-cognitive phenomena.

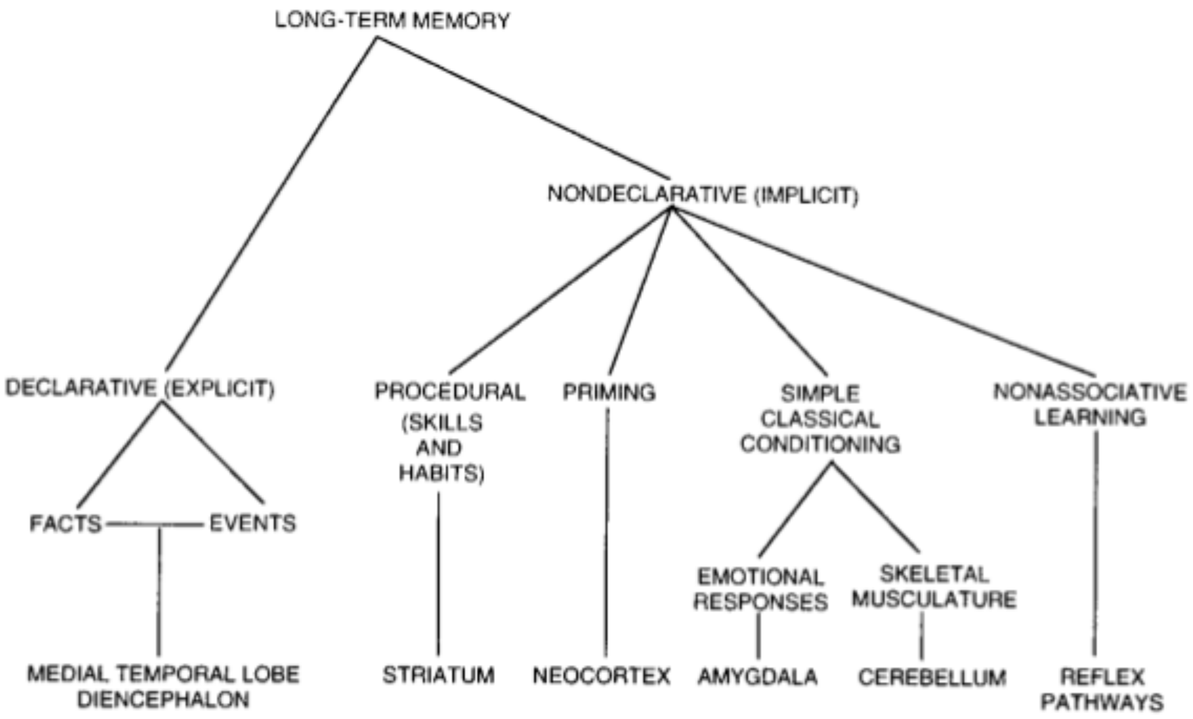

***Figure 3*.** The *MTL model.* On this model, the fundamental dichotomy is between *declarative,* or *explicit,* memory systems and *nondeclarative,* or *implicit,* memory systems. The explicit memory system is a single system that consists of two highly interconnected subsystems: episodic memory and semantic memory. Episodic memory is hypothesized as storing information concerning events relevant to one's autobiography (*e.g.*, *I recall hearing my favorite song being played at the coffee shop earlier today*), while semantic memory is charged with storing all manner of information of an impersonal sort simply referred to as "facts" (*e.g.*, *The Television Personalities released their first record in 1981*). While both subsystems are located in the medial temporal lobe (MTL) episodic memory is comparatively more reliant on the hippocampus whereas the semantic memory is more heavily reliant upon the neocortex. On this model, the contents of the mnemonic structures stored in the explicit memory system are normally such that subjects can come to have some kind of conscious or experiential awareness of them, hence their label. While the MTL model treats explicit memory as a single memory system that has two subsystems as proper parts, it treats nondeclarative, or implicit, memory systems differently. Each implicit memory system operates in parallel with respect to the others, stores different kinds of information, exhibits different learning and extinction rate, have distinct neurological substrates, and distinct functional properties. Each type of implicit learning has been observed in human and non-human animals alike. Importantly, this model assumes that individuals have relatively little, if any, access to the contents of the memory structures stored in these systems and that these systems typically operate automatically and below the conscious threshold. Adapted from Squire & Zola, 1996.

1.2 The MMS Model of Social Cognition

Extending the memory systems perspective to social cognition research yields the view that "we learn about people via multiple systems, encoding information in multiple representations, often simultaneously, and that these systems have complementary influences on judgements, actions and decisions" (Amodio 2018: 4)." SM (see §2.1 for review), by contrast, claims that implicit forms of learning occur via a single memory system, encoding information in the form of associations, and that this system drives automatic, spontaneous, non-conscious judgments and behavior.

Of the many memory systems discussed in the literature, Amodio (2018) tends to focus on three that are of most obvious relevance to the interests of social cognition researchers—*viz.* (1) the semantic-associative (or conceptual priming) system, (2) the Pavlovian aversive conditioning system, and (3) the instrumental (goal driven) system. (See **Table 2**.)

Before I describe the properties of each system, it is worth pausing to appreciate the differences between Amodio's MMS-SC and the classical model. The first thing to note is that the implicit/explicit distinction does not appear in the presentation of MMS-SC (*cf.* Henke 2010). This constitutes a departure—though not necessarily a radical one—from some of the more well-known dual-process theories of cognition (*e.g.*, Kahneman 2011, Strack & Deutsch 2004). On MMS models that eschew the implicit/explicit distinction as an organizational principle, implicit/explicit memories are typically regarded as *forms* of memory. Schacter and Tulving's discussion is particularly illuminating:

> Explicit and implicit memory are not systems. These terms were put forward to describe and characterize expressions of memory: "explicit" refers to intentional or conscious recollection of past episodes, whereas "implicit" refers to unintentional, nonconscious use of previously acquired information. Schacter notes specifically that the implicit/explicit distinction "does not refer to, or imply the existence of, different underlying memory systems" (1987:501). Thus, according to this formulation, implicit and explicit memory, though psychologically and behaviorally distinguishable forms of memory, could either depend on the same underlying memory system or different underlying systems; the question is open and subject to experimental investigation. (1994: 12-13).

So, in the discussions that follow, I'll be adopting this understanding of the implicit/explicit distinction (*cf.* Gawronski & Bodenhausen 1996). With that said, we will more closely examine the relationship between memory systems models of cognition, social and otherwise, and dual systems models of cognition in Chapters 4 & 5.

Second, MMS-SC proposes several refinements to the classical model. One such refinement has to do with *priming*. Where the classical model proposes only one priming system, most of the more recent models posit two (Schacter & Buckner 1998, Tulving & Schacter 1990): There is *repetition priming* (or perceptual priming) whereby exposure to an item facilitates an ability to identify or reproduce it (*e.g.,* subliminal exposure to an umbrella facilitates your ability to recognize umbrellas or draw an umbrella); *semantic priming*, by contrast, is the phenomenon whereby exposure to an item facilitates one's ability to produce a semantically related item (*e.g.*, you study the word 'banana' and exposure, sometime later, to the word 'fruit' causes you to think of bananas). Where repetition priming effects have been shown to occur across the neocortex (largely depending on the exploited sensory modality), the anterior temporal lobe (ATL) appears to be of particular importance for semantic priming.

Another refinement is worth mentioning. While memory systems models share the classical model's assumption that the amygdala has an important role to play in classical conditioning (or Pavlovian conditioning) of affective responses generally, the more recent memory systems models assume that the amygdala plays a comparatively more active role in aversive or fear-related conditioning in particular (Gore *et al.* 2015). While we do not need to make precise the distinction being drawn here, a quick example will help to bring the distinction into relief. One might, for example, be trained to associate the term 'black' or the concept BLACK with negative affect via the repeated pairing of the term 'black' with items that produce the characteristic set of evaluative responses that are widely thought to signal a negative affective experience (*e.g.,* one might be able to identify the term 'black' as being less positive than the term 'white'). And when that evaluative response does signal the presence of an occurrent negative affective experience, amygdala activation helps to explain the occurrence of this negative affective experience. But, crucially, the same evaluative response to the term 'black' could be generated by an individual even when the corresponding affective experience is absent or when the relevant amygdala structures are damaged (see Claire et al. 2016); though for a lengthier discussion, see Chapter 6). Fear conditioning, according to the most recently articulated memory systems models is comparatively special. When one comes to associate a stimulus object with the relevant class of fear related evaluative responses, the amygdala's involvement is comparatively far more likely. Generally speaking, one does not typically find individuals with severe amygdala damage to demonstrate the same range of associated fear-related evaluative responses to items that usually trigger fear responses when compared with individuals with intact amygdalae. Because the mechanism whereby one comes to associate a term or lexical concept with affect differs from the mechanism whereby one comes to associate a stimulus object with a fear response, it is not uncommon for memory researchers to identify several memory systems involved in affective learning and regulation (Murray 2007).

Having said all this, let's now limn the central characteristics of the aversive conditioning, instrumental learning and semantic priming systems. The goal of this discussion is two-fold: First, I want to give the reader a sense of why these systems are treated as distinct memory systems rather than, say, merely different memory-related processes that operate over a single system. Second, it should also provide the reader with a sense of why these systems are relevant to the *explananda* of implicit social cognition research. Readers who are interested in the details should consult any of the surveys already available (*e.g.,* Amodio 2018, Amodio & Berg 2018, Amodio & Ratner 2011, Ferbinteanu 2018).

1.2.1 The Pavlovian Aversive Conditioning System

The Pavlovian Aversive Conditioning (PAC) system is subserved primarily by the amygdala, which is a subcortical structure buried deep in the medial temporal lobe (MTL). A number of studies have demonstrated that a particular structure in the amygdala—the basolateral amygdala (BLA)—represents stimuli with aversive reinforcement value (S*s, also known as unconditioned stimuli (USs)) of any sensory modality (*e.g.,* electric shocks, loud noises, bad smells, bad tastes). S*s elicit a variety of innate emotional and autonomic responses (Ferbinteanu 2018). For instance, a painful electric shock may result in fear and anxiety, cause an increase in one's heart rate, skin conductance responses, cortisol levels and generate any number of avoidance behaviors (LeDoux 2007).

Importantly, the amygdala can form associations between S*s and stimuli that lack innate reinforcement value (*e.g.,* a neutral auditory tone) via associative pairing (*e.g.,* an electric shock is administered concurrently with or immediately after the neutral tone) resulting in an S-S* associative representational structure. Once a particular S has become associated with an S*, there mere presence of S is sufficient to elicit the same sorts of affective, autonomic, and avoidance responses typical of S*. These representational structures are often formed very rapidly, often after only a single S-S* pairing (LeDoux 2000), and are extraordinarily difficult to extinguish via counter-conditioning (*i.e.,* these associative structures are hard to break simply by presenting the S *sans* S*) (for a review, see Maren, Phan, & Liberzon 2013).

Moreover, the hypothesis that the PAC system is distinct from other memory systems has been established via a number of studies that report double dissociations. For example, while healthy individuals demonstrate conceptual learning via self-report (*e.g.,* Blue cards lead to shocks!) and affective learning (*e.g.,* skin conductance responses increase when selecting a blue card) during a classical conditioning task, individuals with amygdala lesions demonstrate conceptual learning *sans* affective learning. Individuals with hippocampal lesions, moreover, show the reverse impairment: They demonstrate affective learning without conceptual learning (Amodio 2018, Bechera 1995).

But what reason do we have to think that the PAC system plays a role in influencing the kind of social behavior that is of interest to social cognition researchers? In the context of social interactions, heightened amygdala activity is strongly correlated with increased anxiety, averted gaze, disfluent speech, closed body posture and increased interpersonal distance (Amodio 2010, Amodio & Devine 2006, Dovodio, Kawakami and Gaertner 2002, McConnell & Leibold 2001). With respect to social decision making, compared to healthy individuals, subjects with selective amygdala damage are willing to invest 100% more money in unfamiliar others even when no return on investment is expected (van Honk, Eisenegger, Terburg, Stein & Morgan 2013). Other studies have shown that the amygdala is far more responsive to angry relative to happy faces, even when exposure to these faces is subliminal (Williams *et al.* 2004). Thus, it is highly likely that the PAC memory system plays an important causal-explanatory role with respect to the phenomena central to social cognition research.

Having said all this, how should we conceive of a PAC-based implicit attitude (IA$_{\text{PAC}}$)? Here's a first pass:

> IA$_{\text{PAC}}$: An amygdala-based associative structure that obtains between a social object, category or property (S) and a stimulus with aversive reinforcement value (S*) (*i.e.,* a S-S* pair) such that the mere presence of S elicits an emotional response, a range of autonomic responses, or an approach/avoidance behavioral orientation.

For instance, a person who has an amygdala-based association between a Black face (S) and a gun (S*) such that the mere exposure of a Black face results in the sorts of aversive responses typical of what occurs when one sees a gun.

1.2.2 Instrumental Learning System[17]

Before I articulate Amodio's conception of the instrumental learning system, it is worth prefacing this discussion by stating that the more general memory systems models tend to identify not one but two instrumental learning systems— *viz.* the model-free instrumental learning system and the model-based instrumental learning system (see Daw & O'Doherty 2014). For present purposes, I'll focus only on describing the model-free instrumental learning system. Aside from the

[17] A quick caveat: The interests of brevity make a nuanced discussion of the operations of the Instrumental Learning (IL) system impossible. Those with a hankering for a more technical discussion should see the references provided in the main discussion.

fact that Amodio does not directly appeal to the model-based instrumental learning system in his model, there are a couple of further reasons for bracketing discussion of it for present purposes: first, the relationship between the model-based instrumental learning system and other memory systems is a topic of active and ongoing empirical and theoretical controversy (*ibid.*). Second, none of the discussions that follow turn on questions of the operation of the model-based instrumental learning system. As such, it's best that we set aside a lengthy discussion of the model-based instrumental learning system at present.

With that said, memory researchers are inclined to locate the model-free instrumental learning (hereafter, simply IL) system in the ventral striatum and the basal ganglia (Amodio 2018, Squire & Zola 1996). It is common for modelers to conceive of IL as operating in accordance to reinforcement learning principles: For any given behavior, B, one comes to associate B with positive or negative value (V) on the basis of past experience. At risk of oversimplifying matters, one may think of representations formed by the striatum as being associative structures that yoke a particular V (which may be graded) with B. It is generally assumed that changes in the value associated with the representation of the relevant stimulus are error-driven: if the production of a behavior in response to a particular stimulus results in greater or lesser expected value, the association between the production of a particular behavior in response to the corresponding stimulus changes in accordance with the magnitude of the difference between the predicted value of the behavior and its actual value.

IL, moreover, can go about influencing an organism's behavior in the absence of conscious awareness or intention. Thus, goal-pursuit can quite often occur automatically and below the conscious threshold. When cognitive resources are abundant, both the striatum and the hippocampus work in concert to influence goal-oriented decision making (Amodio 2018, Foerde *et al.* 2006). However, when under cognitive load or when the hippocampus is damaged, it is typically the instrumental system via the striatum that is the primary driver of goal-oriented behavior. Consequently, the instrumental learning system is relatively unrestricted in its forms of behavioral expression (Amodio & Ratner 2011).

Given that the Pavlovian conditioning systems (of which, PAC is one) also have a role to play in guiding goal-driven behavior, why think that the IL system constitutes a separate memory system? There are several reasons for this. First, Pavlovian systems are selectively attuned to biologically relevant rewards, while the IL system is not so picky (Amodio & Ratner 2011). A reciprocation of a greeting might be interpreted by the IL system as rewarding under the right circumstances. Second, each kind of system implements a different kind of learning/extinction algorithm: A large prediction error might bring about rapid learning in the IL system, but, for instance, once a tone has come to be associated with a shock, pairing a tone with an animal's favorite treat is unlikely to effect much change. Third, double dissociations between the PAC and the IL systems have been observed (MacDonald & White 1993).

But why think that IL should shed any new light on implicit bias? Aside from the role that the IL system plays with respect to goal pursuit, conscious or otherwise, studies have also indicated that this system has a role to play with respect to impression formation (Hackel *et al.* 2015) and can modulate the affective responses generated by the various Pavlovian conditioning systems (Cunningham, Van Bavel, Johnsen 2008).

Having said all of this, here's a first pass at characterizing an IL system-based implicit attitude (henceforth, IA_{IL}):

> IA_{IL}: A striatal/basal ganglia-based associative structure holding between a social-object-directed behavior and a value (*i.e.*, a B-V pair) such that its activation serves to

> explain conscious and unconscious goal pursuit, impression formation, and affect modulation.

To illustrate, imagine that Jones is looking over several applications of prospective hires at his company at the busiest time of year. Jones has had little personal experience working alongside women and is well-aware of the stereotype that women are incompetent with respect to business matters, though this is a stereotype that he would firmly disavow were asked. As such, he passively associates hiring men with positive value and hiring women with negative value. This negative association generates predictions as to the expected negative value of hiring a woman for the job, which, in conjunction with his predictions that hiring a man would yield positive value, leads him to pass on applications with stereotypically female names. In this case, we can discern the operation of at least two distinct IA$_{IL}$s each of which can plausibly regarded as an implicit bias: Jones' attitude that (1) assigns negative value to the hiring of a woman for the position and the attitude that (2) assigns positive value to the hiring of a man for the position.

1.2.3 The Semantic-Associative (or Conceptual Priming) System

Of these three memory systems, the semantic-associative system is likely to be the most familiar. While there is some controversy as to where to locate the neurological substrates of this system, evidence is accumulating for the importance of the anterior temporal lobe (ATL) to automatic semantic-associative processing (Lau, Gramfort, Hämäläinen, & Kuperberg 2013). Through a series of neuroimaging studies, Zahn and colleagues (2007) demonstrated that the ATL represents abstract social knowledge as indicated by the fact that structures in the ATL were selectively engaged when subjects were asked to judge the meaning-relatedness of social concepts (*e.g.*, HONOR-BRAVE) relative to non-social concepts (*e.g.*, NUTRITIOUS-USEFUL). Importantly, activation in the ATL was shown to be valence independent, whereas medial prefrontal cortical regions—which are strongly interconnected to the amygdala— were shown to be especially active for valenced social concepts.

The semantic-associative system operates in accordance with the principles already enumerated in our brief discussion of Strack and Deutsch's (2004) Impulsive system. On this view, the semantic-associative (or conceptual priming) system stores associative representations between stimuli (S-S associative pairs). Many associative models treat the relevant S-S pairs as obtaining between lexical concepts (*e.g.*, SALT-PEPPER). Moreover, as we had already learned from Strack and Deutsch's discussion, it is hypothesized that the associative strength between S-S pairs develops gradually via repeated pairings and extinguishes gradually via repeated non-pairings. The same spreading of activation principles discussed earlier apply here, which allows it to explain the various semantic priming effects that have received a great deal of attention in the social cognition literature (for a review, see Wittenbrink, 2007).[18]

Since MMS-SC posits a semantic-associative system, it can account for everything that can be successfully explained by the standard model. Thus, the MMS-SC leaves intact most of the relevant insights that social cognition research has produced over the last three decades. And since the vast majority of the indirect measures of attitude currently employed by social cognition researchers have been designed to tap into an associative-semantic network, there is good reason to think that very little of this massive body of work needs to be fundamentally re-interpreted in light of the MMS-SC.

[18] It must be noted that *repetition* priming is an altogether different phenomenon from *semantic* or *conceptual* priming (see Dehane *et al.* 2001).

There are other ways, however, in which adopting the MMS perspective does challenge our current conception of the semantic-associative network. The standard model endeavors to explain all implicit social cognition-relevant behavioral effects in terms of the operations of this memory system. On MMS-SC, this explanatory burden is greatly reduced: The semantic-associative system has a role to play in generating priming effects, the classification of stimuli (especially faces), and in impression formation (Amodio 2018).

As such, here's a first pass at defining the implicit attitudes that are stored in the semantic-associative system (IA_{SA}).

> IA_{SA}: A semantic-associative implicit attitude is a stimulus-stimulus (S-S) associative structure that is based in the anterior temporal lobe (ATL). This structure mediates the classification of stimuli, a variety of semantic-priming effects, and impression formation.

So, a BLACK-DANGER association might be located in the ATL (or thereabouts) and may facilitate the identification of semantically related items (*e.g.*, guns) and influence the way in which an impression is formed of an individual on the first interaction.

1.2.4 Interactions Among Memory Systems

While much of this section has emphasized the ways in which the memory systems are distinct, much of the MMS literature aims to describe how these memory systems interact so as to produce intelligent behavior (Ferbinteanu 2018). There are three kinds of interactions between memory systems: (a) cooperative interactions, (b) competitive interactions, and (c) informational transfer. As space is limited, I will only discuss the first two (for a review survey on research relevant to informational transfer, see Ferbinteanu 2018, White, Packard & McDonald 2013).

Cooperative interactions between systems occur when the proper functioning of both systems is necessary to produce fluent behavior (see **Figure 1** for some of the hypothesized *cooperative* interactions between memory systems) (Ferbinteanu 2018). The Iowa Gambling Task (IGT) is frequently used paradigm that assesses instrumental learning. It consists of having subjects draw cards form four decks. Drawing from decks A and B yield large initial gains followed by even larger losses such that drawing from these decks yield net losses; whereas, drawing from decks C and D yield small initial gains followed by smaller losses such that drawing exclusively from these decks yields net gains. Healthy subjects typically learn that A and B are the bad decks and their behavior reflects this. Subjects with bilateral amygdala damage might learn that decks A and B are bad but continue to select from them anyway all the while showing little emotional response to large losses. Subjects with hippocampal damage also perform poorly at this task: While they generate typical affective responses to large losses, they are never able to discriminate, conceptually, the good decks from the bad (Bechera 1999). Consequently, they simply pick from one deck until it yields a loss, at which point they begin to select from another (Gupta, Koscik, Bechera, Tranel 2011). This reveals that successful task performance in normal cases depends on the joint contributions of the PAC and hippocampus-dependent memory systems.

But memory systems may also compete for behavior. One kind of competition between memory systems occurs when the joint contributions of two memory systems *inhibit* task performance. Accordingly, taking one memory system offline should *facilitate* fluent task performance. Such competition between the hippocampus and the striatum has been demonstrated in rats (Poldrack & Packard 2003). Another kind of competition is said to occur when two memory systems—for instance, the PAC and the IL systems compete for control of the organism's behavior.

For instance, think of a scenario in which one can win a large sum of money if only one could overcome one's arachnophobia and climb, *Fear Factor*-style, into a box of spiders. As both recommend different courses of action (*e.g.*, PAC says run!; while IL says get in the box!) there must be some mechanism by which these different systems get channeled into a singular behavioral output. Further, available evidence seems to indicate that the anterior-cingulate cortex plays a central role in the monitoring and resolution of inter-system conflict (Botvinick *et al.* 1999).

This discussion highlights yet another theoretical advantage that MMS-SC has over SM: MMS-SC is better positioned to model the nature of the semantic-affective interface than even the most plausible versions of SM. Here's a standard SM-approach to describing the semantic-affective interface:

> In the APE model [a highly influential version of SM], we assume that principles of similarity matching determine the activation of mental concepts that represent the encountered stimulus (*e.g.*, Afrocentric features of a face activating the concept *African American*), which can spread to other concepts that are associatively linked with the stimulus (*e.g.*, activation of the concept *African American* spreading to the associated stereotypical attribute *hostile*). To the extent that the associated concepts have a positive or negative connotation, their activation is assumed to produce a spontaneous gut response that is in line with the valence of these concepts (*implicit evaluation*). (Gawronski & Bodenhausen 2014: 189)

But what exactly does it mean for a concept to have a positive or negative valence? In keeping with the connotation metaphor, perhaps the valance of a concept is best thought of as a function that maps concepts onto gut responses (*i.e.*, affective responses). This does enjoy some intuitive appeal: I token the concept VOMIT, and I automatically experience disgust. The problem is that such theories struggle to explain how a concept, stored in a semantic-associative network, comes to acquire a valence. One might be tempted to offer an explanation in terms of evaluative conditioning (which is the process whereby an otherwise valence-neutral conditioned stimulus (CS) comes to acquire the valence of an unconditioned stimulus (US)). While this might help to explain how representations of otherwise neutral CSs acquire valence, such theories simply assume that the relevant representations of USs have valence. As such, they do not, by themselves, help the SM-theorist explain how valence enters the semantic-associative system in the first place. Otherwise put, such theories are ill-equipped to explain the semantic-affective interface.

By contrast, an MSP is well-positioned to explain the semantic-affective interface. For instance, attention to underlying neural circuitry should provide clues as to which systems are most likely to interact and how. For instance, the amygdala receives information very early on in processing and, because there are well-established circuits whereby it is connected to the ATL, MMS-SC predicts that the PAC may modulate the activity of the semantic-associative system which, in turn, will facilitate the classification of, say, threat-relevant stimuli when the relevant PAC-based associations are activated (Packard, Cahill, McGaugh 1994). Indeed, existing evidence bears this out: Heightened threat-related activity in the amygdala facilitates classification of threat-related items (*e.g.*, weapons) (Davis & Whalen 2001).

2 Potential Objections

Up to this point, I have used Amodio's (2018) discussion of MMS-SC to motivate the existence of three kinds of implicit attitude, each of which is defined in terms of the operations of a different memory system (see **Table 2** for a summary of the three memory systems). Moreover, as this is neither a complete list of memory systems, nor does it exhaust all the memory systems that are of relevance to social cognition research, there's good reason to expect that there are more kinds of implicit attitudes than the ones described here. Before pitting MSP against Machery's trait view, it is worth heading off two potential objections.

2.1 The Objection from Parsimony

Earlier, I anticipated the objection that MMS-SC is guilty of running afoul of Ockham's razor by positing the existence of multiple memory systems (and multiple kinds of implicit attitude) when one such system (and one kind of implicit attitude) is all that is needed to account for social cognition's *explananda*. I've already explained that this objection is based on the false supposition that MMS-SC commits us to the existence of states and mechanisms that our best cognitive scientific theories do not already commit us to. *Pace* Brownstein (2018) and Levy (2014), implicit attitudes are not *sui generis*; rather, they are the subset of memory structures operated on by the various aforementioned memory systems that help to produce the kinds of thoughts and social behaviors that are of special interest to social cognition researchers. Moreover, the memory systems that MMS-SC posits have been central to a highly productive, inter-disciplinary tradition for the past three decades. Thus, the charge of ontological profligacy sticks not to theories of implicit attitudes that are committed to MMS-SC, but to those theories on which implicit attitudes are a unique, unitary, and hitherto unrecognized mental kind.

In response, the monist might press the following line: The MMS-SC theorist has largely taken for granted that the dominant view in memory research is the only view that can accommodate the data canvased in this section. But all of the research discussed could be accommodated on a view according to which there is a single memory system and multiple distinct learning and retrieval processes. For instance, the sort of hippocampal damage associated with anterograde amnesia can be thought of as a case in which the relevant damage knocks off-line one's capacity to either encode or retrieve new episodic memories. But this explanation is entirely consistent with the single-system perspective (SSP). Should this reasoning generalize— that is, should the SSP be able to accommodate all of the research that motives the MMS-perspective— then it would appear as though MMS-SC is wildly unparsimonious after all. Indeed, Van Dessell, Gawronski and De Houwer (2019) have already made a very similar case.

Squire (2004: 175) anticipates this objection and his response is worth quoting at length:

> Typically, the notion is that there is only one memory system but that there are multiple processes operating on this system or multiple ways of accessing its contents. The difficulty with such views is that they are unnecessarily abstract and make insufficient contact with biology. For example, the findings from eyeblink conditioning provide direct evidence for a kind of memory that can be acquired, stored, and retrieved in absence of the forebrain. Other kinds of memory (*e.g.*, perceptual learning, declarative memory) do require the forebrain. The locus of memory storage is entirely different in these cases, and the learning proceeds by different principles. Perhaps there is some level of abstraction at which synaptic changes within the cerebellum and synaptic changes within the neocortex can be viewed as different expressions of a single memory system. However, such a perspective tends to ignore rather than embrace the enormous amount that has been learned about neuroanatomy, the molecular and cellular biology of synaptic change, and the organization of brain systems.

There is a sense, then, in which Van Dessell, Gawronski, and De Houwer (2019) miss the point. Everyone grants that a single-system theorist can, in principle, accommodate the data that has led to the development of the MMS perspective. The question is whether such theories can integrate research from cognitive neuroscience in a manner that is at all theoretically illuminating. Advocates of the SSP have given us no grounds for optimism.

2.2 The No Non-Affective Thought Argument

I've argued that the MMS-SC is better positioned than SM to account for the ways in which semantic processing interacts with affect. However, Madva and Brownstein (2018) would argue that the situation is actually the reverse based on the following reasoning:[19]

The *No Non-affective Thought* Argument

P1) MMS-SC, but not SM, is committed to the existence of non-affective thought.

P2) There is no such thing as non-affective thought.

C) So, MMS-SC, but not SM, is committed to entities that don't exist.

In defense of P1, Madva and Brownstein (2018) would presumably say that MMS-SC is committed to the existence of non-affective thought because it distinguishes between the semantic-associative system and the Pavlovian aversive conditioning system. SM, by contrast, is silent on matters related to the cognition/affect divide. Regardless of whether MMS-SC is so-committed, I am willing to grant Madva and Brownstein P1 for the sake of argument.

[19] Madva and Brownstein (2018) don't have my version of pluralism in mind as their target. Rather, the argument is offered against Amodio and Devine's (2006) proposal that we think of implicit stereotypes and implicit evaluations as being distinct constructs.

But what about P2? This is an incredibly contentious claim and a full assessment of it is well beyond the scope of this chapter. As such, I'll focus only on the reasons that Madva and Brownstein give in its defense. Oddly, Madva and Brownstein don't offer up much. Here's what they do offer:

> Reviewing the literature demonstrating the role of affect in the processing of conscious experience, language fluency, and memory, Duncan and Barrett (2007) argue that "there is no such thing as 'non-affective thought.' Affect plays a role in perception and cognition, even when people cannot feel its influence" and conclude that the "affect-cognition divide is grounded in phenomenology." (Madva & Brownstein 2018: 622)

But the central argument contained in this passage is clearly invalid. Even if we grant that our best models of cognitive processing are such that affect contributes to cognition across the board, it neither follows that the affect-cognition divide is grounded in phenomenology nor that all thought is affective. The putative fact that cognitive processes and affective processes always jointly contribute to the production of intelligent behavior is entirely consistent with the view that cognitive processes and states are different in kind from affective processes and states (Carruthers 2018). Thus, Madva and Brownstein have not provided us with sufficient reason to accept P2. We do, however, have good evidence to endorse MMS-SC (or, at least, a close approximation of it). Consequently, *sans* further evidence to the contrary, we ought to reject P2.

3 MSP Theory Vs. The Trait View

Are implicit attitudes mental states? MSP implies that they are, and the trait view implies that the very question is predicated on a category mistake. While we have good reason to believe that the trait view outperforms monist versions of the Freudian view, how does Machery's trait view fare against MSP? I contend that it performs relatively poorly. MSP and the trait view converge on the view that a heterogenous collection of mental states drive performance on indirect measures and social behavior more generally. In that sense, Machery was right to say that pluralism is similar to the trait view. However, only MSP offers a structured framework for understanding the weak relationships among indirect measures, low predictive validity and hitherto unexplained variance in test-retest reliability. As such, the adoption of MSP can better contribute to productive empirical and theoretical research than the trait view. And while the proponent of the trait view might be tempted to view MMS-SC as merely describing the heterogenous psychological bases on which the relevant attitudes depend, there is no theoretical payoff for doing so. In other words, there is no theoretical value added by conceiving of MMS-SC as describing the psychological bases of attitudes *qua* traits that is not already added by simply adopting MMS-SC. And since MMS-SC already traffics in mental states that have the sorts of contents and functional properties that have long been of interest in implicit social cognition research, it naturally lends itself to pluralist interpretation (as was argued in §3). Thus, we ought to endorse MSP over the trait view.

3.1 Round 1: Weak Correlations Across Indirect Measures

For the sake of argument, let us assume that weak-correlations across indirect measures cannot be explained exclusively by appealing to considerations of process purity, where process purity here is to be understood solely in terms of the differential contributions of *implicit* vs *explicit* processes (see §1.2.1). Machery's trait view offers this explanation:

> According to [the trait] picture, indirect measures typically tap into different components of the psychological bases of attitudes. There is no reason to expect these components to correlate with one another. For instance, the association between the concepts of a black man and of danger may be strong, even if someone has only a weak automatic fear reaction to black men (or vice versa). As a result, indirect measures should often correlate poorly with one another (2016: 117).

When faced with weak or variable relationships among indirect measures, the MMS-SC has recourse to several different explanations. Assuming that there is no serious concern about a measure's reliability or internal consistency, there are several possible explanations for why weak or variable relations across indirect measures might be observed. One possibility is that different indirect measures, M_1 and M_2, tap different sets of memory systems. For instance, while there is good reason to think that the Startle-Eyeblink Paradigm provides a relatively pure measure of amygdala activity (Amodio & Devine 2006), MMS-SC predicts that performance on the standard evaluative Black-White IAT reflects the output of several memory systems including the semantic-associative and the Pavlovian aversive conditioning system. Why would MMS-SC predict this? While I take it that the role of semantic-associative processing in driving IAT-relevant behavior is uncontentious, several lines of research converge on the importance of the PAC-system in driving IAT-effects: To name just a couple, numerous neuroimaging studies document higher amygdala response when one views Black faces relative to White faces (see Eberhardt 2005 for review) and subjects who ingest propranolol, a beta-blocker known to suppress amygdala activity, prior to taking an IAT, show less bias relative to controls (Terbeck *et al.* 2012). But should we expect indirect measures to weakly correlate when they tap the same set of memory systems? There are at least two scenarios in which we should. One scenario, first raised in §2.2, is that M_1 and M_2 tap the same kinds of implicit attitudes, but that the different attitudes contribute differently to each measure. Such explanations seem especially appropriate when, *ceteris paribus*, two measures only modestly correlate. For instance, it is plausible that the IAT and the EPT (when it is designed so as to capture attitudes toward groups as opposed to individuals, see Cooley and Payne 2017) tap the same kinds of implicit attitudes—*e.g.,* IA$_{SA}$s and IA$_{PAC}$s—but that the IAT is more sensitive to IA$_{SA}$s than IA$_{PAC}$s and the EPT is more sensitive to IA$_{PAC}$s than IA$_{SA}$s. Whether this actually obtains is a topic for future empirical research. But there are also cases in which we would expect performance on two measurement tasks that tap the same memory systems to be either weakly correlated or uncorrelated. In one such scenario, M$_1$ harbors a strong implicit bias toward a social object, whereas M$_2$ harbors no bias toward a social object. For example, suppose that the IAT and the First Person Shooter Task (FPST) tap the contents of the same memory systems—*viz.* the PAC system and the SA system—but an individual S harbors a robust IA$_{SA}$(BLACK-DANGER) association but harbors no (or, perhaps a very weak) IA$_{PAC}$(black-threat) association. In this scenario, we would expect the IAT to reveal strong implicit bias and the FPST to reveal no implicit bias. Thus, in such circumstances, we would expect to observe weak relationships between the two measurement tasks even though the tasks are structured so as to recruit the same memory systems. Whether, or how often, these scenarios actually obtain in the wild, though, is a matter for future empirical investigation.

Having said all of this, there is little direct evidence that either of these possibilities actually obtain with respect to the standard suite of implicit measures (*e.g.*, IAT, AMP, EPT, FPST, etc.). This is largely to be expected given that SM has been the dominant narrative of implicit bias and, as a reminder, SM is committed to a single-system perspective. And while there are formal models that attempt to tease apart the contributions of various processes to performance on indirect measures (*e.g.*, the Process Dissociation Procedure (Yonelinas & Jacoby 2012); QUAD model (Conrey *et al.* 2005)), extant formal models have been designed only to track contributions of automatic versus controlled processing. On MMPS, multiple memory systems typically operate automatically and in parallel. Thus, these formal models are in no position to shed light on the extent to which different automatically operating memory systems contribute to performance on any given indirect task.

Attempts to establish that different memory systems make distinct contributions to different versions of the IAT have shown promise, but further investigation is required to rule out potential confounds. For instance, Amodio and Devine (2006) developed one version of the IAT that was designed to tap primarily cognitive racial stereotypes (Stereo-IAT) (subjects were asked to sort Black and White faces with pleasant or unpleasant words) and another that was designed to measure primarily affective evaluations of racial groups (Eval-IAT)(subjects were asked to sort Black and White faces with words associated with mental and physical ability). Across a series of studies, Amodio and Devine (*ibid.*) were unable to establish significant correlations between the two measures—indeed, implicit stereotypes and implicit evaluations were often found to be doubly dissociable. Further, each measure predicted different kinds of behavior consistent with the functional profiles of IA$_{SA}$'s and IA$_{PAC}$'s respectively. While it would appear as though the two measures tap different memory systems, there are legitimate concerns as to whether the data could be better explained by appealing to differences in "the content of the associations, without recourse to distinct underlying mechanisms" (Holroyd & Sweetman 2016: 93).

Both the trait view and MSP, then, account for the weak correlations among indirect measures by appealing to the heterogeneity of what is being measured. But despite this, there are good reasons for thinking MSP's approach is superior in this regard. The main reason—and this will be a recurring one— is that if the trait theorist were to try to provide detailed descriptions of the "components of the psychological bases of an attitude," then presumably her picture would look very much like the one sketched by MSP (at least, if she wants her goal is to sketch an empirically adequate picture). Moreover, in order to explain how these psychological bases interact to drive behavior, again, the picture that she'll draw is one that will have to look similar to the one provided by MSP. After all, the trait theory, *sans* a wide array of auxiliary assumptions, makes no specific predictions about the states, processes and mechanisms that form the psychological bases of attitude. To fill in these very crucial details, the trait theorist has to turn to the best theories of cognitive architecture of the day and there is a compelling case to be made that MMS models are amongst them. As such, there is a sense in which the explanatory appeal of the trait theory *vis-à-vis* explaining the weak relationships across indirect measures is wholly parasitic on the explanatory appeal of MMS-SC. Without MMS-SC's accounts of the mechanisms, states, and processes that serve as components of the psychological bases of attitudes, the best that the trait can theory offer are *post-hoc*, just-so stories about why measures fail to correlate.

But MSP boasts an additional advantage over Machery's trait view with respect to explaining variable relationships among indirect measures. Machery conceives of attitudes as traits, and traits are multi-track dispositions to feel, cognize, and behave in trait-relevant ways. According to Machery

> if the coreferential, differently valanced mental states do not lead to a broad-track disposition to behave and cognize in a way that expresses either a positive or negative preference, then people simply do not have an attitude toward the relevant object. They will act and cognize in a way that expresses a positive preference in some contexts and a negative preference in other contexts, and their aggregate behavior cannot be predicted (even imperfectly) by positing a trait. (Machery 2016: 124)

So, what are we to say of the individual who lacks an automatic fear reaction toward Black people but harbors strong BLACK-DANGER semantic association? Suppose that such an individual is inclined to enter predominantly Black spaces and shows little concern when he/she crosses paths with a Black stranger at night. And yet, his/her robust BLACK-DANGER association causes him/her to cognize in ways consistent with this associations (*e.g.*, maybe s/he is more easily able to conjure up exemplars of dangerous Black people relative to dangerous White people). For the sake of argument, let's imagine that this individual encounters automatic-fear relevant situations just as frequently as she does the semantic-association relevant situations so that aggregation doesn't help us predict her behavior. On Machery's view, we are to think of her as having no attitude toward Black people. My pluralist view, however, licenses the interpretation that this individual has *ambivalent attitudes*: She harbors one kind of implicit bias and not another. This certainly seems to be the right thing to say about the individual in this scenario and has important implications for discussions concerning the ethics of implicit attitudes (which will be touched upon in §5).

3.2 Round 2: Unexplained Variance in Temporal Stability

Machery claims that views on which implicit attitudes are mental states are unable to account for the ways in which seemingly irrelevant situational factors affect performance on indirect measures. Recall, he claims, that whereas Freudian theories cannot predict that taking an IAT in a dark room should increase implicit bias, the trait theory can. His explanation bears repeating:

> By contrast, the trait picture hypothesizes that attitudes depend on psychological bases that encompass good old-fashioned mental states and processes, such as emotions, self-control, and so on, and that indirect measures tap into some of these components… [B]ecause darkness heightens stress in humans… it should modulate

> implicit association test scores if this test measures, or is somehow influenced by, stress. (*ibid.*)

By appealing to the ways in which memory systems interact to drive behavior, MSP provides a more satisfying explanation of the ways in which contextual factors drive performance on indirect measures than Machery's trait view. To demonstrate the importance of this approach, let's consider how MSP can accommodate the findings that Machery refers to. The relevant study is one in which individuals who explicitly believe that the world is dangerous are shown to have a stronger negative preference for Blacks over Whites when taking an IAT in a dark room relative to a well-lit room (Schaller, Park & Mueller 2003). While we cannot say with any degree of certainty that amygdala activation was enhanced for subjects who took the Black-White IAT in the dark room relative to the well-lit room, as no measure of amygdala activity was obtained in this experiment, a great deal of research on the amygdala's role in threat detection (for review, see Chekroud, Everett, Bridge, Hewstone 2014) renders it likely that this is what we would observe were we to include a measure of amygdala activity in a replication of this study. Not only should ambient darkness serve to enhance PAC-related amygdala activity, but so should the pictures of Black faces that are presented as IAT stimuli (Phelps, Ling & Carasco 2006). If the PAC-system were more actively involved in the ambient darkness condition relative to the well-lit or control conditions, then we likewise should expect to have found proportional increases in the relevant autonomic responses (*e.g.*, higher SCRs, increased startle response, *etc.*) and greater emotional arousal. It is well-established that the amygdala has bi-directional connections with visual processing regions (Whalen *et al.* 1998) and, that higher activation of the amygdala in response to emotionally relevant stimuli enhances the contrast sensitivity of visual stimuli and potentiates attention (Phelps *et al.* 2006). This helps to explain why fear-inducing stimulus objects are subject to faster, more efficient, and more accurate processing than non-fear inducing stimuli (Ohman, Flykt & Esteves 2001). Given all of this, we should expect the PAC-system to exert at least an indirect effect on semantic-associative processing *via* increased attention to threat-relevant items and increased processing fluency, thereby facilitating categorization of the relevant stimuli on trials in which BLACK and UNPLEASANT share a response key relative to trials in which BLACK and PLEASANT share a response key for individuals who have the relevant IA_{SA}.

But, again, while it would appear as though both the trait theory and pluralism accommodate the phenomena in the same way, trait theory only does so in a manner that is parasitic on the explanatory value of MMS-SC. The trait view, *sans* a detailed account of the psychological bases of attitudes, is unable to predict, and much less account for, the ways in which affective responses to stimuli with aversive reinforcement value modulate, even indirectly, semantic-associative processes.

3.3 Round 3: Low Predictive Validity

Recall that Machery also claimed that the low-predictive validity of indirect measures is better explained by the view that attitudes are traits than the view that attitudes are mental states. Again, his explanation bears repeating:

> By contrast, on the trait picture, a particular indirect measure taps into one of the many components that determine behavior, such as emotions, associations between concepts, and so on. So, where the Freudian picture posits a single determinant of behavior (*i.e.,* the implicit attitude), the trait picture posits many (*i.e.,* the components of the psychological basis of an attitude). An indirect measure should be worse at

> predicting behavior on the trait picture than on the Freudian picture (in fact it should be a poor predictor) because it only measures one of the many components of the psychological basis of an attitude. (*ibid*: 120)

At the risk of sounding like a broken record, the MSP offers a similar, though, once again, more nuanced explanation for the low predictive validity of indirect measures. Multiple memory systems jointly contribute to the production of various forms of social behavior. Suppose I want to predict how long an individual is going to spend reviewing the application of a minority before making a hiring decision. This decision is likely to be guided by, for example, the IL system, the conceptual system (which is a declarative memory system) and the semantic-associative system and each to varying degrees. So, if I want to use an indirect measure to predict this type of behavior, I had better make sure that this measure taps the contents of these memory systems. Consequently, MSP predicts that an indirect measure should be less predictive of behavior to the extent that it fails to tap the same sets of states and processes that causally contribute to the behavior of interest.

Again, it is worth stressing once more that while the trait theory and MSP both appeal to the heterogeneity of what is being measured in order to account for the low-predictive validity of indirect measures, the explanatory appeal of the trait theory is parasitic on the explanatory value of MMS-SC.

3.4 Objections and Replies

One might worry that I have been unfair to Machery's view insofar as I have treated it as a bona fide theory of implicit attitudes when, in reality, it functions as a suggestion as to how one ought to conceive of attitudes. As such, I have been speaking falsely when I have claimed that the trait view, *per se*, predicts or explains anything of substance regarding implicit attitudes. Thus, it is entirely misleading for me to weigh the trait view against MSP.

At the end of the day, I do think that it would be better to treat Machery's trait view as a suggestion for how to conceive of attitude research as opposed to an actual theory of implicit attitudes (*e.g.*, the RIM model, or MMS-SC). Nevertheless, it is clear that Machery conceives of himself as offering a rival theory of implicit attitudes that is capable of predicting and explaining an array of otherwise puzzling findings. Others, moreover, have also treated his view as such (see Brownstein 2018). In light of this, there is nothing misleading or unfair about my pitting the trait view against MSP. Having said that, it should by now be obvious that it is not the trait view, *per se*, that does any explanatory/predictive heavy lifting; rather, to the extent that any heavy lifting is being done it is being outsourced to the assumption that the psychological bases of traits are heterogeneous. Indeed, Machery certainly could have coupled the trait view with the assumption that the psychological bases of traits were homogenous, in which case his account would have fared no better than any other monist theory.

Lastly, one might have a nagging feeling that there is only a trivial semantic difference between MSP and Machery's trait view: Why not think of the MMS-SC as simply providing an account of the heterogenous psychological bases of traits? Is it not the case that all of the relevant empirical research can be reframed within the trait view? And, if so, would this not trivialize my proposal?

It is telling that Machery, after arguing for his trait view over monist accounts of implicit attitude, feels the need to explain how his proposal is not trivialized by the fact that so much of the relevant empirical research that motivates monism can be reframed within the trait view. This is how he responds:

> Furthermore, this proposal is not trivialized by the fact that current research can be translated, so to speak, into the framework of the trait picture. How empirical research is conceptualized matters, since erroneous conceptualizations lead to false debates. For instance, if attitudes are mental states, the question arises of their causal relations with other antecedents of human behaviors, such as motives and intentions. Are occurrent attitudes causes of motives and intentions, and, through them, of behaviors? Rather, are they directly causing behavior in addition to motives? Such questions are misguided if attitudes are traits. (Machery 2016: 124)

The putative fact Machery's trait view can accommodate all of the empirical findings that motivate monism just as well as it can accommodate the findings that motivate pluralism lends further support to my claim that Machery's trait view isn't really a theory of implicit attitudes at all so much as it is a suggestion as to how to impose order on what otherwise seems to be a chaotic array of empirical facts. But the order that results is largely superficial: it serves only to smooth over and consequently obscure the details of the processes that produce the phenomena that cognitive science is concerned with accounting for.

But there are other substantive differences between MSP and the trait view. According to Machery, "traits do not occur, and thus do not enter in token causal relations, though they can occur in type causal relations" (2016: 112). This is obscure. Either attitudes *qua* traits do causal-explanatory work or they don't. If they do, then Machery owes us an explanation of the kind of casual-explanatory work attitudes do that is not already done by their psychological bases. Even if there is an explanation to be given, then the trait view still commits us to an extraordinarily controversial thesis about the causal efficacy of dispositions. If, on the other hand, attitudes lack causal-powers, then it is unclear why social cognition researchers should care whether or not indirect measures are good measures of attitudes so long as they can be reasonably confident that they are reliably tapping into structures that cause the relevant forms of behavior. So, while this route would seem to secure a victory for eliminativism about attitudes, the victory is hollow. The concept *attitude* that is purged from scientific discourse is one that nobody really cared (or had reason to care) about anyway. MSP, by contrast, does not require a commitment to a controversial thesis about the causal efficacy of dispositions in order to explain the causal-efficacy of implicit attitudes.

Here's another substantive difference. Much of the philosophical literature on implicit attitudes attempts to make sense of how it is that we can be morally responsible for the behaviors driven by mental states of which we have little access and control. A crucial issue is here is that "our" attributions of moral responsibility appear to be keyed to the posits of common-sense psychology (Levy 2014, Mallon 2016). While traits fit comfortably within this framework, it is highly likely that some of the states and processes that MSP posits do not. As such, the shape of one's theory of moral responsibility—where implicit attitudes are concerned—will likely differ substantially depending on the view that one accepts.

4 Concluding Remarks

Before wrapping up, there are two upshots of this approach to social cognition that are worth drawing attention to. The first is that there are a number of outstanding debates, not touched upon here, which MSP could help resolve. For instance, implicit attitudes appear to behave in ways characteristic of both associatively structured representations *and* belief-like propositional structures. Call this the *representation problem*. There are several monist proposals currently on the table: SM treats implicit attitudes as semantic associations; Mandelbaum argues that implicit attitudes are unconscious beliefs (2014); and, both Levy (2014) and Brownstein (2018) argue that implicit attitudes are *sui generis*, though they disagree about the nature of these *sui generis* states. None of these proposals are theoretically satisfying. On my proposal developed elsewhere (see Sechman manuscript), some kinds of implicit attitudes can be treated as having associative structure (indeed, the kinds described in §3 qualify) and some kinds of implicit attitude can be treated as having propositional structure (*viz.* the kind of implicit attitudes that are formed via the hippocampus and related neurological structures that are rapidly formed, have compositional structure, and can be flexibly deployed (see Henke 2010)). Some of the most puzzling findings can be explained via the complex interactions between associative and propositional memory systems, while others can be explained via the complex interactions between multiple kinds of associative memory systems.

The second is that MSP seems to imply—or, at least, strongly recommends— the view that implicit attitudes are not a natural kind. Implicit attitudes, on MSP, have relatively few properties in common *and* each kind of implicit attitude is underwritten by a different causal mechanism (*viz.* its respective memory system). As such, it is highly unlikely that the scientific investigation of any arbitrarily selected subclass of implicit attitudes will lead to generalizations that reliably hold of the entire class. But this is just to say that the concept of *implicit attitude* is not subject to unified scientific explanation. If so, why doesn't MSP warrant eliminativism about implicit attitudes?

While I do deny that implicit attitudes are a natural kind for the reasons just stated, I think that it would be a mistake—at this juncture, anyway—to thereby conclude that we ought to *do* without the *implicit attitude* concept. For example, there is no reason to think that the concept *implicit attitude* has impeded scientific progress—to the contrary, implicit attitude research, under that guise, has taught us a great deal about the various kinds of memory structures that drive the kinds of social behavior that are of central interest to social cognition researchers. Accordingly, the concept *implicit attitude* may be an important *investigative kind* even if it is not a natural kind concept (*cf*, Griffiths 2004). Another possibility is that the concept implicit attitude earns its keep as a *normative kind*. Much of the literature on implicit attitudes more directly concerns a particular subclass of implicit attitudes—*viz. implicit bias*.[20] Implicit bias is fundamentally a normative concept. As such, its unity may be in no way tied to the alleged psychological unity of the particular memory structures of which each implicit bias is an instance; rather, it is plausible that the unity of implicit bias is tied to its normative character (*cf*, Murphy & Stich 1999). On such a proposal, implicit biases might be conceived as those implicit attitudes that drive unwanted or unintended forms of socially (and/or contextually) undesirable behavior. Thus, understanding implicit bias and how to mitigate its effects involves understanding the natures of the various kinds of implicit attitudes. As such, the concept *implicit attitude* may earn its keep in virtue of the relation it bears to the normative kind concept *implicit bias*. The upshot is that while MSP is certainly compatible with eliminativism, it would be, at any rate, premature to purge the attitude construct from scientific theorizing, pending further theoretical investigation.

[20] 'Implicit bias' is to be differentiated from 'implicit discrimination'. Implicit discrimination *is* behavior that is driven by implicit bias.

CHAPTER 4

The FORMAT PROBLEM

In the previous two chapters, I argued that (a) implicit attitudes are best conceived of as mental states, (b) memory systems pluralism better accounts for the core anomalies than any of its extant rivals, and (c) implicit attitude monism is likely false (on the grounds that the assumption that implicit attitudes constitute a homogenous, natural kind is likely false). In this chapter, we shift our attention away from the psychometric concerns that dominated discussion over the last two chapters and toward a debate that more directly concerns the very nature of implicit attitudes. In doing so, I take the mental state picture of implicit attitudes for granted and assume that the memory systems pluralism is, at very least, as plausible a theoretical stance as implicit attitude monism.

The debate that is at the focus of this chapter is broadly as follows. The long dominant view in social psychology is that an implicit attitude is a type of associative structure. This associative view has been challenged by a long neglected but insurgent view according to which implicit attitudes are beliefs or belief-like propositional structures. However, there are several patterns of data that neither perspective satisfactorily explains (call these the *format anomalies*). While members of both camps have attempted to revise their theories in light of these well-established yet recalcitrant findings, others have opted to (a) deny that disputes over the representational format of implicit attitudes are substantive, or (b) develop theoretical accounts on which implicit attitudes are *sui generis* representational structures. Let's call this debate over the representational format of implicit attitudes *the format problem*.

There are two central aims of this chapter: First, I endeavor to make the case that each of the above views are likely incapable of overcoming the format problem for one of two reasons: in each case, the view either fails to explain all of the relevant data or ought to be rejected on broader theoretical grounds. In doing so, I hope to establish a relatively modest conclusion that, of all the options on the table, memory systems pluralism has the best prospects for resolving the format problem, as memory systems pluralism is consistent with the possibility that different species of implicit attitude differ with respect to representational format. Chapters 5 and 6 are largely devoted to establishing that memory systems pluralism actually delivers the theoretical goods.

This chapter proceeds as follows. I begin by describing the dueling predictions made by the associative perspective and the propositional perspective respectively (§1). We'll then turn to look at how these two sets of predictions fare in light of the relevant data (§2). Next, we turn to look at some of the strategies that have been pursued in attempting to resolve the format anomalies beginning with the view according to which the format problem is mere terminological dispute (§3) and ending with the *sui generis* perspective (§4). The chapter concludes with a brief discussion of the prospects of MSP theory *vis-à-vis* resolving the format problem (§5).

1 The Dueling Perspectives on Representational Format

There are two dominant perspectives in social psychology with respect to the representational character of implicit attitudes— viz. the associative perspective and the propositional perspective. The associative perspective's influence, despite its historical dominance, appears to be on the downswing relative to the influence of the propositional perspective (for a review of this burgeoning area of research, see Corneille & Stahl 2018). In this section, I will characterize these two perspectives in general terms, turning to consider specific models only when warranted.

1.1 The Associative Perspective

The associative perspective assumes that implicit attitudes are semantic associations (Gawronski & Bodenhausen 2006, Rydell & McConnell 2008). The dominant view in social cognition research is that associations are mental states that link two concepts (*e.g.,* SALT and PEPPER) such that the activation of one is typically sufficient to activate the other. Elsewhere, we have called and will continue to call associations that stand between lexical concepts *semantic associations*. While some authors maintain that associations can also stand between a concept (*e.g.,* SPIDER) and a valence (*e.g.,* bad), Carruthers (2017: 69) points out that "the ontological status of these postulated clusters is opaque. How does a semantic representation 'associate' with an affective state, or vice versa?" This question, while rhetorical, retains its force as those who posit such structures have not made any serious attempt at describing the kind of mechanism that would give rise to their existence. Carruthers' concern, however, does not arise for those who maintain that implicit attitudes are sometimes associative structures that hold between, say, a representation of a social category (*e.g.*, BLACK) and a semantic representation of valenced concept (*e.g.,* GOOD) as both relata have the same ontological status on such views. Given the metaphysical import of this distinction, it is unfortunate that this is a distinction that is rarely observed in the literature. In any case, for the purpose of this discussion, we'll labor exclusively under the dominant associative view on which implicit attitudes are associations that link two concepts.[21]

[21] One might be concerned that by restricting our attention in this way, I have unfairly stacked the deck against standard monist versions of associationism in favor of a pluralist associationism. After all, theorists posit associative structures that link concepts with valences in order to do the sort of explanatory work that a view on which implicit attitudes are associations only between concepts could not. But this isn't quite right. For instance, Amodio and Devine (2006) claim that there are two types of implicit attitude— implicit evaluations and implicit stereotypes. On one standard reading of this distinction, implicit evaluations are associations between a concept and a valence, while an implicit stereotype is an association between two concepts. Insofar as Amodio and colleagues' view is best understood as a version of pluralism, then I do just as much to limit the explanatory resources available to certain forms of pluralism as I do monism by insisting that we focus on associations between concepts. What I am urging, then, is that we should only posit a mental state if its ontological status is not opaque in the way that the ontological status of a concept-valence structure is opaque and, by making this plea in no way do I unfairly stack the deck against monism in favor of pluralism full stop.

To say that a mental state that links two concepts in this way has associative structure is to say that the tokening of one such concept (*e.g.*, CAPITALISTS) will, *ceteris paribus*, token the concept that it's associated with (*e.g.*, CRIMINAL). While it is true that, on certain propositional views prevalent in the literature the tokening of the *propositional structured thought* CAPITALISTS ARE CRIMINALS likewise involves the tokening of its constituent concepts CAPITALIST and CRIMINAL, the two perspectives differ with respect to how the tokening of the relevant concepts unfolds. Associationism assumes that activating the concept CAPITALIST results in the activation of the concept CRIMINAL via the process of the *spreading of activation* across a net of linked associations. The propositional view, by contrast, assumes that the tokening of the constituents of a propositional structure does not involve a spread of activation (De Houwer 2018, Mandelbaum 2015).

Whereas the propositional structure CAPITALISTS ARE CRIMINALS is generally regarded as reflecting an agent's belief that capitalists are criminals, a CAPITALIST-CRIMINAL association reflects the frequency with which one's concept CAPITAL has been activated alongside CRIMINAL in the agent's learning history (Brownstein 2017, Mandelbaum 2015). In other words, the presence of a propositional structure reflects the way in which the subject takes the world to be, while the presence of an associative structure merely informs one about the history of the causal and temporal sequences of the activation of the relevant concepts in one's mind (Mandelbaum 2015).

While it should be readily admitted by all parties that the question of whether implicit attitudes are propositional structures or associative structures cannot be answered by direct observation, there is general consensus about the kinds of behavioral effects one should expect to observe on the assumption that implicit attitudes are associations versus the assumption that implicit attitudes are propositional structures. Each camp, in other words, makes different predictions about how implicit attitudes change over time. Generally speaking, those who subscribe to associationism predict that formation, change, and the behavioral expression of implicit attitudes is governed by associative principles. Those who subscribe to the propositional view, by contrast, tend to emphasize the ways in which implicit attitudes are formed, updated, and expressed in accordance to more traditional computational processes.

With respect to attitude *change* the associative camp recognizes at least two associative principles that govern this phenomenon (Mandelbaum 2015a). The first is *extinction*. Extinction is the process whereby the link between two semantic representations is weakened or eliminated via repeated presentations of one stimulus without the associated other (*extinction trial*). *Ceteris paribus*, the strength of an extinction effect is generally regarded as being directly proportional to the number of extinction trials presented in a given learning session. The second associative principle frequently discussed is that of *counterconditioning*. Counterconditioning, in this context, is the processes through which the valence (+/–) associated with a given stimulus (CS) is changed via repeated pairings of the CS and a second stimulus or outcome of opposite valence. For instance, if someone has a SPIDER-BAD association, then the goal of a counterconditioning procedure would be to replace this association with a SPIDER-GOOD association via the repeated paring of spiders with positive stimuli. As with extinction, the going assumption is that the strength of the counterconditioning effect is going to be directly proportional to the number of counterconditioning trials administered over a given experimental session.[22]

Given that these are the two widely accepted principles that govern implicit attitude change on the associationist perspective, we should expect to observe rigid updating patterns. To say that

[22] It must be stressed that I am presenting this discussion as it is typically presented in the literature. And this way of presenting the distinction between extinction and counterconditioning smooths over a great number of controversies concerning the rate of extinction/counterconditioning, the precise nature of the learning algorithms that produce extinction/counterconditioning effects, and critical differences across experimental procedures for bringing about counterconditioning effects in particular. Moreover, this way of presenting the distinction also ignores the possibility that one can produce behavioral effects generally conceived of as typical of counterconditioning/extinction via the formation of distinct token associative structures that, under specific contexts, compete with existing associative structures for behavioral influence. I don't make much of these issues here, but they will play a major role in motivating the superiority of memory systems pluralism relative to its rivals in Chapter 6.

implicit attitudes are expected to change in comparatively rigid ways is, in this context, to mean that they are expected to (a) update incrementally overtime through extinction and/or counterconditioning, (b) be insensitive to the meanings of other mental states, and (c) be sensitive exclusively to various forms of co-occurrence relations between the items, features, and/or categories that the associated relata represent. Put another way, these three more specific predictions would be violated if careful empirical investigation uncovers evidence that implicit attitudes (a*) sometimes update rapidly in exposure to relevant information, (b*) are sometimes sensitive to the meanings of other mental states, and (c*) are sometimes sensitive to relations that cannot be best described as types of mere co-occurrence relations.

Now, why would a theorist want to endorse an account of implicit attitudes according to which they update in these various ways? The rationale for these predictions is derived from traditional dual systems theorizing. On the dual systems perspective, there are two kinds of cognitive system. The first kind, popularly known as System 1, is typically conceived of as a slow learning associative system that tracks pertinent environmental regularities. The second kind of system, popularly known as System 2, is a fast learning and so resource dependent learning system that is capable of flexibly updating its stored representations in response to unexpected changes circumstance. While most dual systems models permit various interactions between systems, they nevertheless posit that these systems are capable of acting autonomously *and* that System 1— due to its being computationally cheap and efficient— operates by default (Kahneman 2011) thereby freeing up the comparatively more resource dependent System 2 to intervene in contexts that require more effortful or deliberative responses.

With that said, dual-systems accounts of implicit social cognition have largely fallen out of favor for what are generally regarded as more parsimonious single-system dual-process accounts of social cognition. Such views attempt to preserve the appeal of dual-systems theorizing while tossing out the theoretically cumbersome assumption that associative and propositional processes operate in parallel over distinct memory stores (*e.g.,* the Associative-Propositional Evaluation (APE) Model; Gawronski & Bodenhausen 2006, and the Reflective-Impulsive Model (RIM); Strack & Deutsch 2004). Though the discussion of the associative perspective that follows will be framed exclusively in terms of dual-process theory, it's being the more liberal perspective implies that many (but not all) of the problems that arise for dual-process theory will also arise for dual-systems theory.

1.2 The Propositional Perspective

Let's now turn to the predictions of the propositional perspective. As I had already mentioned, the propositional perspective assumes that implicit attitudes are propositionally structured mental states. Readers familiar with this literature might balk at this description, wondering why I break from the terminological precedent set by others (*e.g.*, Brownstein 2018) by not calling this something like *the belief view*. This is worth briefly addressing as my reasons for breaking from convention are relevant to the discussion that follows.

One reason for breaking from convention stems from the simple fact that that I'd like to avoid needlessly wading into controversies over the nature of *belief*. Mandelbaum (2015) for instance, has recently argued that implicit attitudes are beliefs but, in doing so, is committed to denying standard assumptions about the normative profile of belief (see Brownstein 2018: 73-81). The standard view, sometimes referred to as the *truth-taking view* of belief, assumes that beliefs are attributed to agents "on the basis of agents' reflective judgements and avowals" (Brownstein 2018: 73). Because Mandelbaum defends view on which beliefs are sometimes (or, perhaps, always) formed automatically and unconsciously the moment one considers or entertains a proposition, the question arises as to whether implicit attitudes so-construed are *really* beliefs. The question of representational format of implicit attitudes, however, is more fundamental than these questions concerning the truth-conditions of belief attribution: should implicit attitudes, for example, turn out to be associations, then the question of whether implicit attitudes are beliefs never even arises. Similarly, as this controversy establishes, the putative fact that implicit attitudes express propositions does not entail that implicit attitudes are beliefs. As I have no interest in adjudicating disputes about the nature of belief and such disputes play no role in the ensuing discussion, there is no downside to framing the discussion in the way that I have.

The second reason is that the working social psychologists who labor under the propositional perspective tend not to describe their models using the language of belief. Jan De Houwer, who has been at the vanguard of the propositional movement for over a decade, offers this description of the propositional approach:

> A propositional model of implicit evaluation postulates that a stimulus can evoke an evaluative response automatically only after a proposition about the evaluative properties of the stimulus has been formed or activated automatically (Hughes, Barnes-Holmes, & De Houwer 2011). As I noted in the previous paragraphs, there might be different processes by which propositions can be formed or activated automatically. For instance, the sight of ice cream on a hot summer day could automatically retrieve memories in which the proposition "ice cream is good" was encoded. (De Houwer 2014: 345),

To those familiar with standard conceptions of belief, this account of implicit attitudes appears to fit those standard conceptions. However, and this brings us back to the first concern, whether such mental states are *really* beliefs depends on a set of background assumptions about the nature of belief that are largely orthogonal to the discussion at hand. Given that the debate between propositionalists and associationists has, arguably, not been hindered by the fact that the conceptual debate over what counts as a belief is without resolution, to frame the theoretical central controversy as one between those who take implicit attitudes to be associations versus those who take implicit attitudes would be to frame the controversy in a way that obscures more than it illuminates.

Lastly, it is worth stressing that as with the associative perspective, the propositional perspective is a theoretical perspective that encompasses multiple distinct theories. At present, there are several versions of the propositional view on the market, each of which generates different predictions as to how implicit attitudes ought to behave in various circumstances. For instance, whereas De Houwer's *single-process propositional model* of implicit attitudes assumes that propositional formation is always a conscious affair, Mandelbaum's (2015) propositional account rejects this assumption. De Houwer's model assumes that all implicit attitudes are encoded in the same memory system so as to form a single web of inferentially integrated propositional structures, which is an assumption that, again, Mandelbaum rejects. These differences, as just alluded to, carry empirical significance. To give just one example, results which show that implicit attitudes do not update inferentially under optimal conditions would pose a challenge to De Houwer's model but not necessarily Mandelbaum's (see discussion in §4 of this chapter). With that said, the predictions of the two accounts largely converge. Thus, in the following discussion, I'll smooth over the major differences between the various propositional accounts, attending to them only when mandated by the discussion at hand.

As stressed earlier, proponents of the propositional perspective tend to look to for empirical support from studies that reveal that implicit attitudes change in content-driven ways over the course of one or many experimental sessions (Brownstein 2017, Levy 2014, De Houwer 2014). But what counts as a *content-driven change* as opposed to a change in accordance with associative learning principles? While there has been no serious attempt at explicating the notion of a content-driven change, proponents of the propositional perspective maintain that content-driven changes are evident when, for example, (a) the content of a mental state is sensitive to the precise semantic relation that obtains between a pair of items, (b) the content of a mental state shifts rapidly in response to newly acquired information, and (c) the content of a mental state changes by virtue of its being appropriately inferentially integrated with the subject's other mental states. The general idea is that content-driven changes occur when the content of one mental state changes in virtue of the content of another propositionally structured mental state (Brownstein 2017, Levy 2014). More detailed descriptions of these kinds of content-driven change will be provided when we discuss the various studies that the propositional theorist marshals in support of this view.

Before moving on, it is worth highlighting a critical difference between the predictions made by the propositional perspective and those predictions made by the associative perspective. There is an important sense in which the associative perspective's predictions are riskier, in the Popperian sense, than the propositional theorist's predictions. While it is generally acknowledged that content-driven changes are inconsistent with the generic associative view, propositional theorists do not claim that propositionally structured representations change *only* in content-driven ways: most models that adopt the propositional perspective also allow propositional structures and their constituents to undergo changes in accordance with classic associative principles (De Houwer 2014, Brownstein 2018, Mandelbaum 2015).[23]

[23] Having said that, propositional models do struggle to generate independently motivated predictions about the conditions under which we should expect implicit attitudes to change in a strictly associative manner. So, while propositional models can often be made consistent with data that reveal associative updating, it's not obvious that these models are able to successfully *predict* associative-friendly patterns of data. This has led some on the associative side of the debate to complain that propositional models are so underspecified that they are unfalsifiable. Interestingly, there are moments at which De Houwer (2018), the most prominent propositional theorist, appears to concede these charges while, in the same breath, defending the explanatory superiority of the propositional perspective over the associative perspective. While this concessionary move may, *prima facie*, seem surprising, it becomes less so when we consider that the various ways in which the associative theorist has attempted to account for the data that most impress propositional theorists credibly exposes the associative theorist to similar charges. We return to this issue in §4.2.1.2.

2 The Format Anomalies

The associative versus propositional debate has given rise to a burgeoning literature that documents the various ways in which implicit attitudes appear to change in ways that accord with the predictions of both the associative and the propositional perspectives (for a review, see Brownstein 2018, Corneille & Stahl 2018, Gawronski & Bodenhausen 2018, Gawronski, Brannon & Bodenhausen 2018). In this section, I will argue that both perspectives fail to account for all of the relevant effects described in this literature by focusing on the three empirical controversies that have received the most attention in the literature. If I am right that neither perspective can yield satisfying explanations of each of the three controversies, then we are justified in moving beyond these two standard perspectives in our search for an empirically adequate theories of attitudes.

2.1 The Rate of Change Anomaly

Recall that the associative perspective, generally speaking, assumes implicit attitudes will update gradually and incrementally in response to newly acquired information. This is, in part, because the associative principles that govern the formation and updating of such mental states track relevant environmentally embedded statistical contingencies between the associated relata. The propositional perspective, by contrast, assumes that implicit attitudes will sometimes update rapidly in response to newly acquired information since implicit attitudes *qua* propositional structures represent states-of-affairs.

Rydell and McConnell (2006) conducted a series of influential studies specifically designed to test the associative perspective's prediction that implicit attitudes will update gradually in response to counter-attitudinal information. In an initial impression formation block, subjects were presented with 100 statements each of which described Bob as performing a positive action or a negative action and were tasked with stating whether each statement was characteristic or uncharacteristic of Bob. In the relevant experimental condition, all statements reinforced a positive impression of Bob. All subjects completed both an implicit and an explicit attitudinal measure to ensure that all subjects formed a positive impression of Bob. In a second block of learning trials, which we'll refer to as the *counter-conditioning* block, subjects were either exposed to 20, 40, 80, or 100 statements each of which reinforced a negative impression of Bob All subjects then completed a second round of implicit and explicit attitudinal measures. McConnell and Rydell found that while explicit attitudes toward Bob showed a complete reversal after only 20 counter-attitudinal trials (*i.e.,* explicit attitudes changed rapidly in response to the newly acquired information), implicit attitudes updated incrementally in direct proportion to the number of counter-attitudinal trials, which is precisely what the associative perspective but not the propositional perspective predicts. Let's call this the *incremental* effect. While the propositional perspective can accommodate this data, no model in that family predicts these results. Moreover, even if some version of the propositional model were to predict these results, such views have difficulty explaining why it is that explicit attitudes (which, are also assumed to reflect propositionally structured representations) tend to update far more rapidly than implicit attitudes.[24]

Rydell and McConnell's studies, however, only tell one side of the story. Cone and Ferguson (2015) conducted a series of studies that do demonstrate that implicit attitudes can undergo

[24] Here I only summarize two sets of Rydell and McConnell's findings that pose trouble for the propositional view. But there are others. For instance, Rydell and McConnell (2006) find that explicit attitudes, but not implicit attitudes, are influenced by the goals that subjects adopt when tasked with forming an evaluation of Bob. The associative perspective better accounts for these results than the propositional perspective. Rydell and McConnell (*ibid.*) also find empirical support for the hypothesis that information presented below the conscious threshold influenced subjects' implicit evaluations of Bob but not their explicit evaluations. Again, these results are better accommodated by the associative perspective than the propositional perspective.

extremely rapid updating over the course of a similarly structured impression formation task. Employing an initial learning paradigm analogous in relevant respects to the one employed by Rydell and McConnell (in which all subjects were trained to form a positive impression of Bob using 100 statements describing Bob's behavior), Ferguson and colleagues found that a single extremely negative behavioral statement that is highly diagnostic of Bob's character (*e.g.,* Bob molests children) was sufficient to produce a complete reversal of indirect measure performance. The propositional perspective, but the not the associative perspective, predicts this result.

A follow up study reported by Cone and Ferguson (2015) lends further support for the propositional perspective over the associative perspective. The associative perspective predicts that the mere pairing of Bob with a negative behavioral description regardless of the actor should be as effective in shifting attitudes about Bob as one that directly attributes a negative behavior to Bob (see the discussion of the Co-occurrence/Relational Controversy below). Put differently, the associative perspective predicts that mere *guilt by association* should result in a shift in attitudes toward Bob. If implicit attitudes are content-sensitive as the propositional perspective predicts, then guilt by association should not be as effective in shifting subjects' Bob-attitudes as one that attributes the behavior to Bob. In a test of these dueling predictions, using the same initial learning phase (100 statements plus feedback designed to reinforce a positive impression of Bob) Ferguson and colleagues exposed subjects in the guilt by association condition were shown a single extremely negative behavioral description attributed to Bob's coach ("Bob's coach was a child molester"), subjects in the diagnostic condition were presented with a single behavioral statement, matched in valence, attributed to Bob ("Bob is a child molester"), and subjects in the control condition were exposed to a neutral statement (*e.g.,* "Bob chews bubblegum"). The propositional perspective's predictions were borne out by the data: the only subjects to show a complete reversal in indirect measure performance were those in the diagnostic condition.

2.1.1 The Associative Perspective and the Diagnostic Effect

The associative perspective readily handles Rydell and McConnell's findings. Indeed, as already stressed above, the associative perspective straightforwardly predicts these findings whereas the propositional perspective is merely consistent with them. The classical associative perspective, however, struggles mightily to account for Ferguson and colleagues' findings. However, some have attempted to do so and a brief discussion of both these attempts and the reasons they fail will prove instructive in what follows.

Those in the associative camp pursue two broad strategies in response to Ferguson and colleagues' findings, the direct strategy and the indirect strategy. The direct strategy attempts to respond to the challenge posed by Ferguson and colleagues' findings head on by deploying the tools afforded to them by dual-process theorizing. The indirect strategy, by contrast, attempts to show that any putative inconsistency between the associative perspective and Ferguson and colleagues' findings is illusory by denying that the associative versus propositional debate is one of substance. The indirect strategy is covered extensively in 4.3, so we'll bracket this approach until then. Here, and for the remainder of this section, we'll consider only those responses that adopt the more direct strategy.

The standard associative assumption is that exposure to the statement "Bob is a child molester" should weaken the BOB-GOOD association. However, one could reject this assumption and claim, as some associative theorists have in other contexts, that exposure to the target statement results in the formation of a new BOB-BAD association (see Gawronski & Bodenhausen 2018). On this picture, it is the latter association that mediates indirect task performance.

This response suffers from two major objections. First, this account doesn't explain why the newly formed, weak BOB-BAD association wins out over the much stronger BOB-GOOD association in the competition for behavioral expression. The associative theorist might respond that each token associative structure is indexed to a particular context (the process whereby this occurs is known, fittingly enough, as *contextualization*). On the contextualization hypothesis, the BOB-BAD associative drives indirect responses because the context in which subjects' indirect responses are recorded is the same context in which they were exposed to the novel diagnostic information. In other words, the two associative structures are not competing for behavioral expression as they are indexed to different contexts.

But the contextualization response is not without its own difficulties. Cone and Ferguson (2014: 53) "made every effort to ensure that the circumstances under which participants learned Time 1 and Time 2 information were essentially identical and that no obvious contextual cues were available to allow participants to make sense of the evaluative inconsistencies to which [the experimenters] exposed them." Furthermore, in a follow-up study, Bannon and Gawronski (2019) replicated Cone and Ferguson's results and found no evidence of contextualization effects.

Here's the second objection: even if the molester statement does result in the formation of a novel associative structure, this hypothesis does not itself explain why the effect of diagnostic information on evaluative responses is more pronounced than guilt-by-association information. After all, recall that the associative perspective assumes that subjects who are exposed to the statement "Bob's coach is a child molester" should have also automatically formed a BOB-BAD association given that Bob is described in that statement as being associated with a child molester. Since each association was formed in response to a single statement of counter-attitudinal information, there is no reason to expect that the BOB-BAD association formed *via* guilt by association information would differ in associative strength from a BOB-BAD association formed via diagnostic information.

Perhaps the dual-process theorist could respond by denying that the guilt-by-association statement induces the formation of a BOB-BAD association; rather, so the story goes, the guilt-by association-information (a) results in the formation of a COACH-BAD associative structure and (b) elicits the propositional inference that that the guilt by association information contains no relevant information about BOB. Since several prominent dual-process models assume that novel associations can be formed via propositional inferences and if (b) led to no change in the structure or strength of the underlying associative network, then this could explain why subjects' implicit evaluations toward Bob in the guilt-by-association condition were not significantly different from those in the control condition. Maybe. The chief concern with this account is that while implicit evaluations of Bob in the guilt-by-association condition were not different from those in the control condition, subjects in the guilt-by-association condition *explicitly* reported liking *Bob* less after finding out that his coach was a child molester. So, if (b) obtained, it is not clear why this inference would have no effect on subjects' implicit evaluations of Bob but a weak but significant effect on explicit evaluations.

2.1.2 The Propositional Perspective and the Incremental Effect

While the propositional account readily handles the diagnostic effect, it struggles to account for the incremental change effect. How is the propositional perspective to explain why indirect task performance exhibits incremental, linear change in response to the number of trials in the counter-conditioning phase but explicit task performance exhibits comparatively rapid change in response to only a counter-conditioning trials? Curiously, propositional theorists have largely ignored the challenge that the incremental effect poses to their view. Nevertheless, at least three strategies are available to the propositional theorist.

The first strategy assumes that the indirect measure performance accurately reflects the subject's true attitudes toward Bob while the explicit measure performance provides a distorted picture of subjects' Bob attitudes. To flesh this out, let's assume for the sake of argument that a Bayesian updating algorithm could successfully explain the linear updating signature revealed by indirect measure performance (Rydell & McConnell 2006). The propositional theorist could then go on to claim that subjects' responses to the direct measure merely reflect subjects' motivations to tell the experimenter what the subjects assume that the experimenter wants to hear. If both assumptions hold, then the incremental effect poses no serious threat to the propositional perspective.

One should find this approach unsatisfactory for a couple of reasons. The first is that the claim that subjects' explicit measure performance were due to demand characteristics is entirely unmotivated. Suspicions about subjects deliberately altering their responses to satisfy experimenter's goals are generally warranted when an accurate expression of one's attitudes may violate some social norm (*e.g.,* when stating how one actually feels about a target might be construed by the experimenter as racist, sexist, or homophobic). But such concerns do not appear to be salient here. Consequently, this response appears entirely *ad hoc*. Second, this sort of skepticism over the value of explicit measures is generally at odds with the assumptions that motivate a great deal of research under the propositional perspective. Generally speaking, propositional theorists tend to be more skeptical of the value of indirect measures with respect to providing insight into subjects' "true" attitudes toward an attitudinal object than they are of the value of direct measures with respect to the same.

A second strategy would be to affirm the value of explicit task performance in understanding subjects' attitudes toward Bob but adopt a skeptical stance toward indirect task performance.

Alternatively put, this response assumes that, in this context, the IAT used in Rydell and McConnell's studies do not provide us with any meaningful insight into subjects' attitudes toward Bob. How might this strategy be pursued? Several prominent propositional theorists have recently proposed that the IAT performance specifically can be shaped by (a) subjects' initial decisions, influenced by experimental context, as to how to classify the targets that appear on the IAT, and (b) subjects' motivation to simplify an otherwise complicated and cognitively taxing task. If both assumptions hold, then the standard assumption that the IAT provides unambiguous and direct insight into subjects' attitudes is unwarranted. This general approach to interpreting IAT performance has been called, for reasons to be discussed later on, the *similarity construction account* of IAT performance (Bading *et al.* 2019).

To illustrate the proposal at hand, let's assume that after completing the counter-conditioning phase of Rydell and McConnell's studies, subjects decide to think of Bob negatively. This initial classification decision might be driven either by a genuine shift in attitude toward Bob as reflected by explicit measure performance or by a demand characteristic analogous to the one described in the context of the first strategy. Either way, this initial decision alone would serve to simplify the IAT task: in completing the IAT subjects will now have to focus only on responding in a particular manner when images of Bob and the presentation of negatively valenced words share a particular response key, thereby resulting in more accurate and rapid classifications of stimuli on blocks where Bob-stimuli and negative stimuli share the same response key compared with the blocks in which they are assigned to separate keys. The implication is that had task demands made salient Bob's positive character, subjects would have performed the IAT differently. This stands in stark contrast to the received view about the drivers of IAT performance. On the received view, it is assumed that subjects are simply motivated to complete the task corresponding to each IAT block as presented, where this does not involve any assumption that IAT performance is mediated by the deliberate adoption of any particular response strategy. Thus, to the extent that the IAT performance is relevant to understanding subjects' Bob attitudes, the IAT only achieves such relevance because subjects have shifted their attitudes toward Bob at the explicit level and adopted a response strategy on that basis. IAT performance, then, is not driven by the activation of BOB-BAD associative structures in memory.

We have good reason to reject this strategy. First, the similarity construction account fails to explain linear change in subjects' IAT performance as the number of counter-attitudinal statements increases. The account, as articulated anyway, seems to predict that that subjects' responses should not shift in direct proportion to the number of counter-attitudinal trials that are included in the various counter-conditioning phases. Perhaps as the number of counter-attitudinal trials increases, so too does the salience of Bob's badness and this increase in salience in turn determines the strength with which the response strategy is adhered to, but this would (a) not only be an ad hoc response to this challenge but also (b) serve to reintroduce the tension between indirect measure performance and direct measure performance. In regard to (b), after all, if we assume that explicit measure performance does accurately capture subjects' attitudes toward Bob and these attitudes are moderately negative after only 20 counter-conditioning trials, then would expect that Bob's negativity is sufficiently salient to produce a proportional shift in indirect measure response especially, after, say 60 trials. But this, of course, is not what we find. Instead, explicit measure performance changes little after 40 trials while IAT performance continues exhibit consistent degrees of attitude change up to 100 trials.

Lastly the propositional theorist might claim that both measures provide important information about subjects' attitudes, albeit different types of information. For instance, De Houwer claims that while all attitudes are propositionally structured propositions that are formed consciously (even if automatically), the process governs *retrieval* of these structures and behavioral *expression* of these structures may end up mimicking the behavior of associative models. De Houwer elaborates on this suggestion in the passage below.

> For instance, the idea that propositions can be activated automatically from memory allows propositional models to mimic the behavior of an association-based network. Assume that a stimulus in the environment (*e.g.*, a particular medicine) automatically activates a proposition in memory (*e.g.*, the proposition that the medicine cures cancer). Under certain conditions, on the related concepts of the proposition might be retrieved (*e.g.,* "medicine" and "cancer") but not the relational information (*e.g.*, "cures). Under those conditions, implicit evaluation would seem to be unaffected by relational information (*e.g.*, the medicine might evoke a negative rather than a positive implicit evaluation) even though the evaluation is mediated by the (partial) activation of a propositional representation).[25] (De Houwer 2014: 346)

Let's bracket pressing questions over why, how, and when such automatically retrieval processes would result in the partial activation (whatever that means exactly) of a propositional structure and focus on the plausibility of the proposal as stated. The proposal, simply, is that an attitude *qua* proposition need not be retrieved, under all circumstances, as such.

One might complain that such a model is empirically indistinguishable from dual process models so as to make such a model unfalsifiable. De Houwer (2014) is sensitive to this concern:

[25] It is worth keeping this proposal in mind as we will revisit this position when we discuss the options available to the propositional theorist in responding to the co-occurrence/relation controversy in 4.2.3 below.

> Although this idea generates a number of interesting testable hypotheses (*e.g.*, that the impact of relational information depends on the extent to which the environment acts as a retrieval cue for information), it would be difficult to disentangle a propositional model that allows for such a partial automatic retrieval from a dual process model that allows for both propositions and associations as mediators of implicit evaluation […]
>
> Nevertheless, even if propositional (or association-activation or dual process) models can never be proven true or be falsified, it remains interesting to exploit the heuristic and predictive power of these models. (*ibid*: 346-347)

Put another way, De Houwer's considered position seems to be he's willing to sacrifice the theoretical virtue of falsifiability for the theoretical virtues of explanatory scope and fruitfulness. While his sanguine response to this sacrifice might strike some readers as odd, it is worth nothing that similar complaints are routinely raised about the APE model, the leading dual-process alternative (Gawronski & Bodenhausen 2018).

With that said, De Houwer too hastily concedes that that his account is unfalsifiable. To see this, consider that unless De Houwer is willing to posit an associative mechanism that tracks the frequency with which Bob is paired with negative statements, De Houwer's model fails to account for the incremental effect. Assume that subjects in Rydell and McConnel's subject, after the counterconditioning phase, form the propositionally structured representation BOB IS BAD and that partial activation of the proposition results in the activation of BOB and BAD. What non-associative mechanism could De Houwer appeal to in explaining why the activation of BOB and BAD produces a linear change in IAT performance as the number of counter-attitudinal statements increases? Without such a mechanism, it's difficult to see how this model could avoid predicting stable IAT performance across experimental conditions. And that's putting it mildly. The good news for De Houwer then, is that a partial activation model is more substantive in this respect than he is inclined to give it credit for; the bad news is that, *sans* an additional bit of associative machinery, the model generates the *wrong* predictions.

2.2 The Inferential Updating Controversy

Now we move to address the empirical controversy that flows from attempts to answer the following question: is indirect task performance influenced or driven by inferential processing? As with each of the controversies, extant empirical data paints a mixed picture. Some of the earliest investigations of this question suggested that the answer is no. For instance, several studies suggest that when subjects are presented with statements that express common cultural stereotypes (*e.g.*, Women are bad at math) each of which is immediately followed by feedback that invalidates the target stereotype (*e.g.,* FALSE!), this has the effect of strengthening the stereotype (Peters & Gawronski 2011). *If* implicit attitudes were influenced by inferential processes (as opposed to mere associative processes) like negation, then we should have observed a weakening of the negated stereotypes. More recently, the results of several studies suggest that implicit attitudes can be influenced by inferential processing in highly robust and surprising ways (see Mann & Ferguson 2015). We'll take a closer look at the results of some of the more compelling studies that have that have been conducted with the aim of answering the above question.

In one influential study in a series of experiments conducted by Gregg, Seibt and Banaji (2006), subjects were presented with positive or negative information about two fictional social groups (the Niffites and the Laapians) before completing measures of implicit and explicit attitudes. Once subjects' responses had been recorded, the experimenter explained that there was something of an experimental mix-up and that the subjects were given incorrect information about the groups: the negative social behaviors that had been attributed to the Niffites are actually attributable to the Laapians and *vice versa*. After being presented with this information, subjects' responses to the same set of measures were recorded. The goal of this second round of measures is to determine what effect, if any, this newly presented information had on subjects' evaluations of the two groups. The researchers found that self-reported attitudes toward the two groups reflected the newly acquired information (*e.g.,* where the self-reported attitudes of the Niffites had previously been positive, they became negative. The same effect was observed, *mutatis mutandis*, for the Laapians.). However, the information had no effect on subjects' implicit evaluations. To rule out the possibility that subjects didn't actually believe that a mix-up had occurred, the same researchers conducted a follow-up study that included a more plausible story (subjects were told that the character of the two groups shifted over time in response to changing social and environmental conditions). This study revealed the same patterns of results. That only explicit attitudes shifted in response to the new information was interpreted as evidence that implicit attitudes, once they had been formed through conditioning, cannot be undone through inferential processes. If they were apt to undergo such changes, then we would not have expected this information to have a differential effect on explicit compared to implicit attitudes. Associative dual-process models, but not propositional models, straightforwardly predict these results.

But, as mentioned at the top of this section, Gregg and colleagues' data do not tell the whole story. Using a novel experimental paradigm, Mann and Ferguson (2015) conducted seven experiments and found in each one that implicit attitudes do update in response to newly acquired information in a manner that seems to implicate inferential processing. Here's how the researchers describe this paradigm:

> In each experiment, participants read a story, presented one sentence at time, about an individual named Francis West who is described as breaking into and causing damage to two homes. Participants' implicit evaluations toward Francis West are then measured. Afterwards, they read a final piece of information which either maintains the gist of what they already read (control conditions in which Francis remains negative) or dramatically reverses is (experimental condition in which Francis becomes positive) by offering a reinterpretation of what was previously learned: The houses were on fire, and Francis was searching for two young children who he knew was inside. (Mann & Ferguson 2015: 826-827)

The researchers demonstrated that this reversal occurs using two different measures of implicit attitude (Experiment 1a exploited the *affective misattribution procedure* (AMP); Experiment 1b exploited the IAT). In Experiment 2, the researchers found that reversal occurs only when the newly acquired information occasions subjects to reinterpret the previously acquired negative information in a positive light. Perhaps most relevant to the discussion at hand, Experiment 3 found that implicit attitude reversal did not occur when subjects were placed under heavy cognitive load (suggesting that ample cognitive resources are necessary for reinterpretation to occur).[26] The remaining two

[26] A standard methodological assumption is that controlled processing, but not automatic processing, is inhibited by cognitive load (Evans 2008). In line with this assumption, when one finds that an increase of cognitive load is correlated with a change in task performance, one tends to infer that the processes involved in indirect task performance are controlled (or, at least, inferential). Recent research on the so-called *default mode-network* suggests that this standard methodological assumption is incorrect (Jenkins 2019). The default mode network (DMN) exhibits higher baseline activity relative to other brain regions and tends to deactivate when people direct their attention to a variety of goal directed tasks (*ibid*: 531). A number of findings suggest that sufficiently substantive increases in cognitive load may disrupt social-cognitive processes *regardless* of whether they are automatic or controlled (see Jenkins 2019 for a survey). This theoretical perspective produces an alternative hypothesis that could explain the reinterpretation effect: when the new information about Francis West was provided to subjects in the high load conditioned, they failed to update their interpretation of Francis West not because subjects were unable to engage controlled inferential processing but because high cognitive load inhibited impression updating *simpliciter*. So long as it is taken for granted that inferential processes are always controlled cognitive processes *and* high load inhibits those brain regions necessary for, among other things, impression formation and updating, then we should not infer from Mann and Ferguson's results that reinterpretation effect is a product of the operations of controlled cognitive processes.

experiments address questions concerning whether reversal occurs because newly acquired information changed how the subjects understand the previously presented information (Experiment 4) and questions concerning the temporal stability of these rapid changes (Experiment 5).

The propositional perspective straightforwardly accounts for the relevant findings. Moreover, Experiment 3 revealed that reinterpretation occurs only under the no cognitive load and low cognitive load conditions does support the claim that inferential processes were instrumental in the updating of subjects' implicit evaluation of Francis West (but see fn26). The associative perspective, by contrast, struggles to account for any of these findings. While Brannon and Gawronski (2019) considered the hypothesis that these results reflected a contextualization effect, data obtained through their own follow-up studies seems to rule out this possibility.

2.2.1 The Associative Response

In an attempt to reconcile Mann and Ferguson's findings with the APE model, Brannon and Gawronski (2019: 281-282) assert that

> the current findings are consistent with the idea that a single piece of new information can lead to rapid changes in implicit evaluations to the extent that this information elicits propositional inferences that affirm a new evaluation rather than merely negate an old evaluation.

On this view, when subjects learn that Francis West is attempting to save children trapped inside a house, they take this new piece of information and combine it with the previously provided negative information, the combination of which serves as the basis for the inference that Francis West is good. This inference results in the construction (the authors call it affirmation) of a new associative structure, say, FRANCIS-GOOD and it is this structure that drives AMP performance.[27] So, in effect, Brannon and Gawronski concede that (a) a single piece of information is sufficient to produce a dramatic shift in indirect measure performance and (b) that this shift is mediated by a structure that is a causally downstream from a propositional inference. What they deny, then, is that the newly acquired information does anything to, strictly speaking, alter or change what was learned previously.

But how plausible is this explanation? Brannon and Gawronski's assertion that the APE model is consistent with Mann and Ferguson's findings is, strictly speaking, accurate. Nevertheless, the concern raised in our discussion of the diagnosticity effect is once more salient. The APE theoretic explanation of the reinterpretation effect assumes that subjects have two distinct Francis West-relevant associative structures, the FRANCIS-BAD association that was developed over the course of the learning phase at t_1 and the FRANCIS-GOOD associative structure formed through propositional inference after the presentation of the single piece of information at t_2. The critical question is which of the two structures should we expect *a priori* to mediate AMP responses?

It is here that proponents of the APE model find themselves faced with a dilemma. Either the FRANCIS-BAD or the FRANCIS-GOOD association mediates response. The former possibility, given the constraints of the APE model, is an obvious non-starter (which explains why Brannon and Gawronski don't suggest it). Now suppose, for the sake of argument that the latter associative structure mediates AMP responding. The difficulty here is that classical associative theories predict that, *ceteris peribis*, the stronger association will mediate AMP responding *and* there is every reason to

[27] Recall that a core assumption of the APE model is that all representations are stored as associations in semantic-associative memory. Under the right circumstances, when certain associations are active (e.g. FRANCIS-GOOD) they either temporarily give rise to, are temporarily transformed into, or are propositional structures to be transformed and manipulated by classic propositional processes.

think that the stronger association is FRANCIS-BAD. So, why, in this context, is FRANCIS-GOOD in the driver's seat? Two *prima facie* plausible responses appear to be available.

The most obvious reply is that the FRANCIS-GOOD structure is actually more likely to drive AMP responses, as it was made more accessible by virtue of the fact that it was association that was most recently active. But while there is strong evidence that while accessibility does influence indirect task performance, there is little evidence that the effect of accessibility is so powerful that it is capable of producing a complete *reversal* of indirect task performance (see Blair 2001 for review). In one highly influential study, Blair and colleagues (Blair, Ma & Lenton 2001) asked subjects to vividly imagine a counter-stereotypical exemplar for over five minutes before having them complete an indirect measure and discovered a reduction in implicit bias but not an elimination of it. Consequently, it is at best unclear whether the relative accessibility of the FRANCIS-GOOD association can do the sort of theoretical heavy lifting that is required of it.

The less obvious possibility appeals, once more, to the process of contextualization. On this picture, each associative structure is indexed to features of a particular context and so become active only when those features are present. On this assumption, the two associations do not enter into competition. The problem here, however, is that this explanation is undermined by Brannon and Gawronski's (2019) own findings. In a conceptual replication of Mann and Ferguson's studies, Brannon and Gawronski explicitly designed their studies to test the hypothesis that the reinterpretation effect is a product of contextualization and found, instead, that subjects implicit attitudes of Francis West were stable across contexts. The associative perspective, then, appears ill-equipped to account for the reinterpretation effect.

2.2.2 The Propositional Response

The propositional perspective readily accounts for the reinterpretation effect but struggles to account for Gregg and colleagues' findings, a difficulty that even those friendly to the propositional perspective often acknowledge. The propositional perspective has the same set of options available here as it does with the rate of change anomaly and runs up against analogous problems. To avoid needless repetition, I'll keep this discussion brief.

The propositional theorist might deny that, in this specific context, the chosen indirect tasks tell us anything of relevance about subjects' attitudes toward the two hypothetical social groups, or they might deny that subjects' reported responses track subjects' actual attitudes toward the two social groups. The first option is unsatisfactory because the propositional camp has no principled explanation for why certain indirect tasks, in specific contexts, tell us nothing about the content of subjects' implicit attitudes while other indirect measures and in specific contexts are informative.

In pursuing the second option, there are two paths available. The propositional theorist might deny that the mix-up manipulation was believable. On this hypothesis, indirect task performance reflects subjects' true attitudes toward the two social groups and subjects' reported attitudes reflect subjects' desire to tell experimenters what they want to hear. This hypothesis, however, is unsatisfactory on the grounds that a number of follow-up studies report replicate Gregg and colleagues' findings under conditions in which the cover story is more believable (Cone, Flaharty, Ferguson 2019). Alternatively, the propositional theorist might maintain that retrieval processes activate only a proper subset of the elements constitutive of propositional structures (*e.g.*, NIFFITES and BAD), the activation of which will produce behaviors that mimic associative systems. This last approach is also unsatisfactory. While this account better coheres with Greggs and colleagues' findings compared to Rydell and McConnell's findings, without a principled explanation of when we ought to expect partial activation to occur and an account of the mechanism that gives rise to this phenomenon, such a move amounts to little more than an ad hoc attempt at theory preservation.

It is worth mentioning, however, that Mandelbaum's (2015a) propositional account is better positioned to explain these findings than is the standard propositional account owing to the fact that Mandelbaum assumes that the mind is fragmented such that not all of an agent's beliefs are part of the same inferentially integrated network. On this view, an agent may hold contradictory beliefs at the same time when each belief is stored in a different network. In order to explain why subjects fail to reverse their attitudes toward the Niffites and the Laapians after post mix-up revelation, Mandelbaum needs to only assume that the newly acquired beliefs about the Niffites and Laapians are not stored in the same network as the previously acquired beliefs *and* it was the previously acquired beliefs that drove subjects' indirect task performance.

While the assumption that the mind is fragmented is a reasonable one, Mandelbaum nowhere attempts to articulate a set of principles that govern the storage of beliefs. Why is it that the information provided during the initial conditioning phase is stored in belief network, N_1, while the information given during the counter-conditioning phase is stored in N_2? Similarly, what are the principles that govern the retrieval of information from different networks during task performance? Why should beliefs stored in N_1 come to exert larger influence than beliefs formed in N_2? Under what conditions should we expect beliefs maintained in N_1 to be retrieved for task performance rather than beliefs in N_2?[28] What explains why this information compartmentalization occurs across the two learning phases in Greggs and colleagues' study but not in the analogous conditions of Mann and Ferguson's studies? The primary issue here is not that Mandelbaum offers no determinate

[28] To be sure, these questions will arise for any account on which the mind is fragmented including memory systems pluralism. But a major difference between the memory systems pluralist account and Mandelbaum's belief account of implicit attitudes is that the empirical investigation of these questions are actually part and parcel of memory systems research. Insofar as Mandelbaum is able to answer these questions, it will likely be because he relies on memory systems theoretical assumptions and research.

answers to these questions; it is that without an appealing to an ongoing generative research program like memory systems research, Mandelbaum's account is simply too conceptually impoverished to provide such answers. As is, any resolution of the inferential updating controversy that Mandelbaum may offer will either be *ad hoc* or end up aping the resolution offered by the memory systems pluralist.

2.3 The Co-occurrence/Relational Controversy

The associative theorist assumes that implicit evaluations are associations that track statistical co-occurrence that obtain between associative relata. Because associations do not encode the specific semantic relationship that obtains between the relata, associative theories generate different sets of predictions relative to the propositional view for those cases in which the semantic relationship that x bears to y differs in its evaluative implications when compared to x's merely frequently co-occurring with y. Here's an illustration. *Fisherman's friend* lozenges only tend to be handy when one suffers from a nasty sore throat. On the assumption that implicit attitudes only reflect co-occurrence relations between stimuli, one should not be surprised to find that people might have *negative* implicit attitudes toward Fisherman's friend lozenges by virtue of the fact that they tend to only be present when one is suffering from a sore throat. In other words, their frequent co-occurrence comes to be reflected by the initial formation and subsequent strengthening of the FISHERMAN'S FRIEND– SORE THROAT associative structure. By contrast, on the assumption that implicit attitudes are propositionally structured representations, one can expect people to have positive implicit evaluations of Fisherman's Friend lozenges by virtue of the fact that people who tend to have them nearby when suffering from a sore throat recognize that Fisherman's Friend lozenges will *soothe* a sore throat. In other words, on this view, performance on the relevant indirect task would be mediated by the FISHERMAN'S FRIEND LOZENGES SOOTHE SORE THROATS propositional structure. Given, then, that the evaluative implications of the two kinds of representational structure differ, such differences should be able to be observed under appropriate experimental conditions. For ease of reference, call the sort of effect that the associative view predicts a *co-occurrence effect* and the sort of effect that the propositional view predicts a *relational effect*.

By now the reader shouldn't be at all surprised to learn that there exists a burgeoning empirical literature on this topic that has uncovered both co-occurrence and relational effects. The earliest studies to pit the associative view's predictions against those of the propositional view were conducted by Moran and Bar Anan (2013) using a novel experimental paradigm. Subjects in their studies were told that their task would be to learn about four distinct alien species (A_1, A_2, A_3, A_4). (Novel stimuli were chosen so as to ensure that subjects' responses would not be influenced by learning that had occurred outside the experimental context). Subjects were told that during each trial of the learning phase, the first alien presented would always be the *cause* of either a pleasant-sounding melody (+) or a human scream (–) and the second alien shown would *end* the melody or scream. The associative view predicts that subjects will have an implicit preference for either of the two alien species that are consistently paired with positive melodies, regardless of whether the particular alien species starts the melody or ends it, over either of the two alien species that are consistently paired with human screams (again, regardless of the relationship that each species bears to the scream). On the other hand, if implicit attitudes are propositions, then we would expect subjects to implicitly prefer aliens that stop human screams over aliens that stop pleasant melodies. The results of each study lend unambiguous support for the associative perspective. These initial findings have since been replicated (Moran, Bar-Anan, Nosek 2017, Moran & Bar-Anan 2020).

Relational effects have also been observed and, on at least one occasion, in the same set of studies that also report co-occurrence effects. In one influential study, Hu, Gawronski and Balas

(2017) developed a novel experimental paradigm that, like the Moran and Bar Anan studies, pitted associative predictions against propositional predictions. In this study, subjects were tasked with learning about various hypothetical pharmaceuticals. Subjects were told that each pharmaceutical either *caused* a particular positive or negative health-related outcome or *prevented* a positive or negative health related outcome. In each learning trial, subjects were shown a pharmaceutical and were explicitly told whether the pharmaceutical causes or prevents the ailment that was subsequently presented. The associative view predicts that subjects would have negative implicit attitudes toward even those pharmaceuticals that prevent negative health outcomes and positive implicit attitudes toward even those pharmaceuticals that prevent positive outcomes. The propositional view, by contrast, predicts that subjects would have positive implicit attitudes toward pharmaceuticals that prevented negative outcomes and negative attitudes toward pharmaceuticals that prevented positive outcomes. The researchers reported that their findings were as the propositional view predicts. It is worth pointing out that while this experiment (Experiment 3) produced a relation effect, each of the other two studies reported in the same article found co-occurrence effects albeit using a slightly modified experimental procedure (Experiments 1 & 2). Lastly, it must also be stressed that several more recent studies have uncovered relational effects using a variety of experimental procedures, stimuli, and measurement tasks (see Corneille & Stahl 2019 for review).

As I dedicate an entire chapter to the relational/co-occurrence controversy (Chapter 6), I spare the reader a detailed description of the failures of the associative and propositional perspectives to resolve this controversy here. Rather, I'll merely highlight what I take to be the main difficulties with each approach below.

2.3.1 The Associative Response

Gawronski acknowledges that relational effects pose a major theoretical challenge to associative theories (see Gawronski & Bodenhausen 2018: 14, Hu et al. 2017). Helpfully, Hu and colleagues describe how the dual-process theorist might attempt to accommodate the existence of relation effects (*ibid*: 29-30). I summarize this potential explanation below.

When subjects are presented with the information that Drug X prevents an eye infection, subjects in infer that X is good. This inference produces a novel DRUG X- GOOD associative structure in semantic memory. However, because Drug X is repeatedly paired with images of an eye infection, subjects also automatically form the association DRUG X- EYE INFECTION and, consequently, a DRUG X – BAD associative structure. If we assume that each subject receives the information contained in this learning trial several times over the course of the learning phase, subjects rehearse the aforementioned propositional inference on each occasion at which they encounter this information and that associative structures can be formed by way of inferential processes, then subjects may form two, equally strong yet conflicting representations of Drug X. If so, then we should expect to find that these structures cancel each other out resulting in the actually observed null-effect.

There are several concerns with this putative explanation. The first is theoretical and is explicitly acknowledged by Hu and colleagues:

> However, in the absence of empirical evidence for these additional assumptions, a single-process propositional interpretation seems superior, because (a) it requires only one auxiliary assumption to explain the current set of findings and (b) this auxiliary assumption led to a novel prediction that was empirically confirmed in Experiment 3 (see Gawronski & Bodenhausen 2015).

In other words, in order to account for Hu and colleagues' relation effect, the dual process theorist must make several theoretically unmotivated modifications to their preferred model. The second is that there is, as of yet, no empirical support for the hypothesis that subjects engage in propositional inferences on each learning trial. Although we cannot conclude from this that subjects do not engage in such propositional inferences, the fact that the propositional inference hypothesis was proposed after the fact to account for antecedently unexpected results should make us wary of accepting this explanation pending further empirical investigation.

2.3.2 The Propositional Response

Despite the fact that some dual-process theorists concede that the propositional perspective is better positioned to resolve the relational/co-occurrence controversy, I think that they have reached this verdict prematurely. While a full discussion of the failures of the propositional perspective to resolve this controversy must wait until Chapter 6, we do not need to engage in a lengthy discussion to show that the propositional perspective is not as well-suited to resolve this controversy as it may seem.

First, the propositional perspective does have a hard time accounting for the *co-occurrence effect* observed by Moran and Bar-Anan. The propositional perspective simply has difficulty explaining the relational effect observed on self-reported measures but not on indirect task measures. To be sure, De Houwer could once again attempt to explain indirect performance in terms of retrieval failure (see §4.2.1.2) but such an appeal in this context would carry with it all the same theoretical baggage as it did in the context of explaining the incremental effect. Indeed, in this particular context, the claim that the constituents of propositional structures could be retrieved associatively does threaten to undermine the falsifiability of the propositional perspective in this particular domain.

Second, while Bading and colleagues' (2019) similarity construction account has explicitly been used to show one could explain Moran and Bar-Anan's findings without appealing to the retrieval of associative structures, the similarity construction account is silent on co-occurrence effects that emerge on indirect measures other than the IAT. Since Hu and colleagues uncovered a co-occurrence effect on Experiments 1 and 2 using the AMP, the similarity construction account fails to provide a unified explanation of the target phenomena.

2.4 Summary

What these three controversies seem to establish is that neither the associative perspective nor the propositional perspective gets things quite right. Implicit attitudes really do behave like associations in some contexts and propositionally structured representations in others, and neither perspective can plausibly explain these patterns. As we are still interested in exploring how much mileage one can get from the mental state picture of implicit attitudes, this rules out a retreat to the view. So, where should we go from here? Before I argue that memory systems pluralism (MSP) has the best prospects for elucidating the data that are at the core of the format problem, it is worth considering two more orthodox attempts at resolving these issues. First, we'll consider Gawronski and Bodenhausen's argument that the debate between propositional theorists and associative theorists is not actually a debate of substance. We will then move to consider the view that implicit attitudes are *sui generis* before sketching the memory systems pluralist solution.

3 The Mere Semantics Reply

In the previous section, I alluded to the fact that associative theorists have pursued two orthogonal strategies potential in attempting to resolve the format problem. Hitherto, we have only scrutinized one such strategy in which the associationist attempts to resolve the format problem by explaining how associationism can accommodate the data that most impresses the propositional theorist (the direct strategy). I argued that this strategy is a dead end.

The other strategy involves the rejection of an assumption on which this whole dispute between propositional theorists and associative theorists is based (I labeled this the indirect strategy). In particular, the associative theorist might deny that the distinction between associations and propositions is one of substance, and, instead, affirm that it is a matter of terminological taste whether one wants to refer to the states that drive evaluative responses to indirect measures as associations or propositions (see Gawronski, Brannon, & Bodenhausen 2017, Gawronski & Bodenhausen 2018). The underlying idea is this: if there is no substantive distinction between associations and propositions, then it appears as though there can be no substantive dispute between whether associations or propositions drive indirect measure performance. Thus, the format problem is not one that needs to be resolved; it needs to be dissolved. I call this the *mere semantics response*.

How does the APE theorist (or dual-process theorist) defend the claim that, on their preferred model, there is no substantive distinction between associations and propositions? The response requires some set up and typically begins with a statement of the orthodox dual-process position attitudes are not stored as two distinct types of structure stored in memory. Given the work the theoretical import of this hypothesis for the purposes at hand, the passage in which this is most clearly articulated by Gawronski and colleagues is worth quoting at length:

> Although some theorists assume that people can have two distinct attitudes toward the same object stored in memory (*e.g.*, Greenwald & Banaji 1995, Wilson, Lindsey & Schooler 2000),

> we argue that a duality account based on the distinction between associations and propositions as two independent memory structures is theoretically implausible. Such an account would imply that propositional statements about states of affairs are stored in a manner that does not involve any kind of associative links. Counter to this idea, most theories that are based on associative-propositional duality do *not* assume two distinct memory stores for associations and propositions (*e.g.*, Strack & Deutsch 2004). Instead, these theories propose a single associative store that provides the basis for propositions about states of affairs in the form of patterns of momentarily activated associations. According to this view, the distinction between associations and propositions does not describe two distinct *types* of stored knowledge structures in long-term memory, but different states of stored knowledge. Associations can be understood as dormant links between nodes that constrain the spread of activation within associative networks. Activated patterns of associations, in turn, are assumed to provide the basis for momentarily constructed propositions about states of affairs. From this perspective, any proposition is based on patterns of activated associations; there is no association-independent storage of propositional statements in a different part of long-term memory. (Gawronski, Brannon & Bodenhausen 2017: 105)

.

In this passage, Gawronski and colleagues are particularly interested in explaining how it is that one can coherently posit the existence of both associations and propositions while also maintaining that all such representations are stored in a single semantic-associative memory store. Their first step is to claim that

> (1) most dual-process models (the APE model included) assume that propositional representations and associations are not different *types* of stored knowledge.[29]

The obvious question to raise here is how it is possible to deny that associations and propositions are different types of representational structure while maintaining that they have radically different properties. It is at this point that the authors go on to characterize what they mean by *association*:

> (2) Associations can be understood as dormant links between nodes that constrain the spread of activation within associative networks.

The reader might be surprised to have read this given this description of an association is unlike any of the descriptions that we have so far encountered. And, indeed, this different use of 'association' is made apparent when one considers how they characterize propositions:

> (3) Activated patterns of associations, in turn, are assumed to provide the basis for momentarily constructed propositions about states of affairs.

This characterization is needlessly opaque (*e.g.,* what precisely do they mean when they say that associations provide the *basis for* momentarily activated propositions?). What they should have said, given their remarks in the passage under consideration and elsewhere (see Gawronski & Bodenhausen 2011), is something like this:

[29] The full sentence, of course, goes on to claim that propositions and associations are different *states* of knowledge but the type/state distinction in this context is needlessly (if not hopelessly) obscure, and, in any case, that additional bit is irrelevant for our purposes. What matters is simply that the APE model does not assume that propositional representations and associations fundamentally different kinds of representation and/or are stored independently.

(3*) Propositions are those momentarily activated patterns of associations that represent states of affairs.

Compared to 3, 3* has the virtue of stating unambiguously the relationship between associations *qua* dormant links and nodes that contain activation spread and representations that reflect states-of-affairs. But if propositions just *are* those patterns of active associations that represent states of affairs, then how ought we characterize those associations that the APE theorist routinely posit to explain, say, IAT performance? A core commitment of the APE model, recall, is that *mere associations* causally mediate performance on indirect measures of attitude. And nowhere in the passage that we've been unpacking do we find a characterization of mere association. While imprecise, we can capture the notion of the mere association given the framework articulated in the following:

(4) A mere association is a momentarily activated pattern of associations that are formed, updated, and activated exclusively according to paradigmatic associative principles.

One virtue of this proposal is that it helps to sharpen the claim that associations and propositions are not different types of representation. To say that they belong to the same type or do not need to be stored separately is just to say that both propositions and mere associations are patterns of activation in an associative network with different sets of properties.

So, how do we get from here to the claim that the distinction between associations and propositions, from the perspective of the APE model, is not substantive? In their response to the objection that 3* is false on the grounds that no actually existing associative network is capable of accurately representing relations between objects and events the same way that humans do, Gawronski and colleagues say the following:

> We argue that this rejection is based on a very narrow interpretation of associative representations that reduces them to primitive links between two concept nodes. After all, multi-layer connectionist models involving both excitatory and inhibitory links are perfectly able to represent complex relations between objects and events (*e.g.,* McClelland et al. 1995). Such models often include a hierarchical structure, in that activated concepts at higher levels specify the relation between activated concepts at lower levels. Mental representations of this kind could be described as *propositional,* because they capture relational information. Alternatively, they could be described as *associative,* because they are based on associative links between nodes. *From this perspective, the preferred label becomes a matter of terminological taste rather than genuine theoretical disagreement.* (emphasis added) (Gawronski, Brannon & Bodenhausen 2017: 106)

If one grants that a connectionist architecture can represent complex relations between objects and events and that that connectionist networks are a special case of associationist network, then it is a matter of mere convention (they call it terminological taste) should one chooses to describe the patterns of activation that represent states of affairs as propositions or associations. And so, the mere semantic reply goes, if there is no substantive distinction to be drawn between associations and propositions on the APE model, then the format problem simply does not arise for the APE model since the format problem presupposes a difference between implicit attitudes *qua* associations and implicit attitudes *qua* propositions.

With the view laid out in full, it doesn't take much work to show that the mere semantics reply fails to dissolve the format problem even if one assumes, with the APE theorist, that propositions just are patterns of activations in an associationist network. Simply put, the format problem arises once more if instead of focusing on the representational structure we focus on classes of activation patterns in the connectionist network. A core assumption of the APE model is that mere associations drive performance across a wide range of indirect measures. But, *contra* the APE model, the evidence canvassed in the previous section provides strong support for the view that propositions drive performance on indirect measures. This is the case even if one supposes with the APE theorist that all propositions are momentarily activated patterns of associations in a connectionist network. In other words, the format problem can now be reframed as a problem of which of the two classes of patterns of activated associations drives indirect measure performance. The APE model assumes that the patterns of activation that capture mere association explains indirect measure performance, whereas its detractors can simply claim that the patterns of activation that represent propositions best explain indirect performance. Now, as I have argued at length, both camps are wrong in their assertions. But the problem arises for the APE theorist all the same, albeit under a different guise. And if this is the case *and* the APE model fails to accommodate the data that impresses the propositional theorist, then the APE model seems incapable to resolving the format problem under any of its guises.

Before moving on, there are a pair of important lesson to be drawn from this discussion. I have hitherto framed the problem that we have been addressing as one about the internal structure of those representations that drive indirect task performance: associative theorists claim that all implicit attitudes are associative structures and propositional theorists claim that all implicit attitudes are propositional structures but since the best available evidence undermines both views, then neither perspective gets things right. What this discussion shows is that one cannot dodge these empirical bullets simply by adopting a model on which mere associations and propositions are differently realized in the same underlying architecture. This lesson generalizes: no matter how one thinks that we actually represent states-of-affairs (whether one be a language of thought theorist, a monist, or a connectionist), if one wants to give an account of implicit attitudes, then one has to have a story to tell about why implicit attitudes sometimes behave as though they are mere associations and sometimes as though they express and are sensitive to states of affairs, that is behave as though they are propositional structures or beliefs. So, the format problem should not be regarded as fundamentally one over the internal structure of stored representations despite this being the most common framing.

The second lesson is that even if one can develop a connectionist model that can (a) fully represent states of affairs while also (b) encoding mere associations, there is no guarantee that such a connectionist model will end up being equivalent, in relevant respects to the APE model. Moreover, as the most viable connectionist architectures *do* assume the existence of multiple memory systems (see O'reilly & Munakata 2000), we are left with little reason to think that an APE model framed as a connectionist model will fare any better than the standard associationist APE model with respect to resolving the format problem.

4 The *Sui Generis* Perspective

We have hitherto assumed that implicit attitudes are either mere associations or propositionally structured representations, which are the two dominant theoretical perspectives of working social psychologists. However, several philosophers have rejected the associative/propositional dichotomy in favor of a view on which implicit attitudes are *sui generis*

(SG). While there are several such views (see Brownstein 2018, Gendler 2008, Levy 2014), this section will focus exclusively on Brownstein's SG account of implicit attitudes.[30]

Before turning to Brownstein's account, the claim that implicit attitudes are *sui generis* requires some explication. What does the claim that implicit attitudes are *sui generis* amount to? On one proposal, some class of mental states, *m*, are *sui generis* if *m* is not recognized by common-sense, or folk, psychology (see Levy 2014). While there may be some theoretical projects for which this standard is legitimate, the present project is not one of them. This standard is much too liberal, as too many theories of implicit attitude would likely count as SG-theories on this proposal (*e.g.,* Mandelbaum's view on which implicit attitudes are a species of belief that lack many of the features deemed essential on the ordinary conception of belief). It mystifies more than it elucidates.

Instead, I propose the following: Implicit attitudes are *sui generis* if implicit attitudes are a kind of mental state not currently recognized, or posited, by any of the other empirically well-confirmed or thriving disciplines in cognitive science. The standard, then, for the introduction of a novel mental kind are demanding but not unfairly so. For instance, this standard does not require that some putative novel mental kind be countenanced amongst the kinds of mental state that are to be featured in the "final theory" of mind. This proposal also has the virtue of being consistent with the way that Brownstein understands his proposal that implicit attitudes are an SG-kind:

> While I'll argue that implicit attitudes are sui generis, I don't mean to suggest that I've discovered anything radically new about the mind or am positing a category that

[30] There are a few reasons for this. First, Levy maintains that implicit attitudes are propositional structures that are not beliefs. He calls them *patchy endorsements*. Patchy endorsements are belief-like insofar as implicit attitudes are propositional structures but lack the functional properties characteristic of beliefs (*e.g.,* inferential promiscuity). But on my characterization of the propositional perspective, Levy's account is best thought of as a type of propositional view. As for Gendler's *alief* theory of implicit attitudes, all of the problems that I raise below for Brownstein's FTBA account of implicit attitudes are problems for Gendler's alief account, thus there is no reason to belabor the discussion by addressing them both.

> must be included in a "final theory" of the mind. The history of Western philosophy is suffused with discussion of the relationship between, on the one hand, animal spirits, appetite, passion, association, habit, and spontaneity and, on the other hand, reason deliberation, reflection, belief, and judgement. But they fit naturally within this long history of dichotomous ways of thinking of the mind and represent, inter alia, a proposal for understanding key commonalities among the phenomena variously called passion, habit, and so on. (Brownstein 2018: 65)

Lastly, the proposal on offer implies that conferring SG status to implicit attitudes is warranted if there is some set of characteristics or properties that implicit attitudes have in common such that are not had by any other of the more ordinary mental kinds that cognitive science already traffics in. This additional feature of my proposal ensures that the introduction of a genuine SG-kind will not be a simple relabeling of some entity or collection of disparate entities that are already posited by our most well-confirmed cognitive scientific theories. With the preliminaries out of the way, let's now examine Brownstein's account.

4.1 Brownstein's FTBA Theory of Implicit Attitudes

Brownstein conceives of implicit attitudes as a sort of *spontaneous inclination*. According to Brownstein spontaneous inclinations, as opposed to the kinds of states that issue from deliberative reflection, have four key components.

> These are (1) noticing a salient *F*eature in the ambient environment; (2) feeling an immediate, directed, and affective *T*ension; (3) reacting *B*ehaviorally; and (4) moving

> toward *A*lleviation of that tension in such a way that one's spontaneous reactions can improve over time. Noticing a salient feature (F), in other words, sets a relatively automatic process in motion, involving co-activating particular feelings (T) and behaviors (B) that either will or will not diminish over time (A), depending on the success of the action. This temporal interplay of *FTBA* relata is dynamic, in the sense that every component of the system affects, and is affected by, every other component. Crucially, this dynamic process involves adjustments on the basis of feedback, such that agents' FTBA reactions can improve over time. (Brownstein 2018: 31)

To illustrate, imagine that Smith is walking down a sidewalk late one night in an urban setting and encounter a person of color approaching. Smith might (a) notice the race of this individual (F), which activates feelings of fear (T) and an avoidance response causing Smith to cross the street in order to avoid crossing paths with this individual (B). The activated behavioral response will in turn either succeed in or fail in alleviating Smith's feelings of fear (A). Importantly, if the activated behavioral routine succeeds in alleviating Smith's fear response, Smith is more likely to engage this behavioral routine under similar conditions. If, on the other hand, the behavior does not alleviate Smith's fear response, then Smith may or may not have this spontaneous inclination to cross the street in the future. Because spontaneous inclinations can update in response to newly acquired information in this way, Brownstein conceives of spontaneous inclinations as *intelligent*.

Brownstein conceives of implicit attitudes as associations with FTBA relata, which is not to say that implicit attitudes are *mere associations* between concepts stored in long term memory (*a la* Gawronski & Bodenhausen 2006, 2009). States with FTBA relata are associative in the sense that tokening of a feature representation co-activates affective responses, and motor routines that "aim" at alleviating the affective responses analogous, but not identical, to the way that thoughts of salt co-activate thoughts of pepper. That is, "when an FTBA is set off, a 'spread' of automatic activation always occurs" (Brownstein 2018: 92) in a manner analogous to the spread of activation that occurs when tokening one's concept SALT results in the co-activation of one's PEPPER concept. So while these states are associative, they are not *mere associations* as they are assumed to update in ways that mere associations do not (*e.g.,* by being sensitive to the contents of other representations and by being prone to change in ways that suggest inferential updating *a la* Mann & Ferguson (2015)).

With that said, implicit attitudes cannot properly be treated as either beliefs or propositional representations either. After all, while the *A* component of an FTBA state allows an FTBA state to update in non-associative ways in response to changes in the environment or one's orientation, the fact that the "spread" of activation is automatic "helps to explain why these states do not take part in the inferential processes definitive of states like beliefs" (Brownstein 2018: 94). In other words, Brownstein assumes that beliefs are defined by their functional properties, properties which states with FTBA relata lack due to the automatic nature of the associative spread that binds the four components.

One might wonder what reason Brownstein provides for his claim, which he insists on in various places, that states with FTBA-relata are a *unified*, *mental-kind* (Brownstein 2018, Madva & Brownstein 2017). Why explain the data by appealing to FTBA-shaped lumps (*cf.* Nagel's (2012) "lumpiness" objection to Gendler's alief theory)? Brownstein's response to this question is obscure. The thought appears to be that because the tokening of an implicit attitude *always* (allegedly) involves a representation of a feature (F), a an affective experience (T), an activation of either a behavior or a motor routine (B) and either a diminishing or not of that affective experience (A), this fact alone justifies our treating implicit attitudes as a type of mental state with these components. To be sure, while one might be able to swap out, say, one type of behavioral response for another (*e.g.,* when attempting to quit smoking, one might reach for a pen to place between one's lips rather than a cigarette), Brownstein alleges that an implicit attitude *qua* spontaneous inclination *always* involves the co-activation of *some* feature, with *some* felt tension, with *some* behavioral response, and *with* some feeling of alleviation or not (Brownstein 2018: 94).

Brownstein conceives of states with FTBA components as being the "conceptual cousins" of Gendler's *aliefs*. According to Gendler (2008), an alief is a *sui generis* mental state that is *a*ssociative, *a*utomatic in its activation, *a*ffective, and *a*rational. This explains why, in her classic example, an individual afraid of heights that steps foot on a sturdy, glass structure that hangs over the Grand Canyon may automatically feel a sense of fear, begin to tremble, and want to retreat even though she knows full well that the structure is safe. The crucial difference between an *alief* and a state with FTBA components, according to Brownstein, is that without the alleviation component, Gendler cannot explain why implicit attitudes is sometimes sensitive to reasons in ways that seem intelligent, which allows one to hone one's implicit attitudes so that they can over time become more harmonious with one's explicit beliefs and goals (Brownstein 2018: 91). In other words, one wouldn't be too far off the mark were one to treat Brownstein's FTBA account of implicit attitudes as a modification of Gendler's alief view (which, explains why we are only concerning ourselves with Brownstein's view in this section; see fn30).

Finally, *if* there are mental states with FTBA components, they would satisfy the desiderata articulated above for treating a state as *sui generis*. After all, on Brownstein's view, implicit attitudes have properties characteristic of both associations and beliefs while being reducible to neither. Moreover, it really is the case that our best theories of cognitive science do not recognize (explicitly, anyway) these, as Madva and Brownstein call them, "semantic-affective-behavioral clusters" (2018: 19).

4.2 Rejecting Brownstein's Account

Brownstein's FTBA account of implicit attitudes ought to be rejected on several grounds. First, it doesn't solve the format problem. Second, it falls victim to Carruthers' ontological opacity objection. Third, Brownstein makes no attempt at describing how these various disparate components come to be inextricably linked in a type-identifying manner. Fourth, the best available data suggests that FTBA relata are dissociable and that there are relatively well understood mechanisms that explain why this is so. And finally, Brownstein's account is both wildly unconservative and unparsimonious in light of these other considerations. I'll develop each objection in turn.

The-fails-to-solve-the-format-problem objection. Despite the fact that Brownstein motivates his FTBA account of implicit attitudes by pointing to the failures of extant perspectives to solve the format problem, Brownstein makes no attempt at using his account to resolve the format problem. Perhaps this is the case because Brownstein assumes that such failures serve to rule out all of the relevant alternatives to his account. If so, then this tacit assumption is false on the grounds that he fails to take seriously the memory systems pluralist account of implicit attitudes.[31] In any case, should we be interested in fashioning a response to the format problem from the FTBA account on Brownstein's behalf, it is far from obvious as to how one ought to accomplish this. What, for instance, about his account allows us to predict, much less explain in a post hoc fashion, why it is that co-occurrence effects are observed in Moran and Bar-Anan's (2013) studies but a relation effect is observed in one of Hu and colleagues' pharmaceutical studies? Simply asserting that FTBA

[31] Though, presumably, Madva and Brownstein (2018) would take issue with this claim as they devote an entire paper to criticizing so-called *two-type theories* of implicit attitude, which they claim are entailed by Amodio and colleagues (Amodio & Ratner 2011) multiple memory systems account of attitudes. Whether or not Madva and Brownstein fairly characterize Amodio's position, the two-type picture that they criticize is fundamentally different from the version of pluralism that I defend.

components have some of the properties of associations and some of the properties of beliefs does not make for an illuminating explanation. The problem, in other words, is two-fold: First, after Brownstein draws various lessons from the format problem in motivating his FTBA account of implicit attitudes, he never attempts to circle back on the data that motivates his account to show how his theory resolves the various issues; second, even if one were to attempt to do so, the account is too underdeveloped to account for the data. Presumably, the latter explains the former.

When outlining the associative perspective, recall that I considered and rejected a view according to which implicit attitudes are associations between lexical concepts and affective states on the grounds that the view is *ontologically opaque*. Here's how Carruthers (2017) expresses the objection:

> Madva and Brownstein (2017) argue against a dissociation between cognitive and affective attitudes, however. They claim that all implicit attitudes are clusters of semantic-affective associations, thus defending a position quite close to that of the APE model of Gawronski and Bodenhausen (2006). […] [T]he ontological status of these postulated clusters is opaque. How does a semantic representation associate with an affective state, or vice versa? One semantic-representation can associate with another, since both are stored representations that can become linked in such a way that activating the one will activate the other. But how does a concept associate with a feeling? Affective states are the outputs of evaluative processes, not stored representations of any kind. (Carruthers 2017: 69)

To be sure, Madva and Brownstein might attempt to get around this objection by denying that there are any cold cognitive states— all states are affectively laden. Further, if all affective states also have representational contents, then perhaps this might be sufficient to strip this objection of its force. I am skeptical that endorsing this highly controversial position will be enough to stave off the ontological opacity objection given that the neural networks that produce paradigmatically cold representational states are largely distinct and dissociable from those that produce affective states. But even if this were so, Carruthers' passage hints at a response in the form of a second objection.

There is actually a second objection this implicit in this passage and the objection is this: even *if* a sensible response to the ontological opacity objection can be provided, by means of what mechanism does a semantic representation come to be associatively linked with an affective state? Brownstein makes no attempt to describe the causal mechanism by which these disparate states come to be associatively linked. Brownstein merely points out that these states regularly cluster together and infers from this that these disparate states are so tightly bound together that they constitute a *sui generis* mental kind._Call this the *linking-mechanism objection*.

Fourth, Brownstein has not sufficiently motivated his claim that the explananda of implicit attitude research is best explained by positing the existence of FTBA-shaped lumps. The concern is that even if one were to grant that the tokening of any implicit attitude *always* involves the type-identifying spread of activation described above, this need not be explained by positing the existence of a kind of mental state over and above those states and processes characterized by each individual component. Why treat each of these as components of a single unified mental state as opposed to treating each component as a dissociable state arising from distinct cognitive/affective mechanisms? This latter position, after all, is how memory systems pluralist account of implicit attitudes might treat each of these components (see Chapter 6). This worry is made more pronounced by the fact that Brownstein concedes that these different components of implicit attitudes may be grounded in the operations of functionally distinct brain networks, each of which can be conceived of as "self-standing systems" (Madva & Brownstein 2017).

The *parsimony objection*, then, is this: our best cognitive scientific theories posit the existence of various dissociable neural systems, some of which are directly relevant to explaining (i) how one comes to assign significance to features in one's ambient environment, (ii) comes to affectively respond to those features, (iii) why one behaves in such and such ways in response to those features, and (iv) and why sometimes behaviors diminish, or fail to diminish, one's affective responses after behaving in such and such way. The successes of these theories, moreover, do not in any way depend on our having to posit the existence of a *sui generis,* unified FTBA-state. So, the introduction of such a state serves to increase the number of theoretical posits without also increasing our explanatory power.

On the parsimony objection, Madva and Brownstein *do* have something of a response available to them. It's worth producing the passage in which they consider the possibility that self-standing neural systems may give rise to each FTBA components in its entirety:

> Our one-type model opposes quasi-phrenological efforts to isolate the specific brain regions that subserve specific types of mental state. Of course, we do not deny that there are meaningful differentiations between brain networks, nor that these networks can in some respects be thought of as self-standing systems. Suppose that, during the shooter bias task, perceiving a black man activates semantic associations with criminality and guns, affective responses of danger, perceptual and attentional biases to detect signs of threat, and motor preparations for fight-or-flight (Correll *et al.* 2015). This co-activating effect can occur "within" as well as "between" different networks. In other words, *at different levels of explanation, the brain can rightly be described as composed of several subsystems, as a unified system unto itself, and as one component of a larger bodily- environmental system. A single semantic network model at a higher-psychological level is*

> *consistent with a multi-system model at a lower-neural level.* (emphasis added) (Madva & Brownstein 2018: 631-632)

The suggestion here seems to be that MSP theory and Brownstein's FTBA account are operating at different levels of explanation. Madva and Brownstein seem inclined to treat a view like memory systems pluralism as offering a theory about how psychological states with FTBA relata are realized at a neurological level. Call this the *levels reply*.

The levels reply is unconvincing. Any view that adopts the memory systems pluralist perspective is committed to providing mechanistic explanations of implicit attitude relevant phenomena and mechanistic explanations span levels. Thus, it would be at best misleading and at worst false to describe the explanations provided by Amodio's multiple memory systems account of social cognition (MMS-SC) as being at the neurological level. Moreover, memory systems pluralism could be regarded as a model of how FTBA states are implemented at the neural level only if memory systems pluralism and the FTBA account do not differ in their predictions with respect to how implicit attitudes behave. But not only do the two accounts make different predictions, memory systems pluralism's predictions are better supported by the available data in the way described in the objection below.

Finally, one can, and should, at least challenge Brownstein's contention that implicit attitudes *always* involve co-activating FTBA relata. There is ample empirical evidence, much of which will be canvassed over the next two chapters, that these putatively co-activating components *can* and, more importantly, *do* come apart. In other words, the claims that Madva and Brownstein articulate in the passage below are false.

> A fundamental feature of implicit mental states is co- activation[…]*Which* particular associations are activated in any given context varies in ways we begin to explain below, but *that* a "spread" of activation occurs does not vary.

With that said, we'll postpone further discussion of the matter until we have more fully articulated the memory systems pluralist response to the format problem.

5 The Format Problem and MSP

The vast majority of this chapter has been dedicated to using the format problem to show what implicit attitudes are not. Categorically speaking, they are neither associations, propositionally structured representations, nor some novel sui generis mental kind. One possibility that we have not yet considered fully considered is that of memory systems pluralism. I will now briefly describe how a memory systems pluralist should approach the format problem, before sketching the argument that will be fleshed out and defended over the course of the next two chapters (Chapters 5 & 6).

Memory systems pluralism, recall, denies that implicit attitudes are a unified mental kind. Rather, there are many kinds of implicit attitude and each kind is subserved by a distinct memory system. Of the memory systems that I had characterized in Chapter 3, each was described as operating over a particular kind of associative structure. One might thereby expect me to develop a solution to the format problem that weds the associative perspective with a memory systems pluralist architecture. This is *not* the approach that that I'll be taking. Rather, the tacit assumption, hitherto made by all parties to the present controversy, that all implicit attitudes are of the same representational format. Positively stated, the working hypothesis is that some types of implicit attitudes are to be modeled associatively while some are to be modeled propositionally.[32] Thus, this approach is pluralist not only with respect to the types of implicit attitude we need to posit in order to account for the relevant data but also with respect to representational format. For short, we'll refer to this view as *format pluralism*.

Motivating the position that this brand of memory systems pluralism has the best prospects of the available theories of resolving the format problem in a satisfactory manner requires that I make plausible the claims that (a) our most successful models of memory in cognitive neuroscience are already committed to the existence of some memory systems that traffic in associations and some that traffic in propositions, (b) some propositional memory systems have the properties necessary to influence evaluative responses on a wide range of indirect measures, and (c) appealing to the joint operations of these systems not only helps us to account for the various controversies that constitute the format problem in a non-trivial and illuminating manner but also generates a number of novel predictions that could successfully guide future empirical research. The first claim needs must be defended to assuage concerns about this being an *ad hoc* solution to the format problem. Defending the second of the three claims is needed to push back against dual-process theoretical concerns that propositional states and processes lack the functional properties necessary to influence evaluative responses to implicit measures. The defenses of these first two claims are staged in Chapter 5. And third claim requires defense lest memory systems pluralism make itself vulnerable to the same kinds of complaints that I had just leveled against Brownstein— namely, that

[32] It is important to emphasize that one reason that this possibility has not been explored is that social psychologists have by and large been reluctant to posit the existence of more than one memory system (see Chapter 2. even when this restriction is relaxed by dual-systems theorists (see Smith & DeCoster 2007, Rydell & McConnell 2006), such theorists tend to be reluctant to posit more than one memory system at either the implicit level or the explicit level. The very possibility that more than one learning system could be driving evaluative responses to indirect measures has been crowded out conceptually by these prior theoretically commitments. Chapters 2 and 3 gave us independent reason to reject the single memory system hypothesis. And A stronger case will be made against the dual-systems theoretic assumption in Chapter 5.

memory systems pluralism is in principle consistent with the patterns of results without helping to illuminate them in a theoretically satisfying manner. This will be the topic of Chapter 6.

CHAPTER 5

PROPOSITIONAL MEMORY SYSTEMS AND IMPLICIT COGNITION

The type of MSP theoretical resolution to the format problem that I favor posits that propositional structures and processes, subserved by memory systems that traffic in these states and processes, and associative states and processes, subserved by associative memory systems, jointly contribute to indirect task performance. The view that implicit attitudes are heterogeneous with respect to representational format is a view that I call *format pluralism*. Though MSP format pluralism breaks decisively from social psychological orthodoxy in several important respects (some of which are further explored in this chapter), the previous chapter motivated this solution primarily by showing that all other attempts to solve the format problem fail to do so.

However, the theoretical groundwork for a format pluralist solution to the format problem has not yet been laid in its entirety. A major concern is that one of the most influential memory systems models, *viz.* Squire and Zola-Morgan's (1996) *medial temporal lobe* (MTL) model, assumes that all states and processes deserving to be called *implicit* are associative. In other words, the MTL model assumes a type of *format monism* with respect to implicit attitudes. Because the MTL model is consistent with MSP theory (see the relevant discussion in Chapter 3), it follows that establishing the superiority of MSP over its competitors alone does not guarantee a resolution to the format problem *if*, indeed, format pluralism is necessary for its resolution. What is left to be shown, then, is that we have good reason to prefer the format pluralist friendly memory systems models over the format pluralist *un*friendly models. This stands as the primary goal of this chapter.

It turns out that on route to satisfying the primary goal of this chapter, we will have also satisfied a secondary goal of this chapter. One dogma of dual-process typology (DPT) is that propositional states and processes lack the characteristics necessary to *directly* influence performance on the various indirect measures that are commonly used in social and cognitive psychology. It turns out that the same well established empirical results that undermine the MTL model also undermine this dogma. In making this case, which is a secondary goal of this chapter, we'll have therefore defused a pressing potential objection to format pluralism.

The chapter is structured as follows. First, I briefly reintroduce the MTL model, highlighting those assumptions that most directly challenge the format pluralism. I also elaborate on the connection between the primary goal of this chapter with the secondary goal (§5.1). I then introduce two format pluralist friendly memory systems models and argue that each model accounts for a body of empirical research that the MTL model cannot accommodate (§§5.2-5.3). The chapter concludes with a brief description of two important limitations of this discussion with respect to the development of a solution to the format problem (§5.4).

1 The Medial Temporal Lobe Model and Implicit Cognition

Back in Chapter 3, I introduced Squire and Zola-Morgan's (1991) *medial temporal lobe* (MTL) *model* with the goal of illustrating how it can help us make sense of the claim, which is at the core of the MSP explanatory project, that there are many species of implicit attitude. However, one might recall that the various types of implicit attitude documented in that chapter are all explicable in terms of associative states and processes (that is, from the MSP perspective, the MTL model assumes format monism). Recall that the MTL model assumes a taxonomy on which there are, fundamentally, two types of memory system (see **Figure 3**; 68). There is a single explicit memory system, which is composed of two highly interdependent subsystems, *viz.* episodic memory and semantic memory. Generally speaking, and in line with standard DPT assumptions, it is assumed that this explicit (declarative) memory system implements propositional states and processes. The MTL model also assumes the existence of several independent implicit memory systems each of which is capable of operating in parallel to both the other implicit memory systems and the explicit memory system. As all memory systems that collectively constitute implicit memory operate over associative structures in accordance with associative principles, it follows that the MTL model assumes that all implicit cognitive states and processes are associative states and processes.[33]

The MTL model, so understood, poses a challenge to the development of a plausible MSP format pluralist solution to the format problem. While the format pluralist solution, strictly speaking, takes no stance on the questions of how many types of memory systems there are or whether there are any memory systems such that they operate exclusively in an explicit mode or an implicit mode, format pluralism is nevertheless incompatible with the MTL model's assumption that all implicit cognitive phenomena are best modeled associatively. Having recognized all of this, in making the world safe for format pluralist theorizing, it is not enough to merely show that the MTL model gets things wrong; it must specifically show that it gets things wrong because it falsely assumes it that propositional memory systems never traffic in states and processes that have the properties widely assumed to be characteristic of implicit states and processes.

Let's bracket for a moment the thorny question of which properties are *the* implicit ones and focus, instead, on outlining the general strategy that I will pursue in challenging the MTL model in a way that supports format pluralism. The strategy is this: the MTL model predicts that sustained and careful empirical investigation will fail to uncover implicit cognitive effects that could plausibly be

[33] Note that the implicit memory systems do not depend on the integrity of the medial temporal lobe. The model is named, then, for the region of the brain on which explicit memory depends.

attributed to the operations of episodic or semantic memory. However, I set out to show that such investigation has uncovered such effects. Therefore, the MTL model falsely assumes that propositional memory systems have no direct causal role to play in the production of implicit cognitive phenomena. Moreover, since the memory systems models that were most instrumental in the discovery of these effects and also critical in accounting for them are format pluralist friendly, it is not unreasonable to suspect that any empirically adequate memory systems model can be given a format pluralist gloss.[34] Finally, note that to the extent the strategy that I just outlined succeeds in these aims, I will also have satisfied the secondary goal of undermining the DPT dogma that only associative states and processes directly contribute to indirect measure performance.

Before considering two of the MTL model's chief rivals and the evidence that each cites in its favor, let's briefly return to the question just set aside of *which* properties are those that are widely assumed to be characteristic of implicit cognition. What makes this question thorny is that there is no universally agreed criteria to appeal to. While nearly everyone agrees that what Bargh (1994) has famously called the *four horsemen of automaticity* play an important role in carving out the implicit states (and/or processes) from the explicit ones, there is little agreement as to which of for properties a state must possess or how many it must possess for it to be considered *truly* implicit.[35] Indeed, the proposals that have historically received the most support are nonstarters on the pain of being either too strong or too weak.

[34] If this conclusion feels like it was arrived at too hastily, it is important to recognize that, for reasons to be provided as we move through the discussion, semantic and episodic memory systems are paradigmatically propositional. In any case, insofar as there are any propositional memory systems, these will be counted amongst their ranks. And since I will be adopting a defensible conception of what should count as an implicit cognitive state or process, this is not an unreasonable conclusion to reach.

[35] The four horsemen (or key features) relevant to classifying a state or process as automatic are awareness, intentionality, efficiency, and control. One lacks awareness of a state or process if either the individual lacks awareness of the contents of the state, lacks awareness of the processes that generated the state, or lacks awareness of how the state influences behavior. A state is unintentional if the state's activation does not depend on agential intent. A state or process is inefficient if the state is incapable of influencing behavior under conditions of high cognitive load. And a state is uncontrollable if one cannot inhibit it after it has been activated or one cannot prevent it from manifesting in behavior.

Here's an example of a view that's too strong: a state is implicit if and only if the agent (a) lacks introspective access to its contents, (b) has little to no control over its activation or retrieval from memory, (c) is unable to inhibit or stop either its activation or its effects on behaviors upon activation, *and* (d) can be retrieved even under conditions of high cognitive load. While this is the classic DPT stance, even the most ardent contemporary proponents of DPT have gone to great lengths to distance themselves from this view and for good reason: if one's goal is to provide an exhaustive taxonomy on which all cognitive processes are of one type or the other, then the classic DPT position fails on the grounds that even the paradigmatically implicit states and processes generally fail to satisfy all four conjuncts (for discussion, see Pennycook et al. 2018).

Other proposals are too weak. Consider the position according to which a state is implicit if and only if it satisfies one or more of the conjuncts just stated. This view, as Bargh (1994) stresses in his classic paper, fails to exclude putative explicit processes from counting as implicit. To make matters worse, an upshot of this disjunctive approach is that the same state or processes can count as implicit by satisfying one of the disjuncts and explicit by virtue of failing to satisfy another.

In an attempt to strike a middle ground between the two proposals in a way that avoids the pitfalls of both, contemporary DPT theorists tend to embrace what some have called a *defining feature and typical correlates* (DFTC) approach. The general idea behind the approach is this: one begins by selecting a pair of mutually exclusive features that are used to categorize any given mental state and/or process, thereby avoiding the internal consistency problem by guaranteeing that the same state/process will be such that it fails to possess both properties, and proceeds to identify a set of features that (putatively) serve as typical correlates of each respective defining feature, which ensures that each category is both robust enough to be scientifically interesting and weak enough to insulate it (not necessarily unfairly) from falsification. To provide just one example, one might define all states/processes as either being propositional or associative, and then claim that associative processes are such that they are *typically*, but not always, unintentional, unconscious, uncontrollable, and efficient. The upshot is that any successful critique of an account of this sort requires that one either show that (a) the defining features aren't actually mutually exclusive or exhaustive, or (b) the features selected as typical correlates fail to correlate with the corresponding defining feature in across a wide enough of expected conditions.

Though there are a number of objections that can and have been raised against this sort of approach to DPT theorizing, a thorough exploration of each critique will take us too far off the main path. Having said that, it must be stressed that, for the purpose of the project at hand— that is, for the purpose of identifying propositional states and processes that relevant to accounting for some implicit cognitive phenomenon—- the defining features and typical correlates approach is of little use. Two of the reasons for this are worth briefly discussing. First, even amongst contemporary DPT theorists, there is little agreement as to which are *the* defining features and correlates of implicit cognitive states and processes, and so the question of which are the properties diagnostic of implicit cognition has not been resolved. As such, the states and processes identified in the discussion that follows may count as implicit/explicit on one such proposal but not another, placing us right back in our original position of having to (fruitlessly) engage in a debate over which states and processes are *really* the implicit/explicit ones.

To illustrate the second concern, suppose we were to settle on a particular view on which all implicit states and processes are associative, and the four horsemen (or some subset thereof) serve as the typical correlates. Now let us suppose that we observe that some propositional state or process such that it possesses all such features. Will we have thereby shown that some propositional states or processes ought to be classified as implicit states or processes? The answer seems to be no, as the proponent of this approach could simply dig in one's heels and claim that any putative instance of propositional implicit cognition is actually an atypical case of explicit cognition. The fundamental problem, then, seems to be that without precisely specifying the conditions under which the states and processes are expected to instantiate the relevant features, such a proposal can be rendered consistent with any set of recalcitrant observations. While this is a general problem that besets any DFTC account, the salient point, for our purposes, is that this leaves us no clear criteria that can be used to determine when we have found a propositional state/process that ought to be counted as implicit.

The best way to proceed, it seems to me, is to leave it as an open empirical question whether propositional states and processes can sometimes be regarded as implicit states and processes. In line with this approach, I'll present a range of findings that suggest that propositional states and processes do exhibit properties have long been recognized as being hallmarks of implicit cognition on some defensible characterization of the implicit/explicit distinction. If one takes issue with any particular labeling of a process or state as implicit, then the burden of proof is them to succeed where so many others have failed in delivering a defensible dual-process taxonomy. To put it crudely, this is a case of put up or shut up. And while I am not in general a fan of such burden shifting arguments, *sans* any widely agreed upon set of criteria that one could use to diagnose instances of implicit cognition, this approach seems warranted in this context.

So much for the set up. Let's now consider the first of the two memory systems models that challenge the MTL model in a way that supports format pluralism.

2 Tulving's SPI Model and Implicit Semantic Memory

Let's now consider the challenge that Endel Tulving's (1995) *serial-parallel-independent* (SPI) model poses to the MTL model. The SPI model differs from the MTL model in at least three respects, each of which are especially relevant to the project at hand. First, the SPI model assumes that the semantic and episodic memory systems are independent memory systems. Second, the SPI model assumes that the semantic memory system is such that it may operate either in an explicit mode or an implicit mode depending on task demands. Third, as suggested by this last assumption, the SPI model treats the implicit/explicit distinction as not being applicable to memory systems *simpliciter* but as psychologically and behaviorally distinct expressions of memory systems, which implies that one and the same memory system engage in either implicit or explicit processing.

The structure of this section is as follows. First, I elaborate on the three differences just identified. Second, I go on to canvas empirical support for the SPI model's assumptions that semantic and episodic memory are distinct memory systems *and* that semantic memory in particular may be expressed implicitly. Along the way, I address various potential objections.

2.1 The SPI model

Tulving and colleagues' SPI model is first and foremost a model of how the mind-brain's various *cognitive* memory systems (where, 'cognitive' here is a technical term to be defined below) are related (Tulving 1995, Tulving and Schacter 1994). In this respect, it is broader in scope than the MTL model (as the MTL only specifies the relationship between the subsystems of episodic and semantic memory). But before we can examine the SPI model in detail, we must first introduce the memory systems that Tulving posits and briefly explore the ways in which the taxonomy that Tulving and colleagues imposes on the various memory systems differs from the MTL model.

Tulving (1995) intends for the SPI model to hold only for cognitive systems. Whether it holds for procedural memory systems, he thinks, is a topic for further study. But how are we to understand this difference?

> [Procedural memory] is a vast category, as yet largely unexplored and unknown. It probably comprises several further major divisions and a large number of rather specific subsystems, only some of which have so far been tentatively identified (*e.g.,* Squire, 1992a, this volume) […] It is involved in learning various behavioral and cognitive skills and algorithms, its productions have no truth values, it does not store representations of external states of the world, it operates at an automatic rather than consciously controlled level, its output is noncognitive, and it can operate independently of the hippocampal structures (Hirsch 1974; Squire 1987). Procedural memory is characterized by gradual, incremental learning and appears to be well-suited for picking up and dealing with invariances in the environment over time (Sherry & Schacter 1987)[…] Because of our present lack of information about the vast terra incognita that we call procedural memory, its most adequate description at the present time probably is by exclusion: procedural memory refers to a system, or systems, concerned with learning and memory functions other than those supported by the other four major systems. (Tulving & Schacter 1994: 26-27)

Now consider the way in which Tulving contrasts procedural memory with cognitive memory:

> One of the major differences between the two kinds of system lies in the feasibility of characterization of the changes that result from learning or acquisition in a propositional or some other symbolic form: it is possible to do so for the four cognitive systems, but not quite possible for the procedural systems. The operations of procedural memory are expressed in the form of skilled behavioral and cognitive procedures independently of any cognition. (Tulving 1995: 840)

So, in claiming that the SPI model holds only for cognitive systems, SPI intends for it to describe the ways in which the various propositional memory systems are related (for a list of the memory systems, see **Table 3**).

Table 3. The SPI Model

System	Alternative Names	Subsystems	Format	Retrieval
Procedural	Nondeclarative	Motor Skills, Cognitive skills, Simple conditioning, Simple associative learning	Nonpropositional	Implicit
(1) Perceptual Representation System (PRS)	Priming	Structural description, Visual word form, Auditory word form	Propositional	Implicit
(2) Semantic	General Factual Knowledge, Encyclopedic memory	Spatial, Relational	Propositional	Implicit
(3) Primary	Working memory, Short-term memory	Visual, auditory	Propositional	Explicit
(4) Episodic	Autobiographical, Event memory		Propositional	Explicit

At this point, one might worry that Schacter & Tulving are merely relabeling the implicit/explicit distinction. If so, this might spell trouble for my attempt to argue that propositional memory systems are involved in the production of implicit cognitive effects. To be sure, This is an understandable concern given how deeply entrenched DPT orthodoxy is in cognitive science, as it is easy to make this inference if one assumes that a state or process is implicit *if* it is associative, automatic, and unintentional and if one also assumes that a state or process is explicit if it is propositional and expressed in the form of skilled behavioral and cognitive procedures.

However, this is not how Schacter & Tulving (1994) draw the implicit/explicit distinction. More specifically, they explicitly deny that is appropriate to apply the implicit/explicit distinction at the level of memory systems:

> References to the explicit memory system and the implicit memory system are not uncommon in the literature. Explicit and implicit memory are not systems. These terms were put forward to describe and characterize expressions of memory: "explicit" refers to intentional or conscious recollection of past episodes, where "implicit" refers to unintentional, nonconscious use of previously acquired information[…]Thus, according to this formulation, implicit and explicit memory, though psychologically and behaviorally distinguishable forms of memory, could either depend on the same underlying memory system or different underlying memory systems; the question is open and subject to experimental investigation. (*ibid*: 12-13)

One finds in Tulving and Schacter, then, a conception of the implicit/explicit distinction that is much narrower in scope than is typically found in the social and cognitive psychological literature in at least two respects. On this view, only the processes that govern the retrieval and/or use of mnemonic structures are those that can be appropriately described as either being implicit or explicit; whereas classical DPT assumes that for any given psychological state or process it is either implicit or explicit. Second, Tulving clearly privileges the properties of intentionality and awareness in his conception of the implicit/explicit distinction. Importantly, whatever its theoretical shortcomings, this account at very least provides us with a defensible set of criteria that can be (comparatively speaking) straightforwardly applied in diagnosing instances of implicit cognition. These are the criteria that we'll be using in the discussion that follows. Should one take issue with this approach again, it is incumbent upon them to provide a set of criteria that could better serve as the mark of the implicit.

With the preliminaries out of the way, let's now turn to describing the SPI mode's core assumptions. Tulving claims that the SPI model's "central assumption is that the nature of relations among different cognitive systems is *process specific*: The relations among different cognitive systems depend on the nature of the processes involved as follows" (Tulving 1995: 843). The SPI model assumes that information is *encoded* by each system *serially* (hence the 'S' in SPI) in that "the encoding in one system is contingent on the successful processing of information in some other system, that is, the output from one system provides the input into another" (*ibid.*). As far as encoding is concerned, it is helpful to think of the cognitive systems as being organized hierarchically, with those memory systems that appear further down **Table 3** being superordinate to those that appear below them.

With respect to the process of *storage*, the SPI model assumes that information is stored in different systems in *parallel* (hence the 'P' in SPI). Here's a particularly vivid illustration of this feature of the model:

> The information in each system and subsystem, even if it all originates in one and the same act of perception or "study episode,' is different from that in others, its nature being determined by the nature of the original information and the properties of the system. Thus, what appears to be a single act of encoding— a single glance at a visual display, or a single short learning trial— produces multiple mnemonic effects, in different regions of the brain, all existing (*i.e.,* available for potential access) in parallel. Thus, with respect to storage, different systems operate in parallel. (Tulving 1995: 843)

Third, with respect to the process of retrieval, the SPI model assumes that information can be retrieved "without any necessary implications for retrieval of corresponding information in other systems" (*ibid*). In other words, with respect to retrieval processes, different systems are *independent* (hence the 'I').

Having described the sense in which the memory systems operate in ways that are serial, parallel and independent, let's illustrate this with a quick example. Let us suppose that we can type a random encounter with a stranger on the street as a single perceptual event, E_1. As the information contained by this makes its way serially though each of the four cognitive memory systems, E_1 is such that it results in the parallel storage of four distinct mnemonic structures; (M_{prs}) The visual representation of the stranger, (M_{sem}) A memory with the content THE STRANGER IS TALL AND HAD SHORT HAIR, (M_{wm}) A representation of some aspect of the encounter temporarily stored in working memory, and, (M_{epi}):A memory with the content I ENCOUNTERED THE STRANGER ON WASHINGTON AT

9PM. Crucially, each mnemonic structure may be retrieved, independently of the others, the service of carrying out specific tasks (*e.g.,* M_{prs} might be retrieved when one encounters the stranger on a subsequent encounter; M_{sem} might be retrieved when one is asked to report a description of the stranger's height, *etc.*).

Having laid the core commitments of the SPI model, we are now well positioned to see how this model's predictions differ from those generated by the MTL model with respect to semantic memory. Because the SPI model assumes that semantic memory provides the necessary inputs for episodic memory, the SPI model predicts a one-way dissociation between semantic memory and episodic memory (*e.g.,* we should expect to observe some cases in which damage to a neurological structure, N_1, would produce impairments to episodic memory and not semantic memory and no cases such that damage to a separate structure, N_2, impairs semantic memory but not episodic memory). More specifically, the SPI model predicts that under the right circumstances, we should find individuals that have intact, fully functional (more or less) semantic memory but with no ability to acquire, and therefore retrieve, episodic memories. By contrast, the MTL model predicts that one's capacity to acquire new semantic memories is dependent upon the integrity of the same neurological structure (viz. the hippocampus) involved in the acquisition (but not necessarily retrieval) of episodic memories. In other words, the impairment of the hippocampus should result in a failure to form new semantic memories and new episodic memories. Consequently, we'll have accumulated strong evidence for the SPI model over the MTL model if we find individuals who are afflicted by a particular sort whose amnesia such that it impairs their ability to form and retrieve new episodic memories but spares their ability to form and retrieve new semantic memories. Moreover, and for our purposes, more importantly, we will have also accumulated evidence against the MTL model to the extent that we can confirm the SPI model's predictions that retrieval from semantic memory is sometimes implicit. We now turn to look at some of the empirical support that has been marshaled in favor of the SPI model's prediction.

2.2 Developmental Amnesia as Yielding Evidence for Implicit Semantic Memory

Over the past three decades, memory researchers have gathered a great deal of evidence that confirms the SPI model's predictions. For instance, early explorations of episodic-semantic dissociations relied on individuals who acquired retrograde amnesia, anterograde amnesia or both as adults (Hayman, Macdonald, Tulving 1993). These studies revealed that it's possible for these individuals to "learn new semantic knowledge and retain it over a period as long as 30 months indistinguishably from control subjects" (*ibid*: 375).

One notable limitation of these earlier studies, with respect to the question at hand, is that these studies attempted to measure the acquisition of new semantic structures by having the subject respond with a target word in response to a cue-phrase, where the cue-phrase and the target word were presented together, repeatedly, over the course of several training sessions. While the authors' aim was to investigate the extent to which individuals with amnesia are able to acquire "'propositional' information about the world" (*ibid:* 376) in a manner that "differs from other forms of learning, such as simple conditioning and learning of skills (procedural memory), short-term memory, and perceptual priming" (*ibid.*), it is not obvious that the kind of learning that these individuals exhibit could not be better explained by appeal to the operations of other forms of associative memory systems (*e.g.*, the semantic-associative memory system, or the instrumental learning memory system). To rule out an associative explanation of these results, one would ideally want to see evidence that individuals with severe episodic amnesia are nevertheless able to acquire the kinds of mnemonic structures of the sort that would feature prominently in tasks that are designed to recruit paradigmatic propositional processes (*e.g.,* reasoning, planning, *etc.*).

It is here that we are better off looking to the burgeoning literature on *developmental amnesia* (DA) for empirical support. DA is typically observed in individuals with bilateral hippocampal lesions induced by neonatal hypoxia (Elward & Vargha-Khadem 2018). Early in childhood, individuals with DA exhibit age-appropriate levels of speech, language, and motor function (Elward & Vargha-Khadem 2018, Jolin et al. 2017). Their cognitive impairments typically come to the attention of parents and teachers around the age of 5, which generally coincides with the beginning of their formal education. While they tend to display age-appropriate levels of intellectual ability, working memory, language, and semantic knowledge (*i.e.,* knowledge of general facts) when formally assessed, they are nevertheless unable to recall autobiographical events when prompted (*e.g.,* "What did you do this weekend?") or remember classroom instructions. Additionally, these individuals frequently get lost and misplace important objects (*ibid*).

Researchers Rachael Elward and Faraneh Vargha-Khadem (2018) relay an anecdote involving a DA patient, Jon, which helps to vividly illustrate this unusual condition:

> Jon frequently visits our laboratory in London. To do so, he travels to an underground train station nearby, then takes the lift to the street level and walks the remainder of the journey. On one such visit, the lift at the underground station was out of order, and Jon had to climb 171 steps to the surface (the equivalent of some 14 floors). When he arrived at the laboratory, he had no recollection of having climbed the stairs, and confidently reported that he had taken the lift as normal. Jon was questioned about his memory of this event, "How do you know that you took the lift today?" Jon declared: "I always take the lift!" Why is Jon so confident that he

> *always* takes the lift, when he has no episodic memory of doing so? If Jon has no recollection of his life events as they occur, how does he learn what he typically does? (*ibid*: 23)

Importantly, Jon appears to be in possession of a wealth of semantic or encyclopedic knowledge stored in semantic memory, which subserves various behavioral and cognitive competencies. Moreover, it is generally assumed that such competences depend on one's capacity to represent states of affairs, and one's capacity to manipulate these representations of states of affairs in the service of planning, navigation, and responding appropriately and intelligently to dramatic changes in one's immediate surroundings. In other words, it is generally assumed that these competences are underwritten by one's capacity to store and manipulate propositionally encoded information. DA's case, together with the findings that individuals with DA progress normally compared to healthy controls in domains of mathematical ability, general reasoning ability, and the like (for review, see Jolin *et al.* 2017), challenges the MTL model's assumption that semantic memory and episodic memory are subsystems of a single explicit memory system. And more specifically, it challenges its assumption that *all* declarative memories start off as being hippocampus dependent and are later consolidated in the neocortex via repeated reactivation (a process known as *semanticization*).[36] Were semanticization necessary for the long term storage of semantic memory structures, it would be puzzling, to say the least, as to how individuals with DA could form semantic memories at all and therefore exhibit the range of competencies

[36] It is worth noting that, the bilateral damage to Jon's hippocampus is severe, but it is not so severe that it prevents him from forming *any* episodic memories. Moreover, fMRI data does reveal hippocampal activation across a variety of tasks, though it is severely limited relative to neurotypical controls.

associated with the deployment of semantic memory given that these individuals were born with severe hippocampal damage.

However, we are not yet in a position to make any inferences about whether either the processes that govern retrieval of structures from semantic memory or the impact of these structures on the execution of these underlying competencies are implicit (*i.e.,* unintentional and unconscious). Afterall, while Jon clearly has no idea how he arrived at the office, his ability to accurately report how he arrived at the office is dependent upon Jon's ability to store and retrieve memories from episodic memory system.

It is then fortunate for anyone looking to develop an MSP theoretic solution to the format problem that several studies involving subjects with DA suggest that such individuals do sometimes lack both content awareness (*i.e.*, introspective access to the contents of their memories) and impact awareness (*i.e.*, awareness of how the successful retrieval of memories drives behavioral performance) of their semantic memories. As a thorough review of this rapidly burgeoning literature is not possible here, I'll report only a small selection of the findings.

It is worth describing the methods and results of one of these studies in some detail. During the learning phase, individuals with DA and non-impaired controls were shown four informative short films on topics with which they are unfamiliar. Both groups were instructed to retain as much of the information presented in these documentaries as possible and that their retention abilities would be tested. After each video, their knowledge was tested using a recognition task (*e.g.,* Which social group believed that mistletoe was sacred? (I) Ancient druids (II) Medieval Priests, (III) The Mariners (IV) Tudor Farmers) in order to ensure that some initial learning had occurred. This initial assessment revealed that DA patients performed similarly to controls. Retention of information for two of the four videos was then tested after one day (short delay) while retention for the other two videos was tested a few days to a week later (long delay). Three types of retention test were administered. Across both conditions, all subjects were given a *free-recall task* (*e.g.,* Tell us whatever you can about the groups discussed in the video), a *cued-recall task*, which used the same questions as the recognition-task without the multiple-choice options, and, a *recognition task*, which consisted of the same set of questions and corresponding responses provided during the learning phase. Here's how the researchers described their findings:

> […]patients with DA struggled to free recall the information from the videos, and became upset when they were unable to remember the contents. However, they showed remarkably good performance during cued-recall […] Using this recognition-based learning paradigm, patients recalled 85% of the semantic information that had been presented in the videos, compared to just 35% in the previous version of the experiment that used recall-learning trials. *These data suggest that repeated tests using multiple-choice recognition can support the long term formation of new semantic memory despite the newly-acquired memoranda being unavailable to free recall.* (my emphasis) (Elward & Vargha-Khadem 2018, 28)

So, subjects with DA required a greater number of learning trials for the successful encoding of the relevant information than healthy subjects, they were nevertheless able to do so.

It must be stressed that the researchers found that individuals with DA required *fewer* learning trials than would be expected if learning were subserved by a merely associative memory system (Elward & Vargha-Khadem 2018). Thus, the kind of learning that DA subjects exhibit is slower than would be expected were they relying primarily on hippocampal-based memory systems, but faster than would be expected were they relying on structures that merely encode statistical contingencies.

These data allow us to draw reasonable, yet defeasible, conclusions regarding whether the retrieval of mnemonic structures from semantic memory and their impact on subsequent behavior (revealed by their response patterns) are implicit. Since the contents of DA patient's semantic memories are unavailable to verbal report, which has, across disciplines, long been regarded the most important standard by which attributions of content awareness are to be judged, we can conclude that the outputs of semantic memory are expressed below the conscious threshold. Put differently, were the processes that govern retrieval of semantic memory to always result in explicit output, we would have expected subjects to exhibit similar levels of performance across all three measurement tasks.

Second, this research allows us to make some reasonable, albeit comparatively weak, inferences about whether the processes that govern the retrieval of information from semantic memory are intentional. The fact that DA subjects become upset when prompted to answer free-recall questions suggest about the relevant material further suggests that (a) DA subjects lack both introspective access to the contents of their memories, and (b) that intention is not sufficient for successful retrieval the relevant information for use in task responding. Since the presence of specific cues is required for successful performance, it is also reasonable to conclude that the processes that govern retrieval of information from semantic memory does not depend on intention. To further explore questions of the role of intention (or lack thereof) in initiating retrieval from semantic memory, researchers may want to place stringent constraints on the amount of time each participant has to respond during cued-recall and recognition tasks (but see Jonin *et al.* 2017 for several studies involving a patient with a particularly severe form of DA that did utilize very strict speed constraints and found analogous response patterns).[37]

[37] One might worry that I have both misrepresented the core commitments of the MTL model. With respect to my claim that the research on DA supports the existence of a dissociation between episodic and semantic memory, one might claim that the MTL model does not assume that semantic and episodic memory should be spared *if, and only if*, the

2.3 Conclusions and Limitations

There are four morals to this story. The first is that there is that there is strong empirical evidence for the SPI assumption that semantic memory and episodic memory are really distinct memory systems. This is supported by evidence for a one-way dissociation between semantic memory and episodic memory. Second, the literature on DA also reveals that the expressions of semantic memory are sometimes implicit.[38] The third is that even if we should ultimately have to reject the SPI model on other grounds, these particular patterns of results can be explained only by memory systems models that assume that propositional memory systems are capable of implicit forms of expression both in terms of retrieval operations and the impact of these operations on behavioral outcomes. Fourth, it is worth emphasizing only by breaking from DPT orthodoxy were we able to predict the results described in this section. In other words, at least one model that breaks from standard DPT assumptions has been shown to generate novel and risky predictions about the

hippocampus is preserved. Rather, the whole declarative (or explicit) memory system depends on the integrity of the *extended hippocampal system*, "which includes the hippocampus, fornix, mammillary bodies, mammillothalmic tract, anterior thalamic nuclei, and the retrolental cortex" (Jonin, Besson, La Joie, Pariente, Belliard, Barillot, Barbeau 2018). The DA research presented does not rule out the possibility that, while hippocampal atrophy is severe, other intact structures in the extended hippocampal system support the acquisition of semantic memory. Additionally, the fact that many of the subjects with DA still have some lingering episodic memory function, it is possible that residual hippocampal function could be contributing to semantic learning (*e.g.,* one might have just enough episodic memory function so as to allow the subject to retrieve their episodic memory of the learning event the requisite number of times so as to result in the formation of a decontextualized semantic memory). If either is the case, then this would weaken my case against the MTL model's assumption that episodic and semantic memory are independent memory systems.

But even if one grants that this research does not decisively point to a one-way dissociation between semantic and episodic memory, it is nevertheless the case that the relevant research supports the existence of implicit semantic memory retrieval. So, even if semantic and episodic memory are two components of the same overarching memory system, the MTL's classification of the semantic memory as a form of explicit or declarative memory would have to be rejected.

[38] For the sake of space and overall flow, I have omitted a lengthy discussion of the prospects of one recent strategy that the proponents of the MTL model have used to push back against the considerations from the DA literature articulated above. I have included this discussion, however, in the addendum that appears at the end of this chapter.

existence of a certain type of cognitive phenomena, which suggests that DPT theorizing has outlived its usefulness.

Before moving on, I should remark that none of the data that I have presented rules out the possibility that Tulving and others have erred in treating semantic memory as a propositional memory system. Indeed, this is a matter of ongoing debate. As such, it is possible that future investigation will show that semantic-associative memory best explains results just presented.[39] In light of this concern, it is worth describing another recently developed memory systems model that predicts that the processes involved in the retrieval and expression of episodic memory are sometimes implicit. Since there is widespread agreement that the episodic memory system is a propositional system, evidence for such a prediction would further strengthen the case for a format pluralist approach. We now turn to this topic.

3 Henke's Processing-Based Model and Implicit Episodic Memory

The SPI model is hardly the MTL model's only competition. Indeed, Katharina Henke's (2010) *processing based* (PB) memory systems model stands as a viable alternative to both the MTL model and the SPI model. Like Tulving's SPI model, the PB model predicts the existence of implicit semantic memory. Unlike Tulving's model, however, the PB model denies that the formation and retrieval of episodic memory structures is always conscious affair. That is, the PB model predicts that the operations of both semantic memory and episodic memory can be expressed implicitly. For

[39] Let us suppose though that all the findings presented in this discussion and in the discussion of implicit episodic memory that follows can be explained associatively, presumably by appealing to the joint operations of multiple associative memory systems. If so, we'll have a situation in which the MSP perspective is true and format pluralism is false. This would be a result that I would be comfortable with, as the defense of the MSP perspective is the primary aim of this dissertation. With that said, I am willing to place my empirical bets with format pluralism.

reasons already given, if these assumptions are correct, then the case for a format pluralist approach would be strengthened considerably.

Even at this stage in our discussion, one might object that the very assumption that episodic mnemonic processes are capable of implicit expression is awash in conceptual confusion. In particular, one might be inclined to treat the claim that the outputs of episodic are always explicit is one that expresses a conceptual truth about episodic memory. Indeed, philosophers and cognitive neuroscientists of memory alike are inclined to treat it as such (see Mahr & Csibra 2018, Wheeler 2000). Tulving has been particularly insistent retrieval from episodic memory is always conscious:

> The most distinctive aspect of episodic memory is the kind of conscious awareness that characterizes recollection of past happenings. This awareness is unique and unmistakably different from the kinds of awareness that accompany perceptual experiences, imagining, dreaming, solving of problems, and retrieval of semantic information. To distinguish the episodic-memory awareness from these other kinds, it has been referred to as autonoetic consciousness (Tulving 1995: 841-842).

To be sure, while Tulving does not come right out and say that the target claim expresses a conceptual truth, it is difficult to interpret his remark on the "unique" and "unmistakable" phenomenology that characterizes episodic recall in a way that would render it a contingent, empirical claim. An upshot of this view is that one cannot even conceive of an instance of *truly* episodic recall that lacks autonoetic consciousness much less empirically demonstrate the existence of such instances.

I raise this objection at this early juncture only to bracket it until after we have presented the PB model and the data that Henke and colleagues marshal on its behalf. For readers who are inclined to accept this claim about episodic memory as expressing a contingent truth about episodic memory, I ask only that one keep an open mind as we proceed. If, on the other hand, one truly thinks that implicit episodic memory is a contradiction in terms, then I invite the reader to recast the following discussion in terms of episodic* memory. Having set this issue aside for the time being, I now turn to presenting the PB model.

Henke's PB model, unlike the SPI and MTL models, assumes the existence of types of memory system, which are typed in terms of what Henke calls their *processing characteristics* (see **Figure 4**). On Henke's proposal, there are memory systems that (i) *rapidly* encode and *flexibly* retrieve *compound* representations (hereafter, RFC-systems); (ii) *rapidly* encode and *flexibly* retrieve single, *unitized,* representations (RFU-systems); (iii) and, memory systems that *slowly* encode and *rigidly* retrieve compound representations (SRC-systems).[40] Crucially, these descriptions omit any reference to consciousness, control, and resource dependence. As such, the model makes no assumptions as to how these properties will be distributed across these three types of memory system. Thus, Henke assumes with Tulving and Schacter (1994) one and the same memory system may support implicit and explicit states and processes. Henke makes no claim that these three categories are exhaustive— further empirical investigation may turn up additional memory systems that do not fall under any of these three types.

[40] It is worth noting that nowhere in Henke's (2010) discussion will you find the term 'compound representation'. Instead, she refers to these representations as *associations*. It should be of no great mystery as to why I would prefer to avoid introducing yet another sense of 'association' in this chapter.

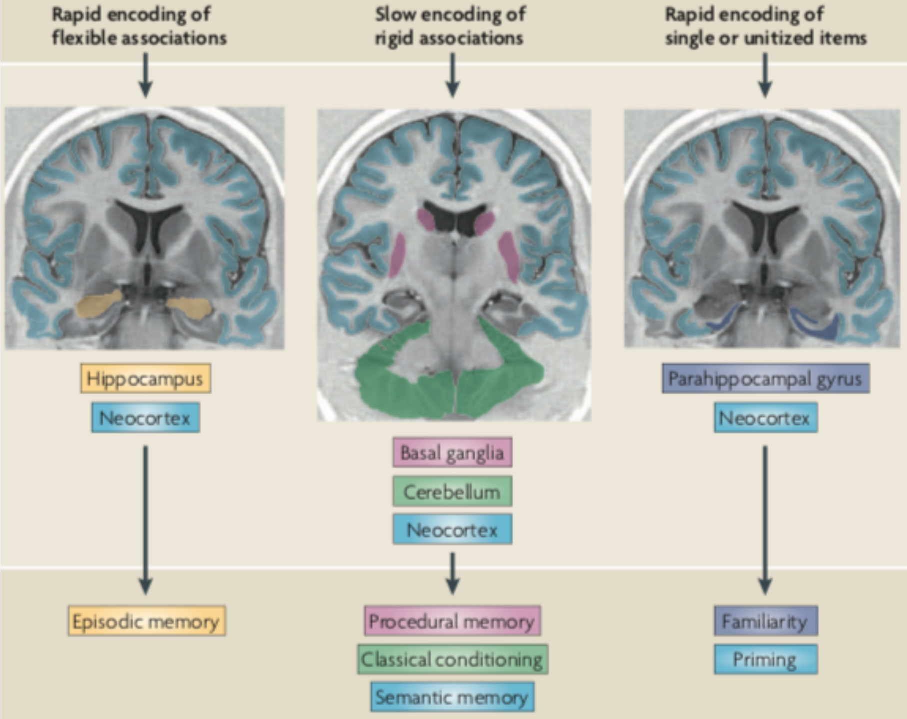

Figure 3. The Processing-Based model. From top to bottom: the distinguishing features of each type of memory system; fMRI images and corresponding labels of the putative neurological substrates of each memory system; and the specific memory systems that fall under each heading. See fn10 about how to understand that specific use of 'association' as it appears in this figure.

Whether a memory system encodes mnemonic structures rapidly or slowly depends on the number of learning trials needed for successful storage in long-term memory (LTM). A flexibly retrievable memory is one on which is can be retrieved in situations that differ from the encoding situation (*e.g.,* one might recognize Smith whether one sees her straight-on or merely has a view of her profile). A representation, R, is a compound representation if and only if R is composed of several elements— say, e_1, e_2, e_3— such that it is possible for e_1 to be activated and processed independently of e_2 and e_3. Consider for instance, my memory structure with the content I JUST HEARD FAVORITE PUBLIC IMAGE LTD SONG AT A COFFEE SHOP. This representation, R, is a compound representation if I can access and retrieve any arbitrarily selected element of R (*e.g.,* the song) independently of the other elements of which the compound memory is composed (*e.g.,* the place in which the song was encountered). A unitized representation, by contrast, is one in which one cannot retrieve any element of R without activating and retrieving of all of R's other elements. On the PB-Model, episodic memory is an RFC-system, semantic memory is an RFU system, and priming is an SRC-system.

Henke (2010) maintains that the processes that govern the formation and retrieval of episodic memories can be initiated automatically and below the conscious threshold. Moreover, Henke argues that the contents of episodic memories can be manipulated in computationally sophisticated ways automatically and below the conscious threshold. In other words, the PB model assumes the existence of a form of episodic memory that lacks autonoetic consciousness.

Why think that Henke has accurately characterized the processing characteristics of episodic memory? Here, Henke exploits three widely shared assumptions about episodic memory's functional properties. The first is that the episodic memory system is capable of binding information about a particular event (*e.g.,* ENCOUNTERED JONES) with representations of the spatiotemporal context in which the item occurred (*e.g.*, AT THE COFFEE SHOP, ON THURSDAY) after only a single learning episode. The second assumption is that this process results in the rapid formation of a single compound representation (*viz.* I ENCOUNTERED JONES AT THE COFFEE SHOP ON THURSDAY) conventionally known as a *What-When-Where* (WWW) *representation.*

These two assumptions enjoy a great deal of empirical support. For instance, individuals with amnesia perform poorly on tasks that can only be successfully completed by being able to accurately indicate where on the computer screen the image had previously occurred) even when they demonstrate preserved memory for the item itself (see Henke 2010 for review). Similarly, those with amnesia tend to be selectively impaired in their ability to rapidly encode and subsequently retrieve how two arbitrarily selected items are related (*e.g., That* face was presented on *that* particular scenic background) compared to their ability to recognize the formerly presented individual items (Hannula, Ryan, Tranel, Cohen 2007). Moreover, the episodic memory system's ability to bind together WWW information into a single mnemonic structure is related to its ability to rapidly bind together, or integrate, otherwise unrelated information more generally. For instance, individuals with hippocampal damage are able to reliably recognize pairs of words that had previously appeared on a study list when the elements are semantically related (*e.g.*, furniture-table) but fail to reliably recognize previously studied word-pairs when their elements are semantically unrelated (*e.g.*, furniture-sky) (Shimamura & Squire 1984).

Having described and motivated the relevant aspects of Henke's PB model, the question now before us is whether there is sufficient empirical support for the assumption that WWW representations can be rapidly constructed and flexibly retrieved to guide behavior independent of intent and conscious awareness.

3.1 Empirical Support for Implicit Episodic Memory

What research does Henke cite in support for the existence of implicit forms of episodic memory? Consider one set of studies conducted by Hannula, Ryan, Tranel & Cohen (2007). During the learning phase, healthy participants studied various human faces each of which was superimposed on a scenic background. All study items appeared with the same frequency over the course of the study blocks to rule out the possibility that subjects were merely responding to stimuli on the basis of relative familiarity. During the test phase, subjects were then shown three faces superimposed over a single scenic background. In each case, at least one of the three faces had appeared against the same background in the study block.

Using eye-tracking gear, researchers found that subjects expressed a non-verbal preference (as measured by looking times) for the face that had been paired with the same background during the learning phase. Importantly, this result was obtained even when (a) subjects were unable to explicitly recall which face had been paired with which scene and (b) when subjects were never even prompted to declare or otherwise attend to the face that had been earlier paired with a specific background. That (a) was observed suggests that studies were able to rapidly encode compound memory structures (viz. arbitrary face-scene representations) despite the fact that this structure could not be consciously retrieved. That (b) was observed suggests that such representations can be automatically, or spontaneously, retrieved from memory even in the absence of any obvious goal. Importantly, when the same experiment was conducted with individuals who suffered from hippocampus-induced amnesia, the researchers found that these subjects failed to exhibit an enhanced indirect preference for the relevant face-scene pairs, providing us with strong evidence that the target capacity critically depends on the integrity of the hippocampus. This last result helps us to rule out the possibility that performance was driven by, for example, semantic memory.

Encouraged by these findings, Henke and colleagues conducted several follow-up studies that investigate the extent to which individuals can encode complex spatial representations from three-dimensional scenes even when the scenes are only ever presented subliminally. In the most recent of these studies (Wuethrich, Hannula, Mast & Henke 2018), not only had researchers found that subjects could form complex representations that encode spatial relations between various objects in a three-dimensional scene that depicts a living-room despite the fact that subjects were unaware of ever having been exposed to these scenes, but that that these representations could be flexibly retrieved in the service of solving puzzles that would appear to involve relatively complex spatial reasoning. The experimental set-up was as follows: the learning phase consisted of the subliminal presentation of a three-dimensional scene, which depicted a living-room with several critical objects (*e.g.,* a vase, a lamp, etc.). After a manipulation check to ensure that subjects were unaware of the stimuli, the test-phase began. During the test-phase, subjects were subliminally presented with either an object that had appeared in one of the study scenes, or a novel object that had not appeared in an earlier training session. Next, subjects were supraliminally presented with one a scene that had been presented during the learning phase *sans* critical objects. Moreover, the scenes that appeared supraliminally were rotated anywhere between 0º-270º. While the scenes were available, subjects were permitted to engage in non-goal directed viewing (free-viewing) for six seconds. Using eye-tracking gear, researchers found that when the subliminal object cue used depicted the same object that had previously appeared in that scene, subjects spontaneously and disproportionately spent more time looking at its previous location than they had at the locations that were missing the other objects. After viewing time was measured, two locations in the scene were highlighted and the subjects were forced to respond with a decision as to which of the highlighted locations the subjects intuitively felt that a particular object should be placed. Researchers found that when the object the subjects were asked about was the same one that had appeared as the subliminal cue, subjects disproportionately responded with the accurate location, adjusted for spatial rotation. Importantly, prior to the test, subjects were generally assessed for whether they made decisions in a more intuition-driven or deliberation-driven manner. While both intuiters and deliberators were observed to have disproportionately viewed the location of the subliminally presented object during the free-viewing task, intuiters were more likely to indicate the correct location of the subliminally presented object during the forced-choice task than their more deliberative counterparts. These findings suggest that there exist memory structures that encode relational information such that they can be formed, activated, and flexibility retrieved involuntarily and without conscious awareness.

3.2 Autonoetic Consciousness and Episodic Memory Revisited

Now that we have canvassed some of the results that impress proponents of the PB model, let's now return to the objection introduced earlier. The argument, recall, went something like this. Tulving claims that (1995) a particular phenomenology (viz. autonoetic consciousness) necessarily accompanies every retrieval of an episodic memory. But autonoetic consciousness was absent in the finding just presented. So, the objection goes, whatever it was that Henke (2010) was investigating, it could not possibly be episodic memory.

Indeed, Mahr and Csibra (2018) contend that empirical investigations of so-called "implicit" episodic memory reveal only that there exists a kind of memory that is related to but distinct from episodic memory—*viz. event memory*. Event memory, like episodic memory, is the output of a hippocampal-based scene-construction mechanism. But on Mahr and Csibra's view, "Hippocampus-based constructions become episodic memories only when they are conceptualized in a certain way, namely, as the outcome of past first-person experience. (*ibid*: 3)" The difference, then, between episodic memory and event memory is that event memory involves hippocampus-based constructions that need not be embedded in this metarepresentational attitude in order to find expression in behavior (*ibid*: 4).

Faced with such an objection, one is presented with the following dilemma: One can concede the conceptual claim that retrieval of episodic memory necessarily accompanies autonoetic consciousness, or one can deny it. If one concedes the conceptual claim, then one must also concede that empirical results just canvassed fail to establish the existence of implicit forms of episodic memory. One upshot would be that Henke and colleagues were actually investigating a different kind of memory. If one denies the conceptual claim, then Henke and colleagues' interpretation of the data holds. By taking the second horn, however, it seems as though we will have reached a dialectical stalemate, with intuitions that conform with folk psychological theorizing on the one hand pitted against various scientifically driven considerations on the other.

There is a sense in which the question of whether retrieval of episodic memory necessarily involves autonoetic consciousness is a red herring. After all, the primary question that we're trying to address is whether the response patterns described in this section are driven by the retrieval of propositional structures operated on by propositional processes. Given that there is no reason why we should think otherwise, the present project would not be jeopardized were one to take the first horn of the dilemma so long as the memory system that subserves what Mahr and Csibra call "event memory" (or what we had above called episodic* memory) is best characterized as a propositional memory system. To be sure, while I am skeptical of the claim that the same memory system that involved in the formation, manipulation, and retrieval of event memory is not the same memory system involved in the formation, manipulation and retrieval of episodic memories, for our purposes we need not settle this debate here.

4 Concluding Remarks

The format pluralist solution to the format problem presupposes that any arbitrarily selected indirect (or implicit) measurement task stands in a one-to-many relation with respect to the mind-brain's memory systems, some of which are associative and some of which are propositional. In this chapter, I identified and addressed two worries that may prevent many from taking this kind of solution to the format problem seriously. One is that while the MTL model is consistent with the assumption that indirect measurement tasks stand in a one-to-many relation with respect to the mind-brain's memory systems, it denies that any of these memory systems subserve propositional states and processes. Without reason to reject the MTL model's format monist assumption, so the worry goes, the format pluralist solution is insufficiently motivated. The other worry is that many cognitive scientists still find themselves within the grip of the DPT dogma that all implicit cognitive states and processes are associative states and processes. The purpose of this chapter was to clear a path for a format pluralist solution to the format problem by removing both obstacles. If in the process I have also succeeded in striking another blow against DPT theorizing in social and cognitive science more generally, then this is just icing on the cake.

Rather than summarize my response to the two worries, I find that it would be more useful to make a pair of additional remarks with the aim of staving some potential misunderstandings. First, I would be remiss were I to leave the reader with the impression that the two propositional memory systems described in this chapter will each have some privileged role to play in accounting for the phenomena that give rise to the format problem. Whether they actually play such a role is an empirical question, one that I have made no attempt to address. This might raise some additional concerns were it either the case that (a) these are the only two propositional memory systems that our best cognitive scientific models posit, or (b) these are the only two memory systems that could potentially influence indirect measures of attitude. It is fortunate, then, that (a) is false and (b) is unmotivated.

Second, and somewhat relatedly, by removing these two theoretical obstacles to the development of a format pluralist solution to the format problem, I do not take myself to have thereby shown that this project will ultimately be successful. Again, this is very much an open empirical question. The success of this approach depends, among other things, on how well it accounts for the phenomena that generate the format problem, whether the explanations it provides are sufficiently theoretically motivated, the degree to which it opens up new and productive avenues of future research, *etc.* These issues are taken up in the following chapter.

CHAPTER 6

TOWARD AN MSP THEORETICAL RESOLUTION OF THE FORMAT PROBLEM

The previous two chapters laid the theoretical groundwork for a format pluralist resolution to the format problem. In Chapter 4, I characterized the format problem and argued that all extant attempts at resolving it have failed. Moreover, I suggested that the reason each view fails to resolve the format problem is that each is committed to IA-Monism. Rejecting IA-monism opens the door to the view that implicit attitudes not only come in multiple varieties (implicit attitude pluralism) but also that implicit attitudes come in multiple representational formats (format pluralism). In the previous chapter, I argued that two format pluralist friendly memory systems models better account for an array of findings pertaining to implicit propositional cognition than its dual-process theoretical rivals.

The primary objective of this chapter is to continue to lay the groundwork for a resolution to the format problem by advancing an MSP theoretic solution of one of the anomalies that animates it, namely the *Co-occurrence/Relational controversy* (hereafter, C/R controversy). A central assumption of the MSP theoretic resolution on offer is that for any effect, E, such that E is obtained by way of an implicit measure and E's presence reflects the activation and/or retrieval operations of attitudinal mnemonic structures in memory, E may be the direct product of either associative memory systems, propositional memory systems, or an interaction of both.[41] That implicit measure effects might be driven by either propositional memory systems or associative memory systems is

[41] The hedge is an important one. The MSP theory is by no means committed to the view that every implicit measure effect reflects attitudinal related states and processes. For instance, some effect may turn out to be driven not by the retrieval of attitudes but by the adoption of a situation specific response strategy (see discussion in §3). The assumption that some implicit measure effects are not genuine attitudinal effects is not unique to MSP theorizing.

what makes this particular version of MSP theory a format pluralist theory of evaluation. It is also what helps to distinguish this particular brand of MSP theory from its conceptual cousins (see Chapters 4 & 5). With this assumption in place, we can take as our initial investigative starting point that the learning and retrieval operations of propositional memory systems are more important than the operations of associative memory systems with respect to explaining the various putative propositional effects (*viz.* the relational effect (Hu *et al.* 2017: Experiment 3), the diagnosticity effect (Cone, Mann & Ferguson 2017), the reinterpretation effect (Mann & Ferguson 2015); we adopt the same assumption, *mutatis mutandis*, for the various putative associative effects (*viz.* the gradual change effect (Rydell & McConnell 2006), the invalidity effect (Peters & Gawronski 2011), and the co-occurrence effect (Moran & Bar-Anan 2013, Hu *et al.* 2017: Experiments 1 & 2)). While our initial classifications of some effect as associative or propositional may upon further investigation turn out to be incorrect, these provisional classifications provide us with as good a starting point for inquiry as any.

This chapter proceeds as follows. I begin by briefly reviewing the central findings that animate the C/R controversy (§1). Next, I describe in greater detail the general MSP theoretic approach I will adopt in an attempt to resolve this particular controversy and highlight some obstacles standing in the way of developing a fully satisfying MSP account of the C/R controversy (§2). I then articulate an MSP theoretic account of the co-occurrence effects, focusing primarily on those reported by Hu, Gawronski and Balas (2017; Experiments 1 & 2). In this section, I also argue that we ought not assume that all observed co-occurrence effects are the products of the same underlying memory systems (§3). Next, the focus shifts to providing an MSP theoretical account of the relation effect obtained by Hu et al. (*ibid.*) in their third experiment (§4). The chapter ends with some brief remarks about how the MSP theoretical approach might help to resolve the other two controversies at the core of the format problem.

1 The C/R Controversy Revisited

Before I sketch my solution to the C/R controversy, it is worth revisiting it briefly. As discussed in Chapter 4 there is now a sizable yet manageable literature documenting both *co-occurrence effects* and *relation effects* with respect to implicit attitudes, and this literature has garnered a great deal of attention from the most prominent researchers in the field (Gawronski & Bodenhausen 2018, De Houwer, Van Dessel, & Moran 2020; for a broader review, see Corneille & Stahl 2019). Co-occurrence effects are a type of classical evaluative conditioning effect, whereby the valence of an initially (or relatively) neutral conditioned stimulus (CS) shifts in the direction of an antecedently valenced *unconditioned stimulus* (US+/–) as a mere function of the frequency of CS-US pairings. What distinguishes a co-occurrence effect from other kinds of evaluative conditioning effect is that co-occurrence effects occur even under those conditions in which the subject is made aware of a relevant CS-US relation, the evaluative implications of which run counter to the evaluative implications of mere CS-US co-occurrence. Under this experimental set up, a *relation effect* is observed when subjects' attitudes toward a CS reflects or is influenced by the evaluative implications of the CS-US relation.

Consider an example. Suppose that Fisherman's Friend lozenges (CS) are reliably within reach only when I have a nasty sore throat (US–). This is no accident; after all, I recognize that Fisherman's Friend lozenges effectively *soothe* sore throats. My attitude toward Fisherman's Friend could be based either on the fact that Fisherman's Friend lozenges are only present when I have a sore throat or on the fact that Fisherman's Friend lozenges *soothe* sore throats. Note that in this case the evaluative implications of co-occurrence are incongruent with the evaluative implications of the relevant relation. If my attitude toward Fisherman's Friend is determined by mere co-occurrence, we would expect my attitude toward Fisherman's Friend to be negative; if my attitude toward Fisherman's Friend were based on my grasping of the relevant relation between Fisherman's Friend and sore throats, then one would expect my Fisherman's Friend attitude to be positive.

An upshot is that all empirical investigations of the co-occurrence effect involve at least two (coarsely typed) forms of learning: (i) the learning of CS-US co-occurrence, and (ii) the learning of the relevant CS-US relation for each CS-US pair. The learning of CS–US co-occurrence is generally facilitated through the use of the *repeated evaluative pairings* (REP) procedure:

> *Repeated evaluative pairings* (REP) refers to a learning procedure in which a conditioned stimulus (CS; *e.g.,* an image of the face of a novel individual, like "Bob' in the work of Rydell *et al.* 2006) is repeatedly paired with a valenced unconditioned stimulus (US; *e.g.,* electric shocks, unpleasant sounds, or words or pictures that denote valence.
> (Mann, Kurdi, & Banaji 2019)

This procedure is a staple of both investigations of co-occurrence effects specifically and investigations of evaluative conditioning (EC) effects more broadly. Recall that what differentiates more co-occurrence studies from more run-of-the-mill EC studies is the inclusion of information about the relevant CS-US relation. Thus, (ii) is typically facilitated by an explicit stipulation of the relevant CS-US relation at some point during the experiment. Frequently, the stipulation of the relevant CS-US relation(s) is provided in the task-instructions prior to the start of the REP/conditioning phase (see Moran & Bar-Anan 2013, Hu, Gawronski & Balas 2017), but other experiments have included relational information in each trial of the REP phase (see Hu *et al.* 2017a: Experiment 3).[42]

The C/R controversy arises from the fact that studies that have satisfied the above conditions have reported seemingly conflicting results. Several studies oft-cited studies report co-occurrence effects (Hu, Gawronski & Balas 2017a: Experiments 1 & 2, Moran & Bar-Anan 2013, Moran & Bar-Anan & Nosek 2015), while others report relation effects (Hu *et al.* 2017a: Experiment 3, Kurdi, Morris, & Kushman 2020, Zannon, De Houwer, Gast, & Smith 2014).[43]

Recall that in Chapter 4, I argued that the prospects of a single-system monist resolution of the C/R controversy are bleak. While every extant account of attitudes predicts that relational information will influence responses to those measures that register one's self-reported likings, accounts diverge in their predictions as to whether relational information should influence performance on the class of indirect tasks depending on whether the account assumes that implicit attitudes are associations or propositions.[44] Associative accounts predict co-occurrence;[45]

[42] While the relevant relation has historically been made explicit by the researchers at some point during the experiment, there is no reason to think that a study could not be designed such that subjects come to reliably infer the relevant CS-US relation(s) despite this information only being implicit. Indeed, pairing a pharmaceutical with images of common side effects of pharmaceuticals such that images of common-side effects follow images of pharmaceuticals, may lead subjects to infer that the pharmaceutical causes the common side-effect, even if this information is at no point made explicit over the course of the experiment and subjects do not report making such inferences.

[43] Sometimes reviews of the relevant literature differ with respect to the kinds of studies that they include. Some make no distinction between studies that document what I have been calling *relation effects* on indirect task performance and studies that document so-called *invalidity effects* (Zanon, De Houwer, & Gast 2014). Other reviews treat these phenomena as separate (Corneille & Stahl 2019). While there are some similarities between the two effects, both in terms of their design and some of the conditions under which they are observed to occur, I am of the mind that, at present, the known similarities are too superficial and the known differences too substantial to warrant treating them as different expressions of the phenomenon. See *6.1.3.2.2* for a more extensive discussion.

[44] Recall that my conforming to the distinction between direct and indirect measures should not be misconstrued as an endorsement of the deeper, theoretical implications that the distinction implies (viz. that, for example, RT measures are any more or less direct measures of attitude than are self-reported measures).

[45] This is a simplification. As we have already noted elsewhere, the APE model allows propositional processes to indirectly influence measures of implicit attitude via the controlled formation of associations. For instance, the belief that (a) hand sanitizer *prevents* infection, might allow me to infer that (b) hand sanitizer is good. So, while the frequent co-occurrence of hand sanitizer and infection could lead to the formation of a HAND SANITIZER-INFECTION association, the frequent tokening of belief (b) would result in the formation of a HAND SANITIZER- GOOD association. So, in this way, relational information could come to have an indirect influence on indirect measures via the formation of an associative structure that reflects the valence implied by the relation. See Gawronski & Bodenhausen 2018 for a more detailed exploration of how the APE model attempts to accommodate various types of EC effect.

propositional models, by contrast, predict indirect measurement outcomes to track the evaluative implications of the relevant CS-US relation. Since neither perspective can account for a full range of findings, I suggested that we turn to developing a format pluralist account of implicit attitudes.

2 An Overview of the MSP Theoretical Solution

This section is devoted to providing an overview of the MSP theoretical approach to the C/R controversy. The section begins with a discussion of several guiding assumptions that, both, help to regiment the discussions in the sections that follow and serve to methodologically differentiate the MSP theoretical approach from its rivals. I then go on to sketch the MSP theoretical resolution to the C/R controversy that is elaborated on and defended in the remainder of this chapter.

2.1 Guiding Assumptions

The format pluralist MSP theoretical resolution of the C/R controversy assumes that each of the two effects may be explicable in terms of the operations of different memory systems. As a starting point, we will assume that the co-occurrence effects are largely driven by learning and retrieval operations carried out by associative memory systems and that relational effects are largely driven by learning and retrieval operations subserved by propositional memory systems. We are by no means beholden to this assumption; it may turn out that, for example, the best model of every hitherto observed relation effect is one that posits only complex interactions of a variety of associative mechanisms.

Furthermore, there is no reason to think that each instance of the co-occurrence effect is the product of the same underlying memory system; the same applies, *mutatis mutandis*, to demonstrations of the relation effect. Put differently, different memory systems may be involved in different demonstrations of the co-occurrence effect. As such, one is not entitled to presume that a memory systems model of, say, Moran and Bar-Anan's (2013) co-occurrence effects will also serve as an accurate model of Hu *et al.*'s co-occurrence effects (2017). Whether this is so is an empirical question.

If these two assumptions hold, it's likely that there is no single model that could account for both effects; that such a model can be given is one of the many false assumptions of the single system monist perspective. The MSP perspective recommends that we tentatively treat each type of effect as involving different memory systems. Thus, with respect to the C/R controversy in particular and the format problem in general, a central question is why experimental protocols that differ only with respect to their seemingly superficial properties sometimes end up recruiting different memory systems during learning and/or retrieval. That is, under what experimental conditions should we expect learning and retrieval to be casually mediated by associative memory systems? And under what experimental conditions should we expect learning and retrieval to be causally mediated by propositional memory systems?

As the MSP theory assigns greater theoretical importance to a broader range features of experimental design and procedure than does single-system monism, a major goal of future MSP driven attitude research should be to identify these features, their boundary conditions, and to give a principled mechanistic account of why these procedural features tend to recruit the memory systems that they do. In any case, our discussion will focus on three such factors that are of general import for MSP theoretical attitude research.

The first two concern task demands imposed on memory systems at the time of attitude formation and/or updating. First, different memory systems selectively attend to specific classes of stimuli. That is, some memory systems are more domain-specific than others. For instance, we have already seen that the amygdala-based Pavlovian Aversive Conditioning (PAC) memory system may be engaged only if CS tokens precede tokens of a highly aversive, biologically salient USs. This implies that the PAC system is of less relative importance to explaining the presence of evaluative conditioning effects involving relatively cold, cognitive semantic-associations (*e.g.,* CAPITALIST-WEALTH). Relatedly, types of conditioning are more often individuated by the types of USs used than by the types of CS used. For instance, a blue square may be used as a CS in either fear-conditioning studies *(i.e.*, those that employ fear inducing USs, such as electric shocks or loud startling noises) or in appetitive conditioning studies (*i.e.*, those that in which the USs are, say, tasty treats). As research shows that different types of conditioning experiments engage different memory systems, we must be sensitive to these differences in interpreting the results of any given conditioning study.

Second, a large body of research shows that the precise temporal ordering of CS-US pairings during conditioning may influence (a) which memory systems are recruited during conditioning, and (b) the general effectiveness of the conditioning procedure. To illustrate, the REP procedure, introduced in the previous section, is instantiated when a CS is repeatedly paired with a US (or multiple USs with the same valence). There are many ways to implement a REP procedure; the researcher may opt to (i) present the CS before the US (*forward conditioning*), (ii) present the US before the CS (*backward conditioning*), or (iii) present the CS simultaneously with the US (*simultaneous conditioning*). But not every memory system is responsive to each of the above conditioning procedures. For instance, a large body of research indicates that the bilateral amygdala prefers CSs that signal threats. On this basis, one would predict the amygdala to be more responsive to forward-conditioning than backward conditioning. There is, in fact, strong empirical support for this prediction (see Lonsdorf 2017). By contrast, the semantic-associative system tracks the relative frequency with which tokens of one concept appear with tokens of another. If so, then we would expect the semantic-associative system to respond to all three of the above conditioning procedures so long as conceptual associations are involved.

The third relevant feature of experimental design is related not to task demands imposed during attitude formation or updating but to task demands imposed in the context of retrieval for measurement. Recall that the MSP theory assumes that measurement tasks generally stand in a one-to-many relation with respect to the mind-brain's memory systems. Given this assumption, there is no reason to assume that the implicit measures that register categorization response-latencies (*e.g.,* the IAT, the evaluative priming task (EPT)) will generally tap the same memory systems as those that measure physiological responses (*e.g.*, eye-blink rates, skin-conductance responses, or cortisol levels). Indeed, as others have demonstrated, even different response latency measures may tap qualitatively different states and processes (see Moran, Bar-Anan & Nosek 2017). Since multiple memory systems may be active during the same conditioning phase, it is not uncommon for researchers in cognitive neuroscience to test the efficacy of the conditioning phase by utilizing a variety of dependent measures. This is a practice that attitude researchers would do well to adopt.

There are at least two methodological upshots of this last assumption worth attending to. If the aim of a study is to understand the effects of a particular form of learning on attitude measurement, then special care must be taken to ensure that the memory systems that are recruited during the learning phase are the same as, or sufficiently overlap with, those tapped during attitude retrieval. If none of the memory systems active during the learning phase are also active during retrieval, then researchers risk erroneously inferring either that (a) no attitude formation had occurred when they encounter null-effects or (2) attitude formation had occurred when, in fact, the processes that drive measurement outcome are non-mnemonic in nature. Moreover, as the conditioning phase of any given co-occurrence study may result in the independent storage of multiple mnemonic effects across a variety of memory systems, two implicit measures administered after a conditioning phase may generate seemingly inconsistent results as a function of the different levels of sensitivity of differing implicit measures to the operations of distinct sets of memory system. Should this occur, researchers operating under the single-system monist framework may erroneously infer that one or both measures are of low quality when, in reality, the different implicit measures are sensitive to the retrieval of different types of implicit attitude.

Due to the fact that attitude researchers laboring under the single-system monist perspective are not as a rule attentive to these factors, studies that appear to control for all of the relevant variables from the single-systems monist perspective seem poorly controlled for from the MSP perspective. As this insight is critical in motivating our approach to developing an MSP resolution to the C/R controversy moving forward, it's worth illustrating how each of the above considerations apply to the topic at hand. In the first set of experiments to have pitted evaluative implications of mere co-occurrence against the evaluative implications of CS-US relations, Moran and Bar-Anan (2013) paired four types alien creature (CS_1, CS_2, CS_3, CS_4) with either a pleasant-sounding melody (US+) or a terrifying human scream (US–).[46] Subjects were told prior to the conditioning phase that

[46] Some may take issue with my claim that Moran and Bar-Anan were the first to demonstrate a co-occurrence effect on implicit evaluations even under conditions in which relational information is provided, pointing instead to Peters and Gawronski's (2011) study that found that feedback as to whether a CS-US contingency is false influenced implicit evaluations as measured when such feedback was provided immediately after CS-US presentation. For instance, Peters and Gawronski (2011) presented subjects with an image of an individual, say, Smith (CS) and paired the individual with evaluative statements (US) (*e.g.,* "Smith is generous"). Subjects were then to decide whether the statement is true or false, after which feedback was immediately provided ("Right!" or "Wrong!"). Importantly, when the feedback invalidated the evaluative statement, subjects were told to *infer that the opposite* about the pictured individual was true (*e.g.,* Smith is selfish). Moran and Bar-Anan (2013) themselves conceive of their study as a follow-up to Peters and Gawronski's study. From a certain perspective, this is understandable. Both studies provide subjects with CS-US contingency information delivered via a REP procedure, and both provide subjects with information about CS-US relations such that the evaluative implications of CS-US co-occurrence are at odds with the evaluative implications of the CS-US relation.

But there's reason to suspect that this similarity is superficial. Whereas Moran and Bar-Anan explicitly state that the aliens that appear at the end of each trial end whatever US they are paired with, Peters and Gawronski tell their subjects to *infer* the opposite CS-US relation whenever feedback describes the initially presented CS-US relation as false.

the type alien that appeared first in the sequence (CS_1,CS_2) would *start* the corresponding US and that the alien that appeared at the end of the sequence (CS_3, CS_4) would *end* the corresponding US. This, as far as we can tell, is a totally novel conditioning protocol that involves both forward conditioning and backward conditioning in a single learning trial. Recall, further, that the associative perspective predicts that subjects would prefer the type of alien that ended US+ over the type alien that ended the US–. The propositional account predicts the opposite preference— that is, subjects should prefer aliens that ended the terrifying human screams over those that ended the pleasant melodies. Experiment 1 employed an IAT; while Experiment 2 employed a sorting paired features task (SPF). Moran and Bar-Anan found a differential effect of relational information on implicit and explicit evaluations; subjects reported liking aliens that ended horrifying human screams more than aliens that ended relaxing melodies, and implicit evaluations revealed a co-occurrence effect.

In the third of a series of three experiments, Hu *et al.* (2017) paired images of novel pharmaceuticals (CS) with images of comparatively mild positive health outcomes (*e.g.*, elderly couple riding bikes, a woman with healthy looking hair; US+) and comparatively mild negative health outcomes (*e.g.*, a mild eye infection, a mild rash; US–). Subjects were told that some of these pharmaceuticals cause health outcomes while others prevent health outcomes. Importantly, this relational information is included in each learning trial, such that the CS would first appear on the screen followed by the appearance of a relation immediately underneath (CAUSES/PREVENTS), after which both the CS and the relation are replaced with the US. The conditioning phase takes the form of a forward-conditioning procedure. Attitudes toward each CS were then measured using an

The hypothesis that these studies probe different types of effect is made all the more plausible once one turns to consider the processes involved in each. In Moran and Bar-Anan's studies, subjects must merely recall that the family of the aliens that appear at the end of the sequence end whatever stimuli they are paired with (*e.g.,* Green aliens end horrifying human screams). In Peters and Gawronski's studies, subjects are first exposed to an evaluative statement about an individual presented, must decide whether the statement is true or false, revise their judgement when feedback presents the statement as false, and lastly integrate the validity information with the content of the evaluative statement so that the evaluative statement implies the opposite of what was said.

EPT. The researchers report that the relational information influenced both implicit and explicit measures, although the effect of relational information on explicit measure performance was larger compared to its effect on EPT performance. More specifically, the authors report that subjects preferred the drugs that prevented negative outcomes over the drugs that prevented positive outcomes.

From the perspective of MSP theory, too many theoretically relevant independent variables have not been adequately controlled for. We simply have no way of knowing the effect of the different types of stimuli, the different types of conditioning procedure, or the different types of implicit measure on the outcome of each study. Consequently, a cross-study comparison of this sort won't allow us to say with any reasonably high degree of confidence which factor, or combination thereof, provoked a co-occurrence effect in Moran and Bar-Anan's studies and a relational effect in Hu *et al.*'s studies.

In light of these interpretative obstacles, how should the MSP theorist proceed? It is important to recognize that the MSP theorist does not take for granted any of the extant single-system monist explanations of the co-occurrence effects. On the associative perspective, evaluative conditioning involves the establishment or strengthening of an associative link between the memory representation of the CS (*e.g.*, an alien) and the memory representation of the US (*e.g.,* a scream). Once this link has been established, the presentation of the CS activates the CS representation in memory, which in turn activates the US representation through a mechanism of activation spread. This activation of the US will then activate the associated response units, leading the subject to respond to presentation of the CS as if the US were presented (Hermans *et al.* 2002). While the MSP theory does postulate a semantic-associative memory system, it posits many other associative memory systems besides and so there is no reason *a priori* to think that this story generalizes to all co-occurrence effects whatsoever. This highlights the need for an MSP theoretical account of co-occurrence effects. Providing such an account is our first task.

By the same token, the MSP theory needs to provide an independently motivated account of the relation effect. While we are assuming with those in the propositional camp that the relation effect is driven by the formation of propositional representations, the nature of the mechanism(s) by which propositional representation formation comes to influence implicit evaluation is very much an open question on MSP theory. For instance, is the proposition with the content DRUG A PREVENTS EYE INFECTIONS retrieved from semantic-propositional memory during EPT performance? Or is there some other route by which proposition formation comes to influence indirect measure response, perhaps by modulating the effects of affective responses to stimuli presented in the EPT?

By giving an independently motivated MSP theoretical account of each effect, we'll have in the process explained why sometimes relational information influences indirect measure performance and sometimes it doesn't. Moreover, we'll have done so in a manner that helps to further integrate social cognitive attitude research with theories of learning in surrounding disciplines.

2.2 The MSP Theoretic Resolution to the C/R Controversy

I'll now briefly sketch the MSP theoretical resolution of the C/R controversy that I will be defending in the following sections. Consistent with the above assumptions, the fact that each finding of a co-occurrence effect is obtained through the use of a different conditioning procedure provides us with some reason to think that each co-occurrence effect may exploit different underlying mechanisms. Since two of the three studies that produce co-occurrence effects employ conditioning procedures that are either entirely novel (in the case of Moran and Bar-Anan (2013)) or poorly understood (Hu *et al.* 2017: Experiment 2), I will focus primarily on accounting for the co-occurrence effect obtained in Hu and colleagues' (2017) Experiment 1, which employs a standard forward-delay conditioning procedure. I will then argue that the co-occurrence effect obtained in this study is driven primarily by amygdala-dependent attitudes that had been established during that study's forward-delay conditioning-procedure, which runs counter to the standard associative account that treats EPT performance as a function of the activation of semantic-associative structures in memory. The defense of this position requires the defense of two further claims—namely, that (1) the amygdala is engaged during the study's conditioning phase and (2) amygdala-dependent representations play a substantive role in EPT performance. After this account is defended, I consider the extent to which this account generalizes to the other two co-occurrence effects under consideration, concluding that more research is needed before any firm conclusions can be reached on this topic.

Whereas the co-occurrence effect observed in Hu and colleagues' Experiment 1 is best understood as a type of conditioning effect, the relation effect obtained in Hu and colleagues' Experiment 3 is best understood as involving a form of instructed learning (see Koban *et al.* 2017). Briefly, instructed learning occurs when subjects encode an explicitly instructed CS-US relation. Instruction, on this view, provokes the formation of a propositionally structured representation of each CS-US relationship (*e.g.*, DRUG A PREVENTS EYE INFECTIONS). Recent work that investigates the differential effects of instruction and conditioning on learning have shown that the memory systems that update in response to instructions differ from those that update in response to repeated CS-US co-occurrence (*i.e.,* conditioning) (Atlas 2019, *Koban et al.* 2017, Koban *et al.* 2019). Crucially, this research shows that while the left amygdala updates in responses to instructions about the CS-US relationship, the right amygdala updates only in response to CS-US co-occurrence (Atlas 2019).

On the view that I favor, the relation effect is primarily driven by left amygdala-dependent instructed learning representations.

3 The Co-Occurrence Effect

The structure of this section is as follows. Our discussion begins with the co-occurrence effect obtained Hu, Gawronski, and Balas' (2017) Experiment 1 (hereafter, HGB_{ex1}). Once we've developed a rough account of the states and processes that drive evaluative responses on the EPT, we'll then shift our attention to the other two co-occurrence effects obtained by HGB's Experiment 2 (HGB_{ex2}) and Moran and Bar-Anan's (2013) experiments.

3.1 HGB: Experiment 1

In HGB_{ex1}, subjects are explicitly informed of the relevant CS-US relation during the task instructions that immediately precede the attitude acquisition phase. I've reproduced the task instructions given to each subject below:

> Your task is to think of the image pairs, such that the pharmaceutical product causes [/prevents] what is displayed in the following image. For example, if a product is paired with a positive image, you should think of the product as causing [/preventing] the positive outcome displayed in the image. Conversely, if a product is paired with a negative image, you should think of the product as causing [preventing] the negative outcome displayed in the image. (Hu *et al.* 2017a: 20)

Here we find that the researchers have presented the relational information in what amounts to a quantified two-place relational formula (*viz.* (x)(y)(if x is presented at t_1 and y is presented at t_2, then x causes (/prevents) y)). We will follow the terminological convention established by HGB in referring this a *rule*. Importantly, the manipulation of relational information is between-subjects; half of subjects are told that relevant drug *causes* health outcomes are presented, while the other half are told that the relevant drug *prevents* health outcomes.

So much for task instructions. How is the REP learning phase structured? Subjects are exposed to five CSs in total. One CS is neutral (hereafter, referred to as CS–) and is therefore never paired with a corresponding image (*i.e.*, it is presented without any sort of reinforcement). Of the remaining four CSs, two (*e.g.*, CS_1: Drug A; CS_2: Drug B) are systematically paired with a particular positive US (US_1+: elderly couple riding bikes; US_2+: woman waving radiant looking hair) and each of other two CSs (CS_3: Drug X; CS_4: Drug Y) are systematically paired with a particular negative US (US_3–: eczema outbreak; US_4–: eye infection). The structure of each trial of the REP phase is as follows: CSs appear in the center of the computer screen for 1,000 ms followed by a 1,000 ms presentation of the US in the same location. This conditioning procedure is commonly known as a *forward-delay procedure*. A delay procedure is one in which US onset is delayed relative to the CS onset without an extended temporal gap between CS and US onset. (Procedures that involve an extended gap between CS termination and US onset are known as *trace conditioning* procedures.) As the intertrial interval (ITI) lasts for 2,000 ms (during which, nothing appears on screen) there is low probability of inadvertent backward conditioning (whereby the CS used in, say, trial two comes to be associated with the US that appears at the back end of trial 1).

The relevant findings are as follows. HGB found that relational information provided during task instructions influenced explicit responses to the conditioned stimuli; however, relational information had no effect on EPT performance. Instead, HGB found a main effect of valence on

EPT performance, meaning that subjects were able to more quickly and accurately identify positive words after being exposed to a CS prime that had previous been repeatedly paired with USs+ relative to CS primes that had previously been repeatedly paired with a USs–. Having said that, it is worth taking a closer look at the patterns of results that realize the above co-occurrence effect. The authors report that subjects on average responded much more unfavorably to CSs that had been systematically paired with USs– relative to baseline CSs and responded only slightly more favorably to CSs that had been repeatedly paired with USs+ relative to baseline. Thus, if the EPT performance tells us anything about the sort of learning that had occurred, these patterns suggest that conditioning effects were stronger for CSs that were paired with aversive USs than for CSs that were paired with positive USs.

Having now reviewed the relevant findings, there are two phenomena that must be accounted for: (a) the effect of relational information on direct responses only, and (b) the nature of this particular co-occurrence effect. Of the two, the MSP theoretical account of the former is comparatively straightforward. The differential effect of relational information on implicit and explicit task performance is most plausibly explained by the fact that subjects are able to ignore the rule until they are presented with the explicit measure, during which they are asked to (i) report how positive/negative they find the pharmaceutical with response options ranging from 1 (*very negative*) to 7 (*very positive*) and (ii) when asked how good/bad each pharmaceutical is, with response options ranging from 1 (*very bad*) to 7 (*very good*). This explanation, note, is entirely consistent (if not identical) to explanation that Hu and colleagues endorse.

But how should the MSP theorist account for the observed co-occurrence effect? While it may be the case that both the semantic-associative memory system subserves a form of conditioning and its operations influence EPT performance, I propose that we think of this particular co-occurrence effect as largely being driven by the retrieval of negatively valenced amygdala-dependent

CS representations. It is possible that during the conditioning phase the Pavlovian Aversive Conditioning (PAC) system is involved in the preferential updating of those CS representations that immediately precede presentations of USs– in accordance with backward-looking, model-free algorithms (see Atlas 2019). On this view, when these CSs appear as primes (hereafter, *prime*$_{neg}$) during the EPT, the activation of these amygdala-dependent CS representations either (a) generates an expectation that a negatively valenced item will appear on screen, or (b) will facilitate the processing of valence-congruent stimuli. If either (or both) of these possibilities are true, then negative targets follow the presentation of a prime$_{neg}$, response times (RTs) to targets are lowered.

Two questions about this account immediately arise. First, what reason do we have to expect amygdala engagement during the conditioning phase? Second, what independent evidence do we have that amygdala-dependent representations are capable of producing the type of priming effect measured by the EPT? We'll address each in turn.

The primary reason for suspecting amygdala involvement during the REP phase is that the experiment has many of the trappings of more traditional aversive conditioning experiments. For instance, the negatively conditioned CSs are systematically and repeatedly paired with negative biologically relevant outcomes, which the amygdala is particularly sensitive to (Atlas & Phelps 2018). To be sure, there are important differences between the USs– used in HGB$_{ex1}$ and those used in fear-conditioning or other aversive conditioning studies. For example, we can assume that the USs– used in this study are not as intense (affectively speaking) as stimuli used in aversive conditioning experiments generally or fear conditioning experiments in particular (see Lonsdorf 2017). Thus, it is an open question as to whether the mildly negative health outcomes used as USs– in this study are sufficiently negative to trigger amygdala involvement. Additional research is needed to address this concern. With that said, USs of low affective intensity and of less biological significance have effectively been used to trigger amygdala involvement in past research (*ibid.*).

Relatedly, the most effective aversive conditioning studies frequently employ a variant of the forward-delay conditioning procedure (Lonsdorf 2017). As will become relevant when we consider the other empirical demonstrations of the co-occurrence effect, amygdala-dependent aversive conditioning is more effective in the sense that it tends to produce stronger conditioned responses (CR) in fewer trials when conditioning is facilitated through the use of a forward-delay conditioning procedure than either simultaneous or backward conditioning procedures (*ibid.*). Moreover, animal research has repeatedly demonstrated that amygdala-dependent aversive conditioning is preserved in animals with hippocampal lesions when a forward-delay procedure but not a forward-trace conditioning procedure is used (Manns, Clark, & Squire 2000). As HGB_{ex1} exploits a forward-delay procedure, conditions are suitable for effective amygdala-based conditioning.

Let's assume, then, the PAC system (among others) is involved in learning. If, in this context, a priming effect on the EPT is the target CR that we want presentation of the CS to elicit, it doesn't follow from the above assumption that CS representations formed via amygdala-dependent learning will provoke this CR. Though there is strong evidence that the amygdala is involved in affective priming when facial expressions are both used as CSs and targets even under conditions in which face-primes are presented subliminally (Yang *et al.* 2012), the EPT is not typically used as a measure of conditioned responses in aversive conditioning studies. So, it is possible that even highly effective amygdala-dependent aversive conditioning has no effect on response times when the target is a word (as opposed to a fearful face). Thus, if amygdala-dependent representations and processing help us explain the co-occurrence effect, we would like some additional evidential support that this system, perhaps in conjunction with others, manifest the kinds of responses registered by the EPT.

Earlier I proposed two hypotheses of how amygdala-dependent representations and processes might produce affective priming effects. The first hypothesis is that when a given CS is repeatedly paired with a biologically aversive US, the model-free processing implemented by the

PAC produce an expectancy that a negative stimulus will appear following the presentation of the CS. This is analogous to how a subject who is repeatedly shocked after blue square presentations might come to anticipate an electric shock (as indicated by, say, increased skin conductance responses (SCRs)) given the presentation of a blue square. Call this the *expectancy hypothesis* of evaluative priming. Given that subjects are instructed to rapidly hit the corresponding key as soon as they recognize a target word as being positive or negative without making too many mistakes when completing the EPT, the expectation of a negative stimulus given the presentation of a CS-prime may result in a pre-activation of the motor routine associated the pressing of the "negative" key. This expectancy driven behavioral response may occur even if the target word is not fully processed (*cf.* Klinger *et al.* 2000, De Houwer & Randell 2002), leading to faster and less error-prone responses on prime-congruent compared to incongruent trials. Following established terminological convention, let's call the pre-activation of a motor command in response to CS-prime presentation *response priming* (Eder *et al.* 2011).

The other hypothesis is as follows: in the context of the EPT, the presentation of an aversively conditioned CS would reduce the time it takes to respond to a negatively valenced target because the CS-prime facilitates the processing of the target word. This is analogous to an account of conceptual or semantic priming according to which a prime word facilitates the recognition of the target word's meaning through the spreading of activation of associated concepts in a semantic-associative network (*e.g.,* Bargh *et al.* 1996, Fazio 2001, Spruyt *et al.* 2007). On this hypothesis, the facilitation reduction of response times comes not from an expectancy of a negative stimulus, but from the fact that the amygdala CS representation itself elicits a negative affective response and this affective response to the CS-prime facilitates the processing of affectively congruent targets leading to faster classifications when the valence of the target is congruent with the prime than when the valence of the target is incongruent with the prime (*cf.* Baeyens 1998). Call this the *shared valence* hypothesis of evaluative priming. Because subjects must process the target sufficiently before issuing the corresponding behavioral response, this view predicts that the onset of the corresponding motor response should tend to occur only after subjects have had sufficient time to process the target.

An additional comment about the shared valence hypothesis is worth making. The plausibility account depends on the plausibility of the claim that the amygdala is critically involved in the encoding and/or processing of a word's valence; if not, there is no reason to expect an increase in the speed it takes to process a word (as opposed to an emotion-laden face) during EPT performance. A large body of research does indicate that the left amygdala in particular is involved in both the encoding of word valence and its subsequent processing (Hamman & Mao 2001, Richardson, Strange & Dolan 2004, Phelps 2004).

A particularly striking demonstration of the left amygdala's role here involves the case of a teenager with severe damage to the left amygdala. While this individual did not differ from healthy controls with respect to judgements about whether the words she was asked to memorize were positive or negative, her performance differed significantly relative to healthy controls with respect to both judgements as to how arousing the positive and negative words were (*e.g.,* she treated them as marginally more arousing than the neutral words she was asked to memorize) as well as in her recognition performance for emotional words compared to controls (Claire *et al.* 2016). While control subjects failed to recognize just 2.5% of the emotional words, the amygdala patient failed to recognize 26.6% of the emotional words (for neutral words, she failed to recognize only 6%) (*ibid*). The researchers also observe that this individual can identify positive or negative words as such without any accompanying affective experience. It is possible that this individual learns which words are positive or negative through social learning (Olsson, Knapska & Lindström 2020). Future empirical research should aim to investigate the extent to which individuals with left amygdala damage would perform similarly to controls on the EPT. If such individuals do not perform as well as controls, this would provide some empirical evidence in favor of the *shared valence* hypothesis.

It is worth noting that the shared valence hypothesis and the expectancy hypothesis are not mutually exclusive; both types of processing may help to shape EPT performance when prior conditioning is aversive. Hermanns and colleagues (2002) demonstrated that a single aversive conditioning phase can give rise to both expectancies that a negative event will follow the presentation of a CS prime *and* a change in the CS representation's valence. Indeed, it may be that negative expectancy provoked by a CS presentation is the basis of the subject's negatively affective reaction to the CS itself. If Smith dislikes rain, then Smith comes to have a negative affective response to the presence of storm clouds overhead because storm clouds signal rain. There is no reason, then, to think that both conditioning effects could not shape EPT performance. Indeed, several studies electroencephalogram (EEG)) studies have produced evidence that EPT performance reflects the combined output of both response priming and shared valence priming (Eder *et al.* 2012, Zhang *et al.* 2006). With that said, additional empirical research is needed before we can confidently say whether evaluative priming arises out of expectancies, shared valence, or some combination thereof. Whichever account turns out tell the correct story, what is important is that the MSP theory has a plausible story to tell.

We are now well-positioned to inquire about the extent to which the account just proposed generalizes to other demonstrations of co-occurrence effects in the literature. Due, however, to the quirks of the procedures used to obtain the relevant co-occurrence effects, we'll have relatively little to say pending further empirical investigation.

3.2 HGB$_{ex2}$

First, let's turn our attention to HGB$_{ex2}$. The procedure of this study is identical to that of HGB$_{ex1}$ in all but one respect; whereas HGB$_{ex1}$ employed a forward-delay conditioning procedure, HGB$_{ex2}$ employed a simultaneous conditioning procedure. Though HGB once again report a differential effect of relational information on implicit and explicit evaluations such that relational information impacted explicit but not implicit evaluations, the patterns of results that realize the co-occurrence effects are markedly different compared to those obtained in HGB$_{ex1}$: relative to neutral CS primes, subjects responded more favorably *both* to CSs that had been paired with USs+ and CSs that had been paired with USs–, though the priming effect was more pronounced for the former than the latter. In other words, the simultaneous conditioning procedure involving the same set of stimuli produced faster than baseline responses to positive target words regardless of previous pairings and regardless of relation. So, while the authors report that "CSs that had been paired with a positive US elicited more favorable responses than CSs that had been paired with a negative US" (Hu *et al.* 2017: 25), it is misleading to frame this as an unambiguous demonstration of a co-occurrence effect given that CSs that had been paired with USs– elicited more favorable responses than baseline.

The procedural change together with this unexpected pattern of priming effects makes these results especially difficult to interpret. The simultaneous conditioning procedure is not regularly used in evaluative conditioning research, so the mechanisms it exploits are poorly understood relative to forward-conditioning. Moreover, a large body of research finds that forward conditioning procedures are overwhelmingly more likely to generate the target conditioned responses than either simultaneous conditioning or backward conditioning (Malone 2002). When the simultaneous conditioning procedure is used in attitude research, it tends to be most successful in provoking the *implicit misattribution effect*, which occurs when subjects mistakenly infer that the CS, as opposed to the US it is simultaneously paired with, is an important causal source of their occurrent affective experience (see Jones, Olson, & Fazio 2010). But this does not appear to explain why HGB's simultaneous conditioning procedure causes subjects to respond more favorably to those CSs that had previously been paired with USs–.

I am inclined to think that this is one instance in which the EPT effects observed are not best explained merely by appeal to the retrieval of attitudinal states or the engagement of attitude-related mnemonic processing. To understand why, consider what happens when we try to explain the results by appeal to either the expectancy hypothesis or the shared-valence hypothesis. It is not easy to see how the repeated simultaneous presentation and termination of CS-US pairs would lead subjects to anticipate the US at t_2 given its presentation at C_1, as this temporal structure is not mirrored in the conditioning procedure. Moreover, even if simultaneous presentation could generate US expectancies, it is hard to explain why repeated CS-US– presentation would lead subjects to expect positively valenced stimuli upon presentation of the relevant CS. Similarly, the shared valence account fails to explain the data as it would predict that, insofar as evaluative conditioning occurs at all, the valence of the US should transfer to the CS that it is presented with. This prediction is inconsistent with the actually observed priming effects.

One might complain that the difficulty of accounting for the EPT effects demonstrated in this particular study is a strike against the MSP theory. But it's important to keep in mind that it fares no worse than its rivals here. The APE model, for instance, would predict that repeated CS-US pairings would result in a transfer of valence from the US representation stored in memory to the CS representation. But this doesn't address why subjects were on average faster to identify positive words after the presentation of a CS-prime that had been repeatedly paired with USs–.. The propositional theorist also has not straightforward explanation of these results. We know from the results that a relation effect was observed for explicit attitudes but not implicit attitudes. So, if only the activation of propositional structures in memory drives priming effects, subjects would need to form beliefs about CS-US relations that do not reflect the relation given in the instructions. What relation would this have been? Presumably, it would have been propositions of the sort DRUG A IS PAIRED WITH EYE INFECTIONS (or DRUG A CO-OCCURS WITH EYE INFECTIONS). But there are two glaring problems with this proposal. First, it is far from obvious that the evaluative implications of this proposition casts Drug A in a positive light. Second, and relatedly, the mechanism by which the activation of this— or any— propositional structure impacts the time it takes to respond to positive or negative target words is, it seems to me, hopelessly opaque.

3.3 Moran & Bar-Anan (2013)

Lastly, let's turn our attention to the first empirical demonstrations of the co-occurrence effect as obtained by Moran and Bar-Anan (2013) across two studies. The extent to which the partial account of the co-occurrence effects demonstrated in HGB_{ex1} generalizes to these findings is, again unclear, suggesting that further empirical research is needed to address this issue. In order to understand why, we will have to briefly review the relevant details of Moran and Bar-Anan's procedure and findings.

Moran and Bar-Anan paired four types alien creature (CS_1, CS_2, CS_3, CS_4), differentiated by color, with either a pleasant-sounding melody (US+) or a horrifying human scream (US–). Here's how the authors describe the procedure:

> The learning included 10 melody trials and 10 scream trials, ordered randomly and separated by a soft ticking sound that played for 10-15 seconds. Each trial began with a presentation of the "starting" creature, appearing in silence for 500 ms. Next,

> the auditory stimulus began, playing for a randomly selected duration of 10–30 seconds. The "starting" creature remained on the screen for the first two seconds of the playback. Then the screen remained blank until the "ending" creature appeared for the last two seconds of the auditory stimulus's playback and remained on the screen for another 50 ms of silence. […] We told participants in advance that the creatures started and ended the stimulus. We instructed participants to learn which family performed each of the four actions for a later memory test. (Moran & Bar-Anan 2013: 746)

In the first study, subjects then completed two versions of the IAT— one that measured "automatic preference" for the groups of aliens that served as starters of the relevant stimulus (hereafter, *starters*) and one that measured automatic preferences for the aliens that appeared as enders of the relevant stimulus (hereafter, enders). For the purposes of testing the effect of relational information on implicit attitudes, the enders-IAT is the critical measure— we'll address this point momentarily. The second experiment differed only in respect to the implicit measure used; in that experiment all subjects completed a sorted-paired-features (SPF) task. Upon completion of the respective implicit measure, all subjects were then asked to self-report the degree to which they liked each alien group on a 9-point scale (1= *Dislike Strongly*, 9= *Like Strongly*). Finally, subjects completed a memory recognition task in which each subject was asked to identify which alien performed which action.

Given this setup, a co-occurrence effect would be observed if subjects prefer the aliens that *ended* the horrifying human screams (US–) to aliens the *ended* relaxing melodies (US+). By contrast, a relation effect would be observed if subjects preferred the aliens that ended horrifying human screams over aliens that ended relaxing melodies (because people should prefer things that end negative stimuli over things that end positive stimuli). This explains the above claim that the enders-

IAT is the critical measure. Moran and Bar-Anan report a co-occurrence effect in both studies. Consistent with the results that HGB obtained their co-occurrence studies, Moran and Bar-Anan found that relational information influenced only explicit responses; subjects reported that they prefer the aliens that ended screams over those that ended relaxing melodies.

That a co-occurrence effect was observed under these circumstances is somewhat surprising as the effect was produced by way of backward-delay conditioning (*i.e.*, the onset of each CS was delayed upon the onset of the corresponding US). As stated in our discussion of HGB_{ex2}'s demonstration of a co-occurrence effect, research on conditioning consistently shows that backward conditioning is generally less effective than forward conditioning in generating the relevant class of CRs. But this is not to say that backwards delay conditioning procedures never produce aversive or reward conditioning effects; the efficacy of backward-conditioning procedures appears to be a function of the US intensity (Burkhardt 1980) and the number of US-CS pairings (Heath 1976). It may be the case that the horrifying screams were of sufficient intensity and/or the number of pairings sufficiently high so as to bring about changes in the CSs that co-occurred with human screams.

Having said that, one might worry that the results obtained by Moran and Bar-Anan do not sit well with past research on the differential effects of backwards-conditioning on implicit and explicit CS evaluations. For instance, at least one study has demonstrated that when a CS is presented immediately upon the offset of a pain-inducing US, subjects implicitly evaluate the CS favorably (as measured by startle responses) but explicitly evaluate the CSs negatively (Andreatta *et al.* 2010). In other words, when a backward-conditioning procedure is used, co-occurrence information influences explicit evaluations more than it does implicit evaluations, which is the inverse of what Moran and Bar-Anan found. But these results are not obviously inconsistent with Moran and Bar-Anan's findings.

Two key procedural differences between the two sets of studies might explain the different patterns of results. First, in Moran and Bar-Anan's findings, the presentation of the fear-inducing stimuli temporally overlaps with the presentation of the CS; there is no such temporal overlap in Andreatta and colleagues' experiments. This difference could help to explain why implicit evaluations of CSs that had systematically been paired with aversive stimuli were favorable in the latter but not the former studies. The idea here is that the CS comes to signal safety when the CS is only ever presented upon the termination of the US presentation. But when there is sufficient temporal overlap in CS-US presentation, the CS is no longer an unambiguous safety cue. Further, This overlap may promote amygdala dependent CS-US– binding. A second critical difference is this: at no point does Andreatta *et al.* explicitly instruct their subjects as to how to think of the CS-US relationship, leading subjects to rely on co-occurrence information when reporting CS evaluations. Because Moran and Bar-Anan do instruct their subjects that the aliens that appear last in the sequence *end* the corresponding US, their subjects are able to retrieve this information during explicit measurement. This procedural difference, then, may have a role in explaining why it is that co-occurrence information exerted a greater influence on explicit attitude measurement in Andreatta and colleagues' experiments relative to Moran and Bar-Anan's experiments.

Even if we were to suppose that the simultaneous conditioning procedure was effective in generating evaluative conditioning effects, we would still need an explanation of how these learning effects manifest IAT effects (and SPF effects). While each measure is a response latency measure, there is no reason to believe that the mechanisms that drive EPT effects are the same as those that drive IAT (or SPF effects), as evidenced by low-convergent validity between the two measures and that, *ceteris paribus*, experimental manipulations that influence IAT performance do not always influence EPT performance (Van Dessel *et al.* 2017, Moran, Bar-Anan, & Nosek 2017). This reinforces the assumption starter earlier that that it would be a mistake to assume, *sans* further empirical study, that all co-occurrence effects share a common underlying mechanism. Given the lack of relevant empirical research and restrictions of space, an MSP theoretical account of the type of learning that occurs in these studies and of how learning influences IAT and SPF performance, I must leave these as issues for future research.

4 The Relation Effect

Let's now shift our attention to what is arguably the most influential empirical demonstration of a relation effect— viz. those reported in HGB_{ex3}. The structure of this section is as follows. After briefly reviewing the relevant procedural elements and findings, I propose an MSP theoretical account of the relation effect. The remainder of this section is devoted to its defense.

The experimental procedure used in HGB_{ex3} differs from that used in HGB_{ex1} in a number of respects relevant to the ensuing discussion. First, in a departure from the previous two studies, Hu and colleagues manipulated relational information as a within-subjects factor, meaning that each subject was shown both causal and preventative CSs. Second, Hu and colleagues made a substantive change to task instructions:

> The instructions were similar to the ones in Experiments 1 and 2, the only difference being that participants received the relational information on a trial-by-trial basis during the presentation of the CS-US pairings rather than in the instructions. (Hu *et al.* 2017a: 27).

The passage above makes direct reference to a third key difference: subjects are explicitly told how to think of the relationship that obtains between each CS and each US during the conditioning phase. Fourth, the researchers doubled the number of unique CS-US pairs that subjects had to track, from four in the previous two experiments to eight (four CSs were systematically paired with a US+; four CS were systematically paired with a US–) in the present study, bringing the total number of CSs to nine (including the neutral, unpaired CS).[47]

The modified learning phase is as follows. On each trial, participants were presented with one of the CSs in the center of the screen for 1,500 ms. Five-hundred milliseconds after the onset of the CS, either the term 'causes' or 'prevents' appears just below the CS. Both the CS and the relational term remain on screen for 1,000 ms, at which point both are replaced by the US. The US remained in the center of the screen for 1,000 ms. For this conditioning procedure, HGB used an inter-trial interval (during which nothing appeared on screen) of 1,000 ms. Structurally, this procedure looks like a forward-delay conditioning procedure. Hu and colleagues found that the relational information moderated both implicit and explicit measure performance, though the effect was weaker for implicit compared to explicit evaluations. Importantly, when causal CSs were presented as primes, no statistically significant priming effect was observed. However, for preventative CSs, subjects responded on average more favorably to CSs that had been paired with USs– than to CSs that had been repeatedly paired with USs+.

[47] Here's another difference, though its importance is debatable. Hu and colleagues did *not* include the recognition memory task that had been included in the previous two studies. The decision to omit a memory assessment from this study (the rationale for which is never provided) may strike some as puzzling, given that this procedure is far more demanding on memory than the previous two. In this experiment, not only are subjects tasked with learning twice as many unique CS-US pairings as subjects in the previous two studies, these subjects also must learn the precise relation that obtains between each CS-US pairing. Recall that in the previous two studies, relational information was manipulated as a between-subjects factor, so that subjects could always be sure that the same relation obtained between each CS-US pair. It is also worth noting that even for the less demanding studies, subjects showed strong but imperfect recollection for each CS-US pair (the average performance across both studies hovered at around 90%, though interaction effects on recollection were observed for both valence and relation). Moreover, as contingency awareness is known to be positively correlated with the efficacy of evaluative conditioning, this omission seems especially glaring.

The reason why the importance of this difference is debatable for the purposes at hand is that subjects' explicit measure performance showed a strong relation effect, suggesting that subjects were able to accurately recall to which CS bore which relation to specific USs. Whether this effect could be explained on the assumption that subjects had poor memory is unclear.

In broad brush strokes, the explanation of these findings that I favor is as follows: by presenting the relational information alongside CS-US co-occurrence information trial-by-trial, the procedure shifts from a standard evaluative conditioning procedure to a type of instruction procedure. In line with standard instruction studies, subjects were explicitly told how to think of the relation that obtains between each CS and US. Crucially, there is strong evidence that the memory systems that subserve conditioning differ from those that subserve instructed CS-US contingency learning. Moreover, whereas the former frequently implement model-free learning algorithms, the latter are best modeled as model-based learning systems (see Atlas 2019). For these reasons, it is reasonable to suppose that relation effect on EPT performance observed in this study is driven primarily by the memory systems that subserve instructed learning as opposed to conditioned learning. If this explanation holds, then we have on hand a plausible format pluralist resolution to the C/R controversy, as the relation effect is driven primarily by propositional memory systems and the co-occurrence effect is driven by amygdala-based associative states and processes. The remainder of this section is devoted to elaborating on and defending this proposal.

When I claim that the present study is best understood as employing an instruction procedure as opposed to a conditioning procedure, what exactly does this claim amount to? To illustrate the difference, contrast a fear learning study in which subjects are merely instructed that presentations of blue squares (CS) will be accompanied by electric shocks (US–) with a study in which subjects come to learn that blue squares accompany electric shocks through repeated pairings of blue squares with electric shocks. Call the former *instruction only* (IO) studies and the latter *conditioning* studies.

This distinction is absolutely crucial for a pair of reasons. First, large body of research shows that instruction-based fear learning is as effective as merely conditioned fear-learning with respect to provoking fear-related CRs (Hugdahl 1978, Hugdahl, Ohman 1977, Chase *et al.* 2015). Analogous effects have been demonstrated across multiple learning domains, including instrumental reward learning, pain-learning, and evaluative learning (see Koban *et al.* 2017 for a review). We discuss some of these findings in greater detail below. Second, an equally compelling body of research demonstrates that the memory systems that respond to mere instruction about how a CS-US pair are related differ from the memory systems update in response to actually observed CS-US pairings. Dissociations between the systems that are sensitive to instruction and those that are sensitive to actually observed CS-US pairings have been reported across various learning domains. For instance, in the domain of amygdala-dependent aversive learning, it has been shown that IO-fear inducing procedures influence left but not right amygdala activation (Phelps, O'Connor, Gatenby, *et al.* 2001, Butler *et al.* 2007, Funayama, Grillon, Davis, Phelps 2001). By contrast, human and animal research shows that the right amygdala updates only in response to CS-US pairings (Atlas, Doll, Li, Daw, Phelps 2016, Braem, De Houwer, Demanet, Yuen, Kalisch, Brass 2017).

Another important source of evidence concerning such dissociations comes from recent research on *instructed reversals* in the domain of fear learning (see Atlas 2019). In such studies, subjects are exposed to two CSs, say, a green circle that is occasionally accompanied by an electric shock and a blue square that is never accompanied by an electric shock. After a number of conditioning trials, subjects are then told that the relationship between electric shocks and the CSs are now reversed (*i.e.,* blue squares accompany shocks and green circles do not). Atlas' (2019) description of her and her colleagues' findings is as follows:

> Individuals underwent aversive reversal-learning in a combined within-subject and between subject design and we measured responses to unreinforced CSs during fMRI scanning. Differential SCRs reversed immediately upon instruction in an Instructed Group, replicated previous work. We also observed immediate reversals

> in the striatum, insula, ACC [anterior cingulate cortex], and the VMPFC/OFC [ventral medial prefrontal cortex/ orbital frontal cortex], and the strength of reversal learning in striatum and VMPFC correlated with dorsolateral PFC (DLPFC) activation during the presentation of verbal instructions. […] Most notably, however, the right amygdala did not update immediately upon instruction. Instead, differential amygdala responses reversed after actual reinforcement, indicating that it was more responsive to experiential, rather than instructed, learning. (Atlas 2019: 123)

It is worth emphasizing that the finding that the VMPFC and the DLPFC are critically involved in the representation of instructions during this task replicates and extends previous findings demonstrating the importance of these two regions in a wide range of other contexts that involve instructed learning (see Koban *et al.* 2017).

Atlas's findings are consistent with a large body of research that demonstrates that even though the brain networks that respond to instruction differ from those involved in conditioning, the memory systems that subserve instructed learning frequently exert powerful top-down effects on the systems that are most responsive to experiential forms of learning (see Koban *et al.* 2017 for a review). In the first empirical demonstration of the effects of instructions on fear learning, Cook and Harris (1937) measured SCRs in response to green light presentations (a) after subjects were told that a green light would be followed by a shock, (b) after these initial instructions were reinforced, and once more (b) after participants were told that shocks would no longer be delivered. The researchers found that SCRs increased in response to instructions, didn't change in response to conditioned reinforcement, and extinguished immediately after subjects were told that there would be no more shocks. Follow-up studies replicated these early findings (Hugdahl 1978, Hugdahl, Ohman 1977, Chase *et al.* 2015). These findings concerning the top-down modulation of networks involved in instructed learning on those involved in conditioning parallel results obtained in adjacent fields. For instance, human research on instructed reward learning, shows that the PFC maintains instructions and modulates learning in the striatum, either by abolishing prediction errors when the instructions are accurate (Li, Delgado, & Phelps 2011), or by biasing learning toward instructions when instructions and feedback conflict (Doll, Jacobs, Sanfey, & Frank 2009). Similar effects have been observed in research on the placebo effect and the nocebo effect (see Koban *et al.* 2017 for an extensive review of these and various other top-down effects).

Suppose one is willing to grant both that instructed contingency learning occurs in HGB$_{ex3}$ as a consequence of their decision to embed relational information on each CS-US presentation, and that the memory system that subserves instructed learning in this case differs from the memory system(s) that drives learning in HGB's previous two studies. One might nevertheless deny that these assumptions help us to explain the relation effect, on the grounds that I have provided no evidence that instructed CS-US contingency learning manifest evaluative priming effects. The idea is that if instructed CS-US contingency learning generally have no effect on EPT performance, then there is good reason to think that instructed CS-US learning does not drive EPT performance here. If so, the objection continues, then the relational effect obtained in this context cannot be understood as involving instructed learning.

The available evidence on the effects of instruction on EPT performance is mixed. One study that investigated the so-called "instructed mere-exposure effect" on different indirect measures attitude found that mere instructions that a novel non-word would frequently occur over the course the experiment influenced performance on the IAT and the *affective misattribution procedure* (AMP) but not the EPT (Van Dessel, Martens, Smith, De Houwer 2017). Other instruction effects on EPT performance in attitude research have been demonstrated (Moran, Bar-Anan, & Nosek 2017, Van Dessel, De Houwer, Gast, Smith 2015). Some have suggested that these inconsistent patterns are likely a function of the EPT's relatively low reliability (Bar-Anan & Nosek 2017). While it would be wise for researchers to take its relatively low reliability into account when designing studies and interpreting results, there is no reason why instruction should give rise to EPT effects in every case even on the assumption that instruction in a particular instance provoked a change in attitude. Whatever the reason for the mixed results, we cannot dismiss the view on offer solely on the grounds that sometimes instruction does not influence EPT performance.[48]

But perhaps there is another objection lurking here. Even if one were to grant that instructed learning *can* influence EPT performance, one might nevertheless insist that I do not have a plausible story to tell about how the activation of instruction-induced CS attitudes come to facilitate responses to similarly valenced words on an EPT. Without such a story, the objection goes, the view ought to be rejected.

First, even if we grant that central premise of the objection, it must be stressed that the propositional perspective (which, recall, is the only other view that is consistent with the relation effect) likewise has no plausible story to tell about how the retrieval of a propositionally structured representation with the content DRUG A PREVENTS EYE INFECTIONS facilitates faster responses to positive related words. Moreover, if there is a plausible story for propositional theorist to tell here, there is no reason, in principle, that the MSP theorist couldn't produce a similar story. So, even if the above objection presses on a genuine weakness of the account, it is a weakness shared by MSP theory's chief rivals.

[48] There's good reason to think that the MSP theory sheds some light on these seemingly inconsistent patterns of effects. As Van Dessel and colleagues acknowledge, neither regular mere exposure to non-words nor merely instructed mere exposure to non-words produces an EPT effect when non-words appear as primes (Van Dessel *et al.* 2017: 25). If the amygdala typically plays an important role in mediating EPT performance, then there is no reason to expect any form of mere exposure would generate an affective priming effect, as there is no reason to think that mere exposure to non-words would induce amygdala-dependent changes. By contrast, Van Dessel and colleagues report that mere instructions that a CS should be approached or avoided do produce corresponding affective priming effects (*e.g.,* subjects are faster to classify target words as positive when they follow the presentation of a CS that subjects were instructed to approach). But we should expect amygdala involvement here.

Having said that, there is no reason to think that the MSP theorist cannot eventually fully meet this challenge. The caveat here is justified: in order to give a full accounting of how instruction induced CS attitudes produce priming effects when the corresponding CS presented in the context of an EPT, one would need to know (a) whether priming effects are mediated by response priming or shared valence (see §3), whether the learning procedure employed in $\mathrm{HGB}_{\mathrm{ex3}}$ also results in the formation of CS attitudes via repeated CS-US pairings, and (b) how the memory systems responsible for instruction effects interact with the mechanisms responsible for conditioning effects during EPT performance. And while I think we are in a good enough position to make a reasonable pass at addressing (a), additional empirical research is needed before we can adequately address both (b) and (c). (After addressing (a), I will go on to recommend some studies that could help shed light on (b) and (c).)

Here's one possible explanation of how the memory systems that subserve instructed learning drive the relation effect that Hu and colleagues observe. On each trial of the learning phase subjects are explicitly informed as to how to think of the relationship between the CS and the US. Consequently, we can think of this study as containing eight sets of instructions (one set for each unique CS-Relation-US triad), such that each instruction is presented multiple times over the course of the learning phase. On the assumption that subjects are able to commit to memory how each CS is precisely related to each US, by the end of the learning phase subjects should have encoded eight *instructed states* (*i.e.,* cognitive representations of instructional content) (see Koban *et al.* 2019: 38). These PFC-dependent instructed states exert top-down influence on the relevant evaluative networks, training them to respond to each CS in a manner congruent with the content of the instructed state. For instance, the instructed state DRUG A PREVENTS EYE INFECTIONS trains the relevant evaluative network to produce a positive affective response to Drug A presentations, while the instructed state DRUG Z PREVENTS HEALTHY HAIR trains the relevant evaluative network to generate a negative affective response to Drug Z presentations. Let's speculate that both of these CS representations are left-amygdala dependent (recall that only the left amygdala been shown to be responsive to CS-US instructions).[49] Because the left amygdala also responds to word valence, the presentation of a negatively valenced CS activates the same region responsible for processing a negatively valenced target word during EPT performance, thereby facilitating the identification of the target as having negative valence. This account of affective priming, then, is a version of the shared valence account described in our discussion of co-occurrence effects.

One virtue of this proposal is that it potentially renders irrelevant questions about whether the repeated CS-US pairings gives rise to right-amygdala-dependent standard evaluative conditioning effects. Recall the relevant empirical research suggests that instruction-induced CS representations exert more influence on CRs than conditioning-induced CS representations, even in cases of conflict either by inhibiting regions that respond to conditioning or by modulating their learning rates (see Koban *et al.* 2019). So, it is possible that the activation of instruction-induced CS representation inhibits the production of right-amygdala dependent expectancies that may shape EPT performance in HGB$_{ex1}$.

One might object that while this account explains the modest relation effect observed with respect to preventative-CSs, it fails to explain why no statistically significant priming effect was observed for causal CSs. In response, there are a number of reasonable explanations of the null-effect consistent with the account just offered. For instance, the null-effect for causal CSs may have been a product of the EPT's unreliability (see Bar-Anan & Nosek 2017). Also, instruction effects on EPT performance are generally modest (see Van Dessel *et al.* 2017); so, it may have been that the study wasn't sufficiently powered to detect a modest effect. Third, recall that Hu and colleagues did

[49] This is not as big of a stretch as it may seem, as the same region appears to track safety-relevant cues in addition to threat-relevant cues (see Olsson, Knapska, & Lindström 2020).

not include an assessment of memory in this experiment. Because the number of unique CS-US pairings that subjects needed to memorize increased two-fold relative to their previous two experiments (both of which had included memory assessments) and left amygdala activation is strongly correlated with contingency awareness during aversive conditioning (Atlas 2019), poor memory for causal CSs in particular could explain the absence of an EPT effect. Of course, none of these possibilities are mutually exclusive with the other. For these reasons, it would be a mistake to infer from the null-effect that the account on offer is flawed *sans* additional research.

Before leaving this topic entirely, I had mentioned as an aside that I would be motioning toward some studies that could aid us in evaluating my proposal. First, recall that we do not know whether the highly idiosyncratic instructed learning procedure used in HGB_{ex1} generates standard evaluative conditioning effects through repeated CS-US pairings in addition to the hypothesized instructed learning effects. Relatedly, we do not yet know whether the same EPT effects would be observed had Hu and colleagues used a more typical instruction only paradigm. But a single study could clear up both sources of ambiguity. This is how the study might look. Holding the types of stimuli used fixed, researchers could simply compare EPT performance across (i) a standard instruction only condition (*e.g.,* The red pill prevents eye infections; the green pill prevents beautiful flowing hair), (ii) an instruction with conditioned reinforcement condition, and (iii) a version of the idiosyncratic learning phase as actually employed in Experiment 3. If EPT performance in (iii) is a function of mere instruction effects, then the results of (i) should not significantly differ from the results of (iii). If EPT performance in (iii) is a function of the interactive effects of both instruction and conditioning, then EPT performance in (ii) should not significantly differ from performance in (iii).

Second, in order to get a better sense of which memory systems are involved, researchers may consider running similar types of study using different types of US. To illustrate, consider there is strong evidence that fearful faces elicit strong amygdala responses. In order to better understand the amygdala's role on evaluative priming performance, investigate relation effects by (a) pairing CSs (*e.g.,* images of pharmaceuticals) with emotion-laden faces (*e.g.*, angry faces (US–), joyful faces (US+)) and (b) explicitly informing subjects precisely how each pharmaceutical relates to each emotion-laden face (*e.g.,* Drug A causes anger; Drug Z prevents anger). By using the different paradigms discussed (*e.g.,* instruction only, instruction with conditioning, and conditioning only), researchers can arrive a better sense of both how each of these factors influence EPT performance and the amygdala's role in driving affective priming.

Lastly, because we should be interested in understanding the relative influence of co-occurrence information and relational information on evaluations when evaluative implications of each conflict in general as opposed to understanding the relative influence of each on EPT performance specifically, researchers investigating these questions ought to address these questions using different types of indirect task (*a la* Moran, Bar-Anan, & Nosek 2017). Indeed, alongside the indirect measurement tasks typically used in attitude research (*e.g.,* the EPT, IAT, and affective misattribution procedure (AMP)), they should also include various physiological measures where appropriate (*e.g.,* cortisol levels, startle-eyeblink responses, SCRs, *etc.*). The use of different types of measure would not only afford us with a better sense of the mechanisms that subserve evaluation, but they should also help us to arrive at more secure conclusions as to the extent to which changes in evaluation actually took place.

5 Conclusion and Future Directions

The central goal of this chapter was to motivate a format pluralist, MSP theoretical approach to resolving the format problem by addressing just one of the three controversies that animates it—

viz. the co-occurrence/relation (C/R) controversy. Rather than use the conclusion to review the details of the proposal on offer, I will instead end the chapter by offering some general remarks regarding how the insights afforded by the discussion of the C/R controversy might help us resolve other two controversies.

I'll briefly review the two controversies, smoothing over the details for the sake of facilitating a broader discussion (see Chapter 4 for more detailed descriptions of each).

> *The Gradual/Rapid change controversy*: In the domain of impression formation, Rydell and McConnell (2006) found that initial implicit evaluations of a fictional individual Jones updated gradually and in direct proportion to the *number* of counter-attitudinal statements presented. By contrast, Cone and Ferguson (2015) found that a single intensely negative and highly diagnostic behavioral statement attributed to Jones could reverse an implicit evaluation of Jones that had been formed after 100 positive conditioning trials.

> *The Inferential Updating Controversy*. Several studies suggest that the contents of implicit attitudes, unlike explicit attitudes, are not influenced by inferential processing. One influential study found that implicit attitudes of fictional groups (*e.g.,* Niffties and Luupites) that are formed through conditioning are not influenced by the newly acquired information that the information that had been paired with one group should actually have been paired with the other group (Gregg *et al.* 2007). By contrast, Mann and Ferguson (2015) observed that rapid inferentially induced changes in implicit evaluations of a fictional individual, Francis West, can occur

> when newly acquired information casts previously acquired information in a new evaluative light.

Upon revisiting these controversies, some commonalities between these two controversies and the C/R controversy are made readily apparent. All three controversies involve questions about the relative influence of, what might be described as, learning through observation of mere CS-US contingencies and some form of social learning on implicit evaluation. Moreover, what animates each debate are the findings that, in some contexts, observational contingency learning seems to be the primary driver of evaluative responses on indirect tasks, and forms of social learning, in other contexts, take precedence over contingency learning in driving evaluative responses on indirect tasks.

Despite these commonalities, the MSP theory does not automatically license the assumption that the mechanisms appealed to in resolving the C/R controversy can be relied upon to solve the other two controversies. (If I thought that the key that unlocks one door unlocks them all, then we would not be having this discussion!) Nothing about the MSP theoretical approach rules out this possibility either. Whether the accounts offered here generalize is, of course, an empirical question.

Having said that, there is good reason to think that the explanations of the co-occurrence and relation effects will bear little resemblance to the explanations of the effects that animate these other controversies. One of the main reasons for thinking this is that there is good reason to believe that the memory systems centrally involved in impression formation tasks differ from those that subserve forms of aversive learning. While the amygdala is centrally involved in aversive learning, the ventral striatum is a critical circuit for both instrumental learning and trait-based impression formation. For instance, Hackel and colleagues (2015) found evidence that representations of generosity and instrumental reward were encoded separately, and while both were correlated with striatal activity, trait representations of generosity also correlated with brain regions previously linked to explicit trait updating. As both systems can be modeled as model-free learning systems and these systems are consistent with both gradual and rapid updating as a function of the magnitude of prediction error together with the quantity of past learning, it is conceivable that both the diagnosticity effects and reinterpretation effects can be accounted for without appeal to propositional memory systems.

Lastly, memory systems research also offers key insights that may help us understand the nature of the reinterpretation effect demonstrated by Mann and Ferguson (2015). The default mode network (DMN) is a network of brain regions that exhibit higher levels of activation than other brain regions even when individuals are at rest (*i.e.*, the DMN exhibits increased activity than other brain regions at baseline). Brain regions that constitute the DMN include the medial prefrontal cortex (MPFC) and the posterior cingulate cortex (PCC). The DMN has been found to be reliably engaged during various tasks that involve self-reflection, mindreading (*i.e.*, inferences about other's mental states), and imagination (Jenkins 2019). Interestingly, heavy cognitive load has been found to deactivate the DMN (*i.e.,* cognitive load results in a decrease in the activity of DMN relative to baseline). Crucially, Mann and Ferguson (2015) argued that the reinterpretation effect (*i.e.*, the effect whereby an implicit evaluation of an individual undergoes rapid change in response to newly acquired information that changes the evaluative implications of previously acquired information) requires minimal cognitive resources on the grounds that the reinterpretation effect is not observed for subjects under heavy cognitive load. While propositional theorists may delight in this finding, as it suggests that even implicit evaluation is driven by propositional states and processes, memory systems research on the DMN suggests an alternative explanation: heavy cognitive load deactivates the regions involved in mental state attribution *simpliciter* (Jenkins 2019). Thus, evidence that the reinterpretation effect does not occur under heavy cognitive load cannot, by itself, be taken as unambiguous evidence in favor of the propositional perspective of attitude formation (*ibid.*).

Attitude researchers ignore memory systems research at their own peril.

CHAPTER 7

CONCLUSION

In this final chapter, I begin by summarizing the content of the dissertation so far (§1), before moving on to address some theoretical loose ends. In particular, I describe the account of explicit attitudes that falls out of memory systems pluralism (§2), urge a memory systems pluralist grounded interdisciplinary investigation of socio-cultural norms (§3), and offer some ecumenical remarks concerning the relationship between MSP theory and its putative rivals (§4). The dissertation concludes with a brief discussion of the models of prejudice reduction that are best supported by MSP theorizing (§5).

1 Summary

Monists about implicit attitudes believe that implicit attitudes form a single homogenous kind. The standard dual-process position, exemplified by Gawronski and Bodenhausen's Associative-Propositional Evaluation (APE; Gawronski & Bodenhausen 2006) model, holds that all implicit attitudes are semantic associations stored in long term memory. The single-process position (De Houwer 2014), by contrast, holds that all implicit attitudes (and explicit attitudes) are propositionally structured representations stored in long term memory. Even those who carve out for themselves positions according to which implicit attitudes are *sui generis* are likewise committed to the view that implicit attitudes are a homogenous kind of representation stored and retrieved from a single memory system (Gendler 2008, Levy 2015, Madva & Brownstein 2018). And while monists largely accept the view that implicit attitudes are functionally heterogenous (*i.e.*, some implicit attitudes may tend to influence judgements or impressions of others more than behaviors toward them and *vice versa* (see Holroyd & Sweetman 2017)), they tend prefer monist theoretical explanations of functional heterogeneity that appeal to differences in content (*e.g.*, stereotyping someone as angry might have a larger effect on behavior than does stereotyping someone as good at math), differences in how they were formed (*e.g.,* were associations directly formed *via* the perceived spatiotemporal contiguity of x and y, or were the associations formed indirectly formed *via* propositional processes (*a la* Gawronski & Bodenhausen 2006)?), or differences in retrieval (*e.g.*, were the attitudes retrieved *via* a process of similarity matching or *via* conscious recall (*a la* De Houwer 2018) over the sort of explanations that would commit us to the existence of many kinds of implicit attitude (Brownstein 2017, Holroyd & Sweetman 2017).

I argued that we ought to reject monism given its failure to adequately address two sets of anomalies. On the one hand, implicit attitudes are poor predictors of behavior (Greenwald, Poehlman, Uhlmann, & Banaji 2009, Oswald, Mitchell, Blanton, Jaccard & Tetlock 2013, Forscher, Lai, *et al.* 2019), and the standard measures of implicit attitude poorly correlate with each other (Bosson, Swan, Pennebaker 2000, Olson & Fazio 2003, Payne, Govorun & Arbukle 2008, Sherman 2008) and are such that the outcome of any given indirect measure seems to covary in theoretically unanticipated ways as a functioning of seemingly trivial changes in experimental procedure (Dasgupta & Greenwald 2001, Schaller, Park & Mueller 2003, Rudman & Lee 2002). Machery (2016) exploited the existence of these anomalies in motivating a trait view, an implication of which is that to even speak of implicit (or explicit) attitudes is to make a category mistake. And while Machery maintains that these anomalies are fall out of an account of implicit attitudes are mental states, I argued that this was a misdiagnosis. The correct diagnosis is that these anomalies are symptomatic of implicit attitude monism, on the grounds that monism, *ceteris paribus*, predicts higher than actually observed predictive power for implicit attitudes, stronger than actually observed correlations across indirect measures of attitude, and less variance in the outcome of a single indirect measure completed on two separate occasions as a function of seemingly irrelevant situational factors.

In addition to the set of anomalies that Machery capitalizes on in motivating his trait view, we have the set of format anomalies that jointly constitute the *format problem*. In brief, implicit attitudes appear to be (a) both insensitive and sensitive to explicitly presented relations between stimuli (Hu *et al.* 2017a, Moran & Bar Anan 2014), (b) capable of gradual and rapid updating in response to newly acquired information (see McConnell & Rydell 2014, and Cone & Ferguson 2015, respectively), (c) both responsive and unresponsive to reason (Ferguson *et al.* 2019). Each set of findings is a source of tension for monist theories of implicit attitude regardless of whether they are committed to associationism, propositionalism, or to a view on which implicit attitudes are *sui generis*. And while an account according to which implicit attitudes are a general type the subtypes of which are such that their instances can be either propositions or associations (but never both) is capable of reconciling these seemingly inconsistent findings, one cannot endorse such an account without thereby abandoning monism, as this would be to admit that a single, unified theory of implicit attitude cannot be provided.

In place of monism and eliminativism, I have urged that psychologists and philosophers adopt memory systems pluralism. Memory systems pluralism is largely inspired by David Amodio and colleagues' Multiple Memory Systems Model of Social Cognition (MMS-SC; Amodio 2019, Amodio & Devine 2006, Amodio & Ratner 2011). Memory systems pluralism holds that (a) there exists in the mind-brain multiple independent yet interactive memory systems, (b) these memory systems contribute in many important ways to the production of the various explananda phenomena that social cognition researchers have posited implicit attitudes to explain, and (c) implicit attitudes are the general type of thing each subtype of which is individuated by the memory system that governs the encoding, maintenance, updating, and retrieval of its instantiations. Whether MMS-SC, as developed by Amodio (2019) and colleagues is an instance of memory systems pluralism turns on a further question of whether MMS-SC is committed to (c), as it is clearly committed to both (a) and (b). Memory systems pluralism, then, is best understood as a theoretical framework or a general approach to accounting for phenomena of central interest to social cognition researchers rather than a fully developed model of social cognitive phenomena. Having said that, I have taken care to ensure that, wherever possible, my applications of memory systems pluralism do not turn on controversial views in cognitive neuroscience. When the conclusions that I do reach are inconsistent with specific memory systems models, as was the case in Chapter 5, I have gone to great lengths to provide independent evidence that these memory systems models ought to be rejected anyway. So, while I neither endorse nor articulate a specific model of memory systems pluralism, this does not (I hope) undermine the plausibility of any of the more substantive conclusions at which I arrive.

I argued that memory systems pluralism yields the best available account of the three anomalies that drives Machery's eliminativist challenge and is best positioned amongst its rivals to resolve the format problem. With respect to Machery's eliminativist challenge (articulated in Chapter 2), I argued that memory systems pluralism not only vindicates the view that implicit attitudes are mental states, it also reveals that the predictive-explanatory power of Machery's trait view is wholly parasitic on the very research that animates memory systems pluralism (Chapter 3). With respect to the format problem (articulated in Chapter 4), I have attempted to pave the way for a memory systems pluralist solution by pursuing two aims; the first aim involved establishing that propositional states and processes can make important and direct contributions to various kinds of indirect measure, all of which have been understood as tapping states and processes that are either (a) unavailable to self-report, (b) not-dependent on working memory, (c) unintentional in their activation, or (b) uncontrollable once activated (Chapter 5). The second aim was to showcase just what goes into a memory systems pluralist explanation of any given controversy in social cognition research that involves the contributions of multiple memory systems by taking a narrow focus on the co-occurrence/relational effect (C/R) controversy (Chapter 6). While I do not pretend to have offered a complete solution to the C/R controversy and, much less, to the format problem *simpliciter*, I do hope to have demonstrated the general utility of the memory systems approach and the superiority of this approach over its rivals.

In sum, I have argued that the anomalies that have accumulated over the past four decades of sustained social cognitive research on implicit attitudes serve as an indictment of the dominant picture in implicit social cognition research according to which (a) implicit attitudes are a natural kind (*i.e.*, the term 'implicit attitudes' picks out a kind, K, such that the members of K have scientifically relevant properties in common) and (b) the single memory system in which all implicit attitudes are stored plays the role of the causal mechanism that explains the many scientifically relevant properties shared by all implicit attitude tokens. In place of this picture, I have urged that we instead embrace a picture according to which (a*) implicit attitudes are not a natural kind; rather, it is much more likely that many scientifically relevant generalizations are true of the members of subclasses of implicit attitude and (b*) each subclass of implicit attitude is underwritten by a different causal mechanism, namely the specific memory system involved in the encoding, updating, and retrieval of each kind of implicit attitude. Not only does this picture either resolve a number of the more pressing and longstanding psychometric anomalies that many people (both within the field of social cognition research (Forscher, Mitamura, Dix, Cox, & Devine 2017, Machery 2016, 2017a, 2017b) and without (Bartlett 2017, MacDonald 2017)) have seized upon in an attempt to discredit or dismiss this critical area of social cognitive research, it also provides us with the most promising path toward resolving many anomalies concerning the representational format of implicit attitudes.

Worries about parsimony are largely misplaced, as the architectural assumptions that ground the memory systems pluralist approach are already built into the theoretical foundations of many of the most successful programs in cognitive and social cognitive neuroscientific research. Insofar as social cognitive theorizing about attitudes should attempt to not only produce internally (relative to the field) consistent theories of implicit attitudes but also theories of implicit attitude that cohere well with the most well established theoretical perspectives in adjacent fields of cognitive scientific research, there is no reason to think that the memory systems approach necessarily commits us to any theoretical entities over and above those that best theories in cognitive science already commit us to (*c.f.,* Amodio 2020). The virtue of the memory systems approach, then, is that it resolves a range of otherwise crisis-inducing anomalies and maximizes external consistency of implicit social cognitive theorizing, all while carrying the added benefits of preserving the most well established empirical findings of "old school" social cognitive research, increasing the explanatory power of social cognitive theories of attitude without also increasing the overall number of theoretical posits, and promoting the very sort of interdisciplinary collaboration that cognitive scientists of various disciplines had long ago recognized as being necessary for understanding the mechanisms that produce intelligent behavior social or not. The remainder of this chapter is devoted to tying up some loose ends and to drawing connections between philosophical topics and future work in the memory systems paradigm.

2 What of Explicit Attitudes?

In this dissertation, discussions about the nature of explicit attitudes have taken a backseat to discussions about the nature of implicit attitudes. When explicit attitudes were the subject of more or less sustained theorizing, this theorizing typically occurred in contexts in which understanding why implicit attitudes and explicit attitudes sometimes dissociate was necessary for deepening our understanding of the mechanisms that underwrite various subclasses of implicit attitudes and processes through which implicit attitudes are expressed in behavior. To be sure, throughout the dissertation one finds various strands of discussion about the nature of explicit attitudes, which, if collected, reveal how explicit attitudes and their relation to implicit attitudes should be treated by the memory systems approach. It is now time to collect these strands to get a better sense of how the memory systems pluralist ought to approach explicit attitudes.

Since implicit attitudes are often characterized in opposition to explicit attitudes, and *vice versa*, let's kick off this discussion by posing the question, how should we best understand the relation between implicit attitudes and explicit attitudes given memory systems theoretical framework? This is one question whereby the answer that memory systems theorist gives may depend on whether one prefers, say, Squire and Zola-Morgan's (1991) medial-temporal lobe (MTL) model, to the family of memory systems models that most impress Schacter and Tulving (1993; see also Tulving 1995). Recall that Squire and colleagues MTL model divides memory systems into two broad categories, namely (a) explicit/declarative, and (b) implicit/non-declarative. It is the MTL model, moreover, that most directly contribute to the rise of dual-systems (as opposed to dual-process) theories of social cognition (Smith & DeCoster 2000). If one accepts the MTL model, then it looks as though we have reason to preserve the conventional view of explicit attitudes according to which explicit attitudes as a kind of mental state distinct from the perhaps motley assortment of states that we call 'implicit attitudes.' In other words, the MTL model is consistent with both a rejection of IA-monism and a view on which explicit attitudes do form a natural kind.[50] With that said, I argued at length in Chapter 5 that we have good empirical and theoretical grounds for rejecting a Squire-esque division between implicit memory systems and explicit memory systems.

[50] The MTL model, recall, treats explicit attitudes as those that are stored, maintained, and expressed via the operations of the semantic and/or episodic memory systems. Implicit attitudes are those that are stored, maintained and expressed via the operations of the non-declarative memory systems (*e.g.,* the priming system, the procedural memory system, the conditioning system, *etc*.). Since explicit attitudes depend on different memory systems from implicit attitudes, we should think of these different types of attitudes as having distinct representational bases.

Once this taxonomy is rejected, where does this leave us? As noted earlier, I am inclined to endorse Schacter and Tulving's (1994) view according to which the terms 'implicit memory' and 'explicit memory' pick out different ways in which a mnemonic representation might be expressed in a particular context (Tulving and Schacter call these different *forms* of memory, though I find the notion of a 'form of memory' to be somewhat obscure). On their view, an output of a memory system retrieval process is explicit when, for instance, the retrieval involves a voluntarily initiated search (*e.g.*, when I search memory for the name of the pop album that the Vatican had rated as the second best of all time) or when the memory is expressed such that I have conscious access to its output (*e.g.*, the search of memory yields David Crosby's solo-debut *If I Could Only Remember My Name*). If one would like, one could also tack onto this list another standard disjuncts (*viz.* explicit expressions of memory are working memory dependent, an expression of memory is explicit if the search that led to its expressions could have been halted after it had been initiated). By contrast, an expression of memory in a given context is implicit if (a) the retrieval process that led to its behavioral expression was initiated without intent, (b) the output of the retrieval process does not drive verbal report, (c) the expression of the memory was not heavily dependent on working memory capacity, or (d) the retrieval process is ballistic *(i.e.*, its expression cannot be halted once the retrieval process has been initiated (short of things like knocking the person unconscious prior to its expression, or applying transcranial magnetic stimulation (TMS) to a critical pathway prior to its expression, *etc.*). It is an open empirical question as to whether there are any memory systems such that their retrieval processes and their outputs are always best characterized as being either implicit or explicit (or, weaker, whether there are any memory systems such that their retrieval processes or their outputs are typically characterized as implicit or explicit). Should this turn out to be the case, then there is a sense in which it would be appropriate to describe the memory system as explicit or implicit as elliptical for describing its characteristic mode of expression. On this view, it is possible for the same mnemonic representation to be expressed explicitly in one context and implicitly in another.

Proponents of the propositional view have argued that while a retrieval process or its output may be described in either implicit or explicit terms, the process of learning is always explicit (De Houwer 2014)— that is, learning is in all cases involves the conscious formation and assent to some proposition or another. While this position on explicit learning can be accommodated by a memory systems pluralist framework, there is good reason to think that at least some propositional memory systems sometimes engage in implicit learning (see Chapters 5 & 6). Though, again, which specific memory systems are capable of learning implicit/explicitly is question that can only be settled empirically.

Upon returning to the guiding question of this section, we find that it is underspecified to the point that it is difficult to even know what it is that's being asked. Understood as a question about the difference between implicit modes of expression and explicit modes of expression, the question has already been answered. But if the question is understood as asking about the general relation between the mnemonic representations that are expressed implicitly and those that are expressed explicitly, then it begins to sound as though it is premised on a category mistake. So, what is the nature of explicit attitudes on this account? If empirical investigation leads to the discovery of one type of mnemonic structure that is only ever expressed explicitly and another type of mnemonic structure that is only ever expressed implicitly, then I suppose that we might be able to offer an answer to this question. If, however, it turns out that there are no such memory systems the outputs of which are only ever expressed implicit/explicitly, it is not at all obvious that this question admits of an answer. An upshot of this discussion, then, is that there may be no meaningful distinction to be drawn between implicit/explicit *representations*. And if attitudes are simply mnemonic representational structures that causally contribute to various categories of evaluative response in the right sort of way, then there is no distinction between implicit and explicit attitudes. Attitudes can be learned implicitly or explicitly, retrieved implicitly or explicitly, and/or expressed implicitly or explicitly, and we can meaningfully inquire as to the nature of these different kinds of process. But we should nevertheless abandon the view that explicit attitudes are a kind of attitude.[51]

A related question worth addressing is this: when subjects are asked to self-report their attitudes and subjects comply by offering sincere attitude reports, what exactly are they reporting? Put another way, which kinds of causal structures and processes generate attitude reports? Moreover, assuming that subjects are able to accurately report on the contents of certain types of attitude, what explains this? These are important questions for future research, and a full exploration of these issues would take us well-beyond the scope of the dissertation. With that said, I suspect that question about the types of structure and process that drive attitude reports admits of multiple substantively different yet nevertheless compatible answers. It may be that, under certain conditions, sincere verbal attitude reports are primarily driven by the outputs of a single memory system (*e.g.,* if one were to put a live tarantula in close proximity to a severe arachnophobe and were to ask this person to rate their attitude toward tarantulas on a scale of 1 to 7 (1 = extreme disliking, 7= extreme liking), then it may be that the '1' rating reflects the relevant contents of the PAC system). In other contexts, sincere verbal reports of liking may be tracking states that represent the socio-cultural norms or mores that govern attitudes about the relevant domain (Hesslinger *et al.* 2017*)*(*e.g.,* I report that I like *The Office* (U.S.) because disliking *The Office* (U.S.) is apparently a socio-cultural more). In yet other contexts, it may be that sincerely self-reported attitudes are more aspirational than merely descriptive (*e.g.*, it may be that self-reporting egalitarian attitudes conveys information that the individual aspires to have or is, in some sense, committed to having egalitarian attitudes even if the

[51] One might complain, however, that I have not offered answers to any of these more specific questions about implicit/explicit learning, retrieval processing, or expression. And the basis of the complaint is accurate. But these are best treated as questions for future research.

individual does not presently possess such attitudes) (*cf.* Alfano 2014). Perhaps self-reported attitudes are best conceived of as the output of a self-report module that receives its inputs from various memory systems, where the set of memory systems that feed into the self-report module vary with context. This mechanism (or these mechanisms) could one of any number of computational algorithms. For instance, it may be that the self-report module takes the average of the evaluative strength of its inputs, and it is the product of this averaging process that contributes to verbal report. This may vindicate the classical view of attitudes as *summary evaluations* that are spontaneously constructed, on-the-fly, in response to task demands (*a la* Krosnick, Judd, & Wittenbrink 2014). Perhaps the self-report module operates on a winner-takes-all basis, thereby generating a self-report that reflects the contribution of the most active memory system (see O'reilly 1998). These are just a few of the many possibilities about the causal structures that drive attitude reports, and there is no reason to think that theorists are limited to picking just one. It may come as a shock to you, but I am drawn toward pluralism here.

Second, future empirical and theoretical investigations should aim to account for the factors that determine whether a mnemonic representation stored in a specific memory system is expressed implicitly in one context and explicitly in another. And, again, memory systems pluralism is consistent with a range of theories. It may be that whether the output of a memory system drives verbal report (*i.e.*, the output of a memory system is explicitly expressed) in some context is partly determined by the number of redundant representations (*i.e.*, the number of psychological-level units that possess the same representational content) active in that memory system (see Rupert 2011: 112). Perhaps whether an attitude is expressed implicitly or explicitly depends on whether a vehicle that with that attitude's representational content makes its way into working memory, which, on this account might serve as a global workspace (Baars 1988, 1997, 2002). Maybe only those attitudes that can be explicitly expressed are those that are stored in a memory system that has some sort of content monitoring mechanism as a part (*a la* Nichols & Stich (2003)). Perhaps the representational structures that contribute to verbal report are never themselves expressed explicitly; rather, we either come to self-attribute likings and dislikings on the basis of some theory, developed from infancy on, about what we like and/or dislike (*a la* Gopnik & Meltzoff 1997) or come to make these self-attributions on the basis of some faculty that infers our likings and dislikings on the basis of the input it receives from the minds various memory systems (*a la* Carruthers 2011). Some of these accounts are compatible with each other, while others are not. Sorting these issues out is a matter for future research.

Before we leave the topic of the nature of explicit attitudes entirely, a related question is worth addressing. Peter Carruthers (2017: 70) has recent defended the view that "the causal structures underlying both explicit and implicit attitudes are the same." This view, if true, has important implications about whether attitude research can teach us anything new about the ontology of the mind. He summarizes his position thusly:

> I have argued that it is the same types of underlying representation that give rise to both explicit (communicative) and implicit (non-communicative) behavior. Because the behavior-types dissociate, this can create the impression that there are two kinds of representation on play: explicit and implicit. But this may be an illusion, resulting from our lack of familiarity with the different causal processes involved in each case. Once these are detailed it becomes plausible that the underlying representations don't really differ in type. Indeed, in many cases it can be the very same token representation that is involved in the causation of both kinds of behavior under different conditions. (*ibid.*)

Memory systems pluralism can accommodate this view—nothing articulated in this passage is inconsistent with memory systems pluralism. Moreover, it appears as though Carruthers and I agree that the implicit/explicit distinction, insofar as it is worth preserving, cannot accurately be employed to identify two distinct kinds of representation.[52] But it is one thing to note that memory systems pluralism can accommodate this view, and another thing to assert that the memory systems pluralist should accept this view. So, should we maintain that the types of states and processes that factor into self-report are the same as those that drive indirect performance?

The short answer is, "probably not." I am skeptical of this position for three reasons. The first is that, as argued in the previous chapter, it might be the case that some indirect measurement tasks like the EPT do not draw from the same set of memory systems as self-report measures. Whereas self-report measures may reflect contributions from, say, the episodic memory system, it is difficult to see how the subset of episodic memory structures that express evaluations or stereotypes could exert influence on EPT performance.[53] If so, then this would be a case in which the set of causal structures that drive indirect measurement task performance is not the same set of causal structures that drive self-report. Now, can one in principle design an indirect measurement task that draws from the same memory systems as direct tasks? This certainly is not inconceivable. But the proof is in the pudding.

The second reason for skepticism is this: Kurdi and Banaji (2018) had subjects complete an instrumental learning task in which subjects were conditioned to associate interactions with the fictitious social group called "the Luupites" with reward and another fictitious group called "the

[52] I hedge as I am most inclined to quibble with the penultimate claim. But I'll expand on this below.

[53] To be sure, episodic memory might influence EPT performance in other ways. For instance, I might retrieve information from episodic memory in generating response routine such that whenever a prime of the target category is presented, I hit the "negative" key. If the implementation of this response strategy is driving responses, then EPT performance in this instance tells us nothing about what the subjects thinks of or feels toward members of that category. So, episodic memory may contribute to EPT performance in ways that reveal nothing substantive about the construct of interest.

Niffites" with punishment. At a certain stage in the experiment, an instructed reversal was introduced (*i.e.*, subjects were told that the expected value of interacting each group is now reversed) and subjects then completed both an IAT and a measure of self-reported likings. The researchers found that while model-free learning algorithms predicted subjects' IAT performance both before and after reversal, self-reported attitudes of the Niffites and the Luupites seemed to reflect the contributions of both model-free and model-based learning algorithms. These findings suggest that while there may be a great deal of overlap between the causal structures that contribute to performance across both measurement tasks, self-report measures sometimes draw on causal structures that are not implicated in IAT performance. Again, it is possible, in principle, to develop an indirect measurement task that draws on the same types of model-based structures as those that drive self-report responses on Kurdi and Banaji's (2018) studies? And, again, I provide the same response: I have no reason to doubt this. The point, however, is that our best available empirical evidence suggests that performance on direct measurement tasks sometimes recruit causal structures that do not influence performance on indirect tasks. Moreover, since the memory systems framework gives us reason to be skeptical of the hypothesis, each actually used measure is such that it is in principle capable of drawing from the same set of memory systems, we should not share Carruthers' confidence that the causal structures that drive indirect task performance are the same as those that drive direct task performance.

Here's the third reason. Above, I suggested that the putative mechanism that transforms the outputs of various memory systems into a product that can be used to guide verbal report may vindicate the summary evaluation account of attitudes. On one version of such an account, the relevant mechanism maps the representational outputs of the contributing mechanisms onto a representation that serves, in this context, as a summary evaluation. This summary evaluation is a representational structure that is generated on-the-fly in response to task demands. Is it possible that such representations could be constructed on-the-fly in response to demands imposed by some an indirect measurement task? For the last time, there is no principled reason for rejecting this possibility. But the current state of empirical evidence is such that we have no good reason for thinking that this actually occurs.

The lesson here is that because we don't have a firm grasp of which memory systems tend to be recruited by the various indirect tasks, and which memory systems tend to be recruited by self-report tasks, it follows that we do not have a firm grasp on whether both types of task tend to recruit the same causal structure. Therefore, without further research, we simply have no way of knowing the extent to which implicit attitude research has ontological import. *Contra* Carruthers (2017), I am willing to place my empirical bets on the claim that further empirical research will eventually reveal that this area of attitude research has led to the discovery of at least one type of mental state that may not have been otherwise discovered.

In sum, the MSP theoretical stance toward explicit attitudes mirrors its stance toward implicit attitudes. That is, memory systems pluralism recommends that we reject the hypothesis that explicit attitudes form a homogenous, unified kind. Future research should be aimed at explaining (a) which memory systems tend to contribute to self-report tasks under a variety of contexts, and (b) how these memory systems contribute to self-report, and (c) whether there are any representational structures that uniquely contribute to self-report measures or indirect measurement tasks. But the primary lesson to take away from this discussion is this: the standing assumption that explicit attitudes are far better understood than implicit attitudes is false.

3 The Psychology of Socio-Cultural Norms

One might find objectionable that, until now, there has been little to no discussion of the role that socio-cultural norms play in shaping all manner of responses, evaluative and otherwise. Here are just a few examples. There is strong evidence that the social norms play an important role in governing aesthetic judgements (Hesslinger, *et al.* 2017), fear responses (Mu, Han, & Gelfand 2017), the degree to which a stimulus is considered painful (Koban & Wager 2016), perceptual judgements (Sherif 1936), moral judgements (Henrich, *et al.* 2010), *etc.* Some of the studies on the effects of social norm acquisition, moreover, rely on measures of self-report while others do not (see Sripada & Stich 2005 for review). Similarly, there is strong evidence that socio-cultural norms also help to shape mechanisms, presumably early on in development, that have long been thought to be resistant to exogenous influences (*ibid.*). In other words, the effects that one's culture has on the development of various perceptual and cognitive faculties, one's evaluative responses to a range of stimuli, and one's perceptual judgements run deep.

I share this concern. A full accounting of the mechanisms that drive various types of evaluative response must also account for the role played by socio-cultural norms in shaping these responses. In defense of my neglect of this critically important topic, research on the impact of socio-cultural norms, described as such, on indirect measure performance is scant. To be sure, debates about whether stereotypes reflect one's "true attitudes" toward members to which the stereotype applies or merely socio-cultural *associations* are ubiquitous in the literature. But such debates typically conflate the encoding of *statistically normal regularities* found in one's socio-cultural environs (*e.g.*, that Blacks are more likely to be arrested, charged, and convicted of crimes than Whites) with the encoding of socio-cultural norms (*e.g.*, one ought to use "sir" or "ma'am" when addressing an unfamiliar elder). While it may be that the encoding of statistical normal regularities may serve an important causal role in coming to acquire or encode socio-cultural norms, it is nevertheless the case that the constructs are distinct (*e.g.*, Smith may respond with disgust upon hearing that Jones, an American citizen, regularly uses an American flag to clean his toilet, despite Smith's never having directly experienced anyone engage in this behavior). Despite the long and venerable history that the social-norm construct has played in social psychological theorizing (Ascher 1951, 1956, Sherif 1937), there is comparatively little empirical and theoretical research on the mechanisms that underly social-norm acquisition and the role norms play in shaping various types of evaluative response. As such, there are few if any major puzzles about social norms in the implicit social cognition literature. Since I have been primarily concerned with addressing outstanding puzzles in attitude research, I hope this helps to explain my neglect of this critical topic.

MSP theory promises to play an important role in regimenting future empirical and theoretical investigations of social-norms and the mechanism (or mechanism) that subserve their encoding, maintenance and expression. While I will not offer my own view here (it is far from being fully developed), I will make some brief suggestions as to where the memory systems pluralist ought to focus her attention.

Sripada and Stich (2006) argue for the existence of a norm system. This norm system contains a mechanism for extracting norms from one's immediate socio-cultural environment, a database that stores the acquired norms, a mechanism that governs norm expression or compliance, and a mechanism that governs norm enforcement (which, in some cases, involves taking punitive action— expressed either in terms of overt behavior or in terms of private judgement) when socio-cultural norms are violated. Having said this, it appears as though Sripada and Stich are positing the existence of a type of memory system that differs from the other memory systems that we have discussed. An immediate question, to be addressed by future research, is this: Should this type of special purpose norm system be included in the final memory systems model of social cognition? If so, then how does it interact with the other memory systems models that also govern similar kinds of evaluative response. For instance, I might know that I ought to remove my shoes before entering a traditional Japanese household, and yet fail to do so as a result of habit (which, perhaps, is an expression of the model-free, striatum-based, instrumental learning system). Of course, it may be that the memory system that supports knowledge of a socio-cultural norm (where, 'knowledge' here is used as psychologists use it) is not the same as the norm's having been successfully encoded in the norm-system's data base, in which case my failure to comply with the norm can be regarded as (defeasible) evidence that the norm has not been successfully registered in the hypothesized norm-data-base. On the other hand, it may turn out that there is no special purpose norm-system at all–– rather, behaviors that we would countenance as either norm-compliant or norm-violating in a particular social context emerge from either the expression of (i) a single more domain general memory system (*e.g.*, the instrumental learning system) or (ii) the interaction of multiple memory systems. Let's call the refer to the former possibility as a *reductive account* and the latter as a *distributed account*. Future theoretical work within the memory systems tradition should aim to identify which of the three accounts is most plausible.

Empirical investigations of the influence of socio-cultural norms on various types of evaluative response need not, and should not, wait until the above theoretical questions have been addressed before proceeding. In the meantime, researchers can begin to address the role that norms play in shaping indirect task performance by engaging in more cross-cultural research. A researcher might also investigate the roles that norms play in social norms by having subjects complete direct and indirect measures of attitudes after taking part in a learning procedure in which subjects are first conditioned to associate an individual with, say, high instrumental reward before revealing that the individual routinely violates certain norms (*e.g.*, it may be revealed that Jones engages in safe, consensual sex with his sibling). Moreover, previous studies have shown that perceptual judgements mediated by the acquisition of social norms are temporally stable (Endler 1960). Further empirical research ought to be conducted so as to determine whether the same is true of various types of social-norm mediated evaluative responses. Lastly, several theories of norm-acquisition identify stable cross-cultural patterns of ontogenesis (see Sripada & Stich 2005). For instance, it is found that the norm system tends to come online between the ages of three to five. Social-psychologists, therefore, ought to collaborate with developmental psychologists in investigating whether children are able to exhibit the relevant forms of norm-relevant behavior before the period in which the norm systems come online, which may yield critical insight as to which forms of seemingly norm-compliant behavior are actually governed by social-norms and which forms of norm-compliant behavior are explicable without appeal to social-norms.

4 A Theory of Everything?

Though I have been highly critical of how attitude researchers have gone about investigating attitudinal phenomena, I would like to end by stressing that at a sufficiently abstract level of analysis, MSP theory is largely compatible with many, but by no means all, of the most influential theories of attitudes that have been developed. To illustrate, consider the MSP theory's relationship between the associative view, the propositional view, the summary evaluation view, and the situationist view.

The MPS theory assumes that some implicit attitudes are associative structures, and so extant associative accounts may help us to account for those behaviors that such structures manifest. Such views may even offer valuable insights as to how the semantic-associative system interacts with the systems that govern self-report *a la* the APE model. We ought to reject, however, the standard assumptions that all implicit attitudes are associative structures encoded in a single long-term memory system; some attitudes are associations between concepts and affect; and all measures of implicit evaluation tap such structures.

The MSP theory that has been developed here assumes that some implicit attitudes are propositional structures, and so extant propositional accounts may explain the properties of these states and the behaviors they manifest. The MSP theory recommends that we reject, however, De Houwer and colleagues' assumptions that all attitudes are the products of conscious propositional formation and assent, and that all such states are encoded in a single long-term memory store. By contrast, MSP theory better coheres with Mandelbaum's (2015) belief theory of implicit attitudes, on which implicit attitudes are sometimes unconscious beliefs the tokens of which are not all stored in the same belief-box.

Recall that the summary evaluation view is classically conceived of as a monist theory of attitudes, according to which various components of an attitude come to be crystalized into a summary evaluation of an attitude object. The summary evaluations are expressed, under the right circumstances, on self-report measures and its putative components manifest evaluative responses on the various indirect tasks (Krosnick, Judd & Wittenbrink 2011). The MSP theory is consistent with the view that summary evaluations manifest responses on self-report tasks. It denies, however, that the representations that drive behavior across various indirect measures are, in any sense, *components of* summary evaluations. A more plausible view is one on which the putative mechanism that produces summary evaluations takes as its inputs the globally broadcast outputs of some class of memory systems (*cf.* Carruthers 2011). To be sure, it may be that a stable summary evaluation is never itself stored in long-term memory and that, instead, summary evaluations are constructed on the fly in response to situational demands (Schwarz 2005). Whatever the case may be, the MSP theory is consistent with both views. Indeed, by framing this debate in terms of the interactions of various memory systems and the mechanisms that operate over their respective outputs, the hope is that we'll be better positioned to make new progress on this relatively idle yet longstanding theoretical dispute.

Back in Chapter 1, I had briefly described Payne, Vuletich and Lundberg's situationist theory of implicit attitudes (or the *bias-of-the-crowds* model) according to which measures of implicit bias are "meaningful, valid, and reliable measures of situations rather than persons" (Payne, Vuletich, & Lundberg 2017: 236) and suggested that this theory of bias is not obviously at odds with MSP theory. It may be that these accounts operate at different levels of analysis— whereas MSP theory, properly understood, is meant to account for intelligent social behavior and their underlying mechanisms, the situationist view is primarily interested in accounting for social-level phenomena and their causes (which may or may not ultimately be reducible to properties at the individual level) by aggregating responses measures of bias.

` Though the MSP theory may be unable to account for all phenomena of interest to the situationist attitude theorist, the MSP theory helps to explain both (a) why individual-level evaluations so-often supervene on environment-based contingencies, and (b) why, in some cases, individual-level evaluations float-free from environment-based contingencies. To illustrate, recall the research on instructed reversals described the previous chapter. There we discussed how the right amygdala tracks actually experienced aversive outcomes and is not responsive to social instruction (Atlas 2019). Moreover, recall that certain forms of amygdala-based representations, once formed, are highly resistant to extinction. These bodies of research jointly account for why evaluative responses elicited by indirect measures of attitudes sometimes shift dramatically as a function of the situation and why evaluative responses sometimes fail to update in response to sudden and dramatic changes in the relevant environment-based contingencies. Because the bias-of-the-crowds model treats implicit tests of, say, racial evaluations to "reflect the net accessibility of all attributes linked to the social categories of black people and white people (Payne, Vuletich & Lundburg 2017: 237)" shared by a *sample* of research subjects (as opposed to individuals) the situationist research will not lead to the discovery of important forms of cross-situationally persistent individual bias much less explain them. The moral of this story is not that situationist attitude research is not worth pursuing but that this line of research is incomplete without the kind of individual-level theory that MSP offers.

5 Fighting Racism with Solidarity

Finally, let's shift our focus to describing some of the implications of MSP theory for the development tools or *interventions* for mitigating real world social harms caused by the activation of prejudicial attitudes, implicit or otherwise. MSP theory predicts that the most effective interventions are those in which individuals are afforded the opportunity to retrain multiple memory systems in parallel, resulting in the production of multiple pro-social attitudes toward members of the target social group. For instance, a successful intervention may provide an individual with an opportunity to (i) form positive episodic memories of encounters with outgroup members, (ii) to treat interactions with outgroup members as being instrumentally valuable, (iii) reduce any learned fear or threat responses to outgroup members, (iv) associate outgroup members with appetitive rewards, and (v) create a sense of community, perhaps facilitated via categorization processes, amongst what would otherwise be a mere aggregate of members of disparate groups. Because these mnemonic representations are subserved by different neural systems for storage and expression, a targeted approach in which one intervenes on only a subset of memory systems may have little or no effect on the representations in the other.

This assumption serves as a partial explanation for why targeted, implicit bias training seminars have proven to be ineffective at producing sustained change to subjects' implicit attitudes and reducing discrimination. To illustrate, consider the standard suite of prescriptions for reducing bias that include (i) imagining counter-stereotypical exemplars of stigmatized groups (Blair, *et al.* 2001), (ii) maintaining constant vigilance for when one might be tempted to act in a way that is contrary to one's egalitarian values, (iii) forming implementation intentions (*e.g.*, IF I SEE A BLACK PERSON, THEN I SHOULD THINK 'SAFETY'!) and (iv) acknowledging one's own sources of privilege. From the MSP theoretical perspective, these interventions may fail because they (a) produce changes in the kinds of attitude that are not functionally yoked to the discriminatory outcomes that we're most interested in addressing, (b) leave too many forms of prejudicial attitude intact, (c) produce changes in attitude that are quickly extinguished upon leaving controlled conditions, (d) result in the generation of new context-specific attitudes that are operative only in the environment in which they are formed (Bouton 2004), or (e) some combination thereof.

On a more positive not, MSP theory recommends more promising forms of prejudice-reducing interventions. Consider *intergroup contact theory* (Allport 1954, Sherif 1954, 1958, 1961), which has historically been and continues to be the gold-standard for prejudice reduction (see Pettigrew and Tropp 2006 for review). This theory assumes that sustained intergroup contact under favorable conditions will reduce prejudice between members of the relevant groups. Gordon Allport (1954) identified four factors, which are by no means intended to be exhaustive, that tend to facilitate prejudice reduction as a consequence of repeated intergroup contact: (1) equal status of groups (*i.e.*, one should not be in a position of authority over another), (2) common goals, (3) intergroup cooperation, (4) contact is supported via law or custom.

Just as MSP theory correctly predicts the failures of the implicit bias training approach to prejudice reduction, it also predicts the successes of intergroup contact theory. But in order to see this, first consider the following scenario. Imagine that both Smith and Jones are workers at a distribution facility for a delivery company, Shamazon, owned by the richest person on the planet, Jed Bozos. Smith, a black woman from a residentially segregated neighborhood, is highly suspicious of Whites and generally avoids interacting with them whenever possible out of fear of discrimination. Jones, a White man, also lives in a residentially segregated neighborhood and subscribes to various, reactionary, pernicious stereotypes about Black people in generally. Suppose that Smith and Jones are members of the same labor union, and this labor union calls a strike in protest of Shamazon's failure to provide dignified working conditions and a living wage. Out of self-interest, both Smith and Jones join the picket line. Day in and day out, Smith and Jones chant and march in unison. As the strike stretches on, Smitch notices that Jones' morale begins to flag to which she responds by giving him an encouraging pep-talk. Perhaps Smith takes this action only because he knows that bad morale is contagious, and so her prospects for better financial conditions depends on this admittedly out of character behavior. Days later, Jones reciprocates and shares with Smith some of his meal; they bond over the fact that they each learned to make variants of the same dish as children. The next day on the picket line, Jones notices a scab berating Smith as he prevents the scab from crossing the picket line, causing the scab to hurl at her racial epithets. Out of solidarity for his fellow worker, he quickly steps into defend her, though she quickly makes clear that she is quite capable for handling herself. Nevertheless, she's grateful for the Jones' act and acknowledges Jones as someone who she can trust to have her back. Though the strike is ultimately crushed by Bozos' unlimited financial resources, Smith and Jones each find that they no longer harbor the same prejudicial attitudes. They now see each other and the rest of their fellow workers, regardless of race or gender, as engaging in a common struggle while pursuing shared goals.

How would the MSP theorist account for these attitudinal changes facilitated through (more or less) optimal intergroup contact? Consider the change in attitudes that Smith undergoes. While mere exposure to Jones may influence her attitudes toward Jones and her fellow White workers (see Zajonc 1968), frequent positive exposure to Jones may result in the extinction of an amygdala-mediated threat response. The unexpected pep-talk might help to retrain Smith's striatal based model-free instrumental learning system, causing Smith to gradually associate interactions with Jones as having positive value, which is then reinforced by Smith's coming to Jones' defense. Her acknowledgement that her own economic gains, and therefore those of her family and community are inextricably linked to his, exerting a top-down influence on the learning rate of the instrumental learning system. We further can assume that their shared meals facilitate liking through appetitive reward and through the cultural understanding it affords. Finally, when Smith retrieves from episodic memory particular positive experiences of Jones while regaling her family with stories about her time on the picket line may further reinforce the positive affect that she associates with him and her fellow workers. I leave it as an exercise for the reader to consider how this form of intergroup contact may have influenced Jones attitudes toward Smith and his black comrades. Note that none of the monist accounts of attitudes is capable of telling a similarly rich story about the mechanisms that subserve the various forms of attitude change afforded by intergroup contact.

Lest one think that this story is too fanciful to describe actual forms of prejudice reduction, there is a growing body of research on the effects of various forms of bottom-up collective action on prejudice that corroborates this basic picture (see Dixon, Levine, Reicher & Durrheim 2012 for review). Of this body of research, one study offers a particularly vivid look at the various ways in which intergroup contact in the context of engaging in a common struggle reduces intergroup prejudice. Acar and Ulug (2016) interviewed a number of activists from marginalized groups who came together to protest a developer-friendly urban renewal project in Turkey:

> On May 27, 2013, a small number of activists gathered in Gezi Park after bulldozers arrived to cut down trees in the park. Over the next few days, more and more people, who witnessed what they considered disproportionate force on the part of the police, joined the

> activists. In the end, thousands gathered in the park and pushed out the developers and the police. Activists spent the next 15 days camped out in the park, where they created discussion forms, classes, cooked, cleaned, and lived together communally. The protests spread to 79 of 81 provinces in Turkey, without at least three million people participating in the protests around the country. […] The protests were unique in their ability to bring together a number of different groups whose central— and sometimes only— common ground was their opposition to the AKP [the ruling political party in Turkey at the time]. Religious and ethnic minorities (*i.e.* Kurds, Alevis, and Armenians had a large presence in the protest, as a part of and alongside leftist political parties and organizations (*e.g.,* socialist and main opposition parties), as well as nationalist and Kemalist-minded groups who, until the AKP came into power, maintained a more comfortable, advantaged position in society. Although participants had different backgrounds and different political positions, they were able to come together for a common cause, and managed to prevent the destruction of the part and resist injustice inflicted upon them. (*ibid*: 169)

Those interviewed described the ways in which the disparate, sometimes antagonistic groups, came to express solidarity with each other. For instance, one nationalistic group, members of which initially chanted homophobic slurs aimed at the police later stood in solidarity with the LGBT groups by chanting their pro-LGBT slogans in unison with the LGBT groups (*ibid.*). Indeed, the effects of these protests were found to have lasting effects long after the protests ended as indicated by the emergence of various leftist political coalitions and fronts composed of disparate groups that had participated in the Gezi Park demonstrations.

That MSP theory is particularly, if not uniquely, well-positioned to explain both the successes of the intergroup contact and collective action models of prejudice reduction and the failures of the various more targeted implicit bias interventions carries with it at least two important implications for the future of attitude research in particular and social cognition research more generally. First, it further bolsters the case for an unabashedly interdisciplinary approach to attitudes of the sort pursued in this dissertation and championed by Amodio (2019). Second, the field of social psychology has historically had an uneasy (to put it mildly) relationship with the very programs that the best available prejudice reduction models recommend for combatting various forms of prejudice. The fields' present preoccupation with the development of the kinds of failed intervention have helped to shield Fortune 500 companies from legal liability for racial and gender-based discrimination (Newkirk 2019). Moreover, corporations routinely hire social psychologists to screen out likely union members among job applicant and to administer surveys to existing employees to identify workplaces that are at risk of unionization (Lott 2013). The field's neglect of the effects of labor-based collective action on prejudice reduction is particularly appalling given people of color are disproportionately represented amongst the working class (Chowkwanyun & Reed, Jr. 2020) and the existence of evidence that present and past union members exhibit less racial prejudice relative to non-union members (Frymer & Grumbach 2020). In any case, the successes of collective action models of prejudice reduction taken together with a theoretical perspective on attitudes that accounts for their successes suggests that the field ought to turn now to the development of what we might call *solidarity* models of prejudice reduction and the implementation of the interventions that such models prescribe.

REFERENCES

Acar, Yasemin Gülsüm, and Özden Melis Uluğ. 2016. "Examining Prejudice Reduction through Solidarity and Togetherness Experiences among Gezi Park Activists in Turkey." *Journal of Social and Political Psychology* 4 (1): 166–79. https://doi.org/10.5964/jspp.v4i1.547.

Adelman, James S., and Zachary Estes. 2013. "Emotion and Memory: A Recognition Advantage for Positive and Negative Words Independent of Arousal." *Cognition* 129 (3): 530–35. https://doi.org/10.1016/j.cognition.2013.08.014.

Adolphs, Ralph. 2003. "Cognitive Neuroscience of Human Social Behaviour." *Nature Reviews Neuroscience* 4 (3): 165–78. https://doi.org/10.1038/nrn1056.

Alarabi, Khamis Faraj, Abdul Wahab, and Izza Karim. n.d. "Measuring Effectiveness of Memory Through Affect," 6.

Allen, Richard J. 2018. "Classic and Recent Advances in Understanding Amnesia." *F1000Research* 7 (March): 331. https://doi.org/10.12688/f1000research.13737.1.

Alter, Adam L., and Daniel M. Oppenheimer. 2009. "Uniting the Tribes of Fluency to Form a Metacognitive Nation." *Personality and Social Psychology Review* 13 (3): 219–35. https://doi.org/10.1177/1088868309341564.

Amodio, David M. 2014a. "Dual Experiences, Multiple Processes: Looking Beyond Dualities for Mechanisms of the Mind." In *Dual-Process Theories of the Social Mind*, edited by Jeffrey W. Sherman, Bertram Gawronski, and Yaacov Trope, 560–77. New York: The Guilford Press.

———. 2014b. "The Neuroscience of Prejudice and Stereotyping." *Nature Reviews Neuroscience* 15 (10): 670–82. https://doi.org/10.1038/nrn3800.

———. 2019a. "Social Cognition 2.0: An Interactive Memory Systems Account." *Trends in Cognitive Sciences* 23 (1): 21–33. https://doi.org/10.1016/j.tics.2018.10.002.

———. 2019b. "Applying Interdisciplinary Innovations to Advance Theories of Social Behavior: Response to Van Dessel and Colleagues." *Trends in Cognitive Sciences* 23 (6): 449–50. https://doi.org/10.1016/j.tics.2019.03.004.

Amodio, David M., and Jeffrey J. Berg. 2018. "Toward a Multiple Memory Systems Model of Attitudes and Social Cognition." *Psychological Inquiry* 29 (1): 14–19. https://doi.org/10.1080/1047840X.2018.1435620.

Amodio, David M., and Patricia G. Devine. 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): 652–61. https://doi.org/10.1037/0022-3514.91.4.652.

Amodio, David M., and Chris D. Frith. 2006. "Meeting of Minds: The Medial Frontal Cortex and Social Cognition." *Nature Reviews Neuroscience* 7 (4): 268–77. https://doi.org/10.1038/nrn1884.

Amodio, David M., and Holly K. Hamilton. 2012. "Intergroup Anxiety Effects on Implicit Racial Evaluation and Stereotyping." *Emotion* 12 (6): 1273–80. https://doi.org/10.1037/a0029016.

Amodio, David M., Eddie Harmon-Jones, and Patricia G. Devine. 2003. "Individual Differences in the Activation and Control of Affective Race Bias as Assessed by Startle Eyeblink Response and Self-Report." *Journal of Personality and Social Psychology* 84 (4): 738–53. https://doi.org/10.1037/0022-3514.84.4.738.

Amodio, David M., Eddie Harmon-Jones, Patricia G. Devine, John J. Curtin, Sigan L. Hartley, and Alison E. Covert. 2004. "Neural Signals for the Detection of Unintentional Race Bias." *Psychological Science* 15 (2): 88–93. https://doi.org/10.1111/j.0963-7214.2004.01502003.x.

Amodio, David M, and Christian Keysers. 2018. "Editorial Overview: New Advances in Social Neuroscience: From Neural Computations to Social Structures." *Current Opinion in Psychology* 24 (December): iv–vi. https://doi.org/10.1016/j.copsyc.2018.10.017.

Amodio, David M., and Kyle G. Ratner. 2011. "A Memory Systems Model of Implicit Social Cognition." *Current Directions in Psychological Science* 20 (3): 143–48. https://doi.org/10.1177/0963721411408562.

Amodio, David M., and Jillian K. Swencionis. 2018. "Proactive Control of Implicit Bias: A Theoretical Model and Implications for Behavior Change." *Journal of Personality and Social Psychology* 115 (2): 255–75. https://doi.org/10.1037/pspi0000128.

Anan, Bar. n.d. "RUNNING HEAD: RELATIONS AND SPEEDED EVALUATION," 35.

Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton, N.J: Princeton University Press.

Anderson, J. R. 1980. "Concepts, Propositions, and Schemata: What Are the Cognitive Units?" *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation* 28: 121–62.

Anderson, John R. 1976. *Language, Memory and Thought*. The Experimental Psychology Series. Hillsdale: Lawrence Erlbaum Associates.

———. 1996. *The Architecture of Cognition*. Mahwah, NJ: Lawrence Erlbaum Assos. Publ.

Andreatta, Marta, Andreas Mühlberger, Ayse Yarali, Bertram Gerber, and Paul Pauli. 2010. "A Rift between Implicit and Explicit Conditioned Valence in Human Pain Relief Learning." *Proceedings of the Royal Society B: Biological Sciences* 277 (1692): 2411–16. https://doi.org/10.1098/rspb.2010.0103.

Ashby, F. Gregory, Leola A. Alfonso-Reese, And U. Turken, and Elliott M. Waldron. 1998. "A Neuropsychological Theory of Multiple Systems in Category Learning." *Psychological Review* 105 (3): 442–81. https://doi.org/10.1037/0033-295X.105.3.442.

Atlas, Lauren Y. 2019. "How Instructions Shape Aversive Learning: Higher Order Knowledge, Reversal Learning, and the Role of the Amygdala." *Current Opinion in Behavioral Sciences* 26 (April): 121–29. https://doi.org/10.1016/j.cobeha.2018.12.008.

Atlas, Lauren Y, Bradley B Doll, Jian Li, Nathaniel D Daw, and Elizabeth A Phelps. 2016. "Instructed Knowledge Shapes Feedback-Driven Aversive Learning in Striatum and Orbitofrontal Cortex, but Not the Amygdala." *ELife* 5 (May): e15192. https://doi.org/10.7554/eLife.15192.

Atlas, Lauren Y., and Elizabeth A. Phelps. 2018. "Prepared Stimuli Enhance Aversive Learning without Weakening the Impact of Verbal Instructions." *Learning & Memory* 25 (2): 100–104. https://doi.org/10.1101/lm.046359.117.

"Awareness Is Essential for Differential Delay Eyeblink Conditioning with Soft-Tone but Not Loud-Tone Conditioned Stimuli | SpringerLink." n.d. Accessed January 30, 2020. https://link.springer.com/article/10.1007/s12264-013-1400-5.

Aylward, Jessica, Vincent Valton, Franziska Goer, Níall Lally, Sarah Peters, Tarun Limbachya, and Oliver J Robinson. n.d. "The Impact of Induced Anxiety on Affective Response Inhibition," 11.

Azevedo, Ruben T., Sarah N. Garfinkel, Hugo D. Critchley, and Manos Tsakiris. 2017. "Cardiac Afferent Activity Modulates the Expression of Racial Stereotypes." *Nature Communications* 8 (1): 13854. https://doi.org/10.1038/ncomms13854.

Baars, Bernard J. 2002. "The Conscious Access Hypothesis: Origins and Recent Evidence." *Trends in Cognitive Sciences* 6 (1): 47–52. https://doi.org/10.1016/S1364-6613(00)01819-2.

Baeyens, F., E. Díaz, and G. Ruiz. 2005. "Resistance to Extinction of Human Evaluative Conditioning Using a Between-Subjects Design." *Cognition & Emotion* 19 (2): 245–68. https://doi.org/10.1080/02699930441000300.

Baeyens, Frank, Geert Crombez, Omer Van den Bergh, and Paul Eelen. 1988. "Once in Contact Always in Contact: Evaluative Conditioning Is Resistant to Extinction." *Advances in Behaviour Research and Therapy* 10 (4): 179–99. https://doi.org/10.1016/0146-6402(88)90014-8.

Ballard, Ian, Eric M. Miller, Steven T. Piantadosi, Noah Goodman, and Samuel M. McClure. 2017. "Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning." Preprint. Neuroscience. https://doi.org/10.1101/115253.

Banaji, Mahzarin R., and Anthony G. Greenwald. 2013. *Blindspot: Hidden Biases of Good People*. New York: Delacorte Press.

Banaji, Mahzarin R., and Curtis D. Hardin. 1996. "Automatic Stereotyping." *Psychological Science* 7 (3): 136–41. https://doi.org/10.1111/j.1467-9280.1996.tb00346.x.

Bar-Anan, Yoav, and Tal Moran. 2018. "Simple First: A Skeleton for an Evaluative Learning Model." *Social Psychological Bulletin* 13 (3): e28761. https://doi.org/10.5964/spb.v13i3.28761.

Bar-Anan, Yoav, and Brian A. Nosek. 2014. "A Comparative Investigation of Seven Indirect Attitude Measures." *Behavior Research Methods* 46 (3): 668–88. https://doi.org/10.3758/s13428-013-0410-6.

Bar-Anan, Yoav, and Brian A Nosek. n.d. "A Comparison of the Sensitivity of Four Indirect Evaluation Measures to Evaluative Information," 37.

Bargh, John A. 2016. "Awareness of the Prime versus Awareness of Its Influence: Implications for the Real-World Scope of Unconscious Higher Mental Processes." *Current Opinion in Psychology* 12 (December): 49–52. https://doi.org/10.1016/j.copsyc.2016.05.006.

Bargh, John A., Shelly Chaiken, Paula Raymond, and Charles Hymes. 1996. "The Automatic Evaluation Effect: Unconditional Automatic Attitude Activation with a Pronunciation Task." *Journal of Experimental Social Psychology* 32 (1): 104–28. https://doi.org/10.1006/jesp.1996.0005.

Barnes-Holmes, Dermot, and Ian Hussey. 2016. “The Functional-Cognitive Meta-Theoretical Framework: Reflections, Possible Clarifications and How to Move Forward: REFLECTIONS AND CLARIFICATIONS.” *International Journal of Psychology* 51 (1): 50–57. https://doi.org/10.1002/ijop.12166.

Barrett, H. Clark, and James Broesch. 2012. “Prepared Social Learning about Dangerous Animals in Children.” *Evolution and Human Behavior* 33 (5): 499–508. https://doi.org/10.1016/j.evolhumbehav.2012.01.003.

Barrett, Lisa Feldman. 1998. “Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus.” *Cognition & Emotion* 12 (4): 579–99. https://doi.org/10.1080/026999398379574.

———. 2006. “Valence Is a Basic Building Block of Emotional Life.” *Journal of Research in Personality* 40 (1): 35–55. https://doi.org/10.1016/j.jrp.2005.08.006.

Bartholow, Bruce D, and Cheryl L Dickter. n.d. “Social Cognitive Neuroscience of Person Perception: A Selective Review Focused on the Event-Related Brain Potential.,” 27.

Bassett, Danielle S, and Olaf Sporns. 2017. “Network Neuroscience.” *Nature Neuroscience* 20 (3): 353–64. https://doi.org/10.1038/nn.4502.

Bauer, Patricia J, Tracy Deboer, and Angela F Lukowski. n.d. “In the Language of Multiple Memory Systems,” 31.

Bayley, P.J., R.C. O’Reilly, T. Curran, and L.R. Squire. 2008. “New Semantic Learning in Patients with Large Medial Temporal Lobe Lesions.” *Hippocampus* 18 (6): 575–83. https://doi.org/10.1002/hipo.20417.

Bechara, Antoine, Hanna Damasio, and Antonio R. Damasio. 2006. “Role of the Amygdala in Decision-Making.” *Annals of the New York Academy of Sciences* 985 (1): 356–69. https://doi.org/10.1111/j.1749-6632.2003.tb07094.x.

Bechara, Antoine, Hanna Damasio, Antonio R. Damasio, and Gregory P. Lee. 1999. "Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making." *The Journal of Neuroscience* 19 (13): 5473–81. https://doi.org/10.1523/JNEUROSCI.19-13-05473.1999.

Bechtel, William. 2005. "THE CHALLENGE OF CHARACTERIZING OPERATIONS IN THE MECHANISMS UNDERLYING BEHAVIOR." *Journal of the Experimental Analysis of Behavior* 84 (3): 313–25. https://doi.org/10.1901/jeab.2005.103-04.

Beer, Jennifer S., and Kevin N. Ochsner. 2006. "Social Cognition: A Multi Level Analysis." *Brain Research* 1079 (1): 98–105. https://doi.org/10.1016/j.brainres.2006.01.002.

Benedetti, Fabrizio, Antonella Pollo, Leonardo Lopiano, Michele Lanotte, Sergio Vighetti, and Innocenzo Rainero. 2003. "Conscious Expectation and Unconscious Conditioning in Analgesic, Motor, and Hormonal Placebo/Nocebo Responses." *The Journal of Neuroscience* 23 (10): 4315–23. https://doi.org/10.1523/JNEUROSCI.23-10-04315.2003.

Benedict, Taylor, Jasmin Richter, and Anne Gast. 2019. "The Influence of Misinformation Manipulations on Evaluative Conditioning." *Acta Psychologica* 194 (March): 28–36. https://doi.org/10.1016/j.actpsy.2019.01.014.

Benishek, Lauren E., Sallie J. Weaver, and David E. Newman-Toker. 2015. "The Cognitive Psychology of Diagnostic Errors." *DeckerMed Neurology*, November. https://doi.org/10.2310/NEURO.6288.

Bennett, Marc, Ellen Vervoort, Yannick Boddez, Dirk Hermans, and Frank Baeyens. 2015. "Perceptual and Conceptual Similarities Facilitate the Generalization of Instructed Fear." *Journal of Behavior Therapy and Experimental Psychiatry* 48 (September): 149–55. https://doi.org/10.1016/j.jbtep.2015.03.011.

Berger, Christopher C., Tara C. Dennehy, John A. Bargh, and Ezequiel Morsella. 2016. "Nisbett and Wilson (1977) Revisited: The Little That We Can Know and Can Tell." *Social Cognition* 34 (3): 167–95. https://doi.org/10.1521/soco.2016.34.3.167.

Berger, Jacob. 2018. "Implicit Attitudes and Awareness." *Synthese*, March. https://doi.org/10.1007/s11229-018-1754-3.

Bergström, Fredrik, and Johan Eriksson. 2018. "Neural Evidence for Non-Conscious Working Memory." *Cerebral Cortex* 28 (9): 3217–28. https://doi.org/10.1093/cercor/bhx193.

Berthoz, S., J. Grèzes, J.L. Armony, R.E. Passingham, and R.J. Dolan. 2006. "Affective Response to One's Own Moral Violations." *NeuroImage* 31 (2): 945–50. https://doi.org/10.1016/j.neuroimage.2005.12.039.

Bicchieri, Cristina. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York, NY: Oxford University Press.

Biderman, Natalie, and Liad Mudrik. 2018. "Evidence for Implicit—But Not Unconscious—Processing of Object-Scene Relations." *Psychological Science* 29 (2): 266–77. https://doi.org/10.1177/0956797617735745.

Binder, Jeffrey R., and Rutvik H. Desai. 2011. "The Neurobiology of Semantic Memory." *Trends in Cognitive Sciences* 15 (11): 527–36. https://doi.org/10.1016/j.tics.2011.10.001.

Binney, Richard J, and Richard Ramsey. 2019. "Social Semantics: The Role of Conceptual Knowledge and Cognitive Control in a Neurobiological Model of the Social Brain." Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/36tm5.

Blair, Irene V., Jennifer E. Ma, and Alison P. Lenton. 2001. "Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery." *Journal of Personality and Social Psychology* 81 (5): 828–41. https://doi.org/10.1037/0022-3514.81.5.828.

Block, Ned. 2011. “Perceptual Consciousness Overflows Cognitive Access.” *Trends in Cognitive Sciences* 15 (12): 567–75. https://doi.org/10.1016/j.tics.2011.11.001.

Blumenthal, Anna, Devin Duke, Ben Bowles, Asaf Gilboa, R. Shayna Rosenbaum, Stefan Köhler, and Ken McRae. 2017. “Abnormal Semantic Knowledge in a Case of Developmental Amnesia.” *Neuropsychologia* 102 (July): 237–47. https://doi.org/10.1016/j.neuropsychologia.2017.06.018.

Bonner, M. F., and A. R. Price. 2013. “Where Is the Anterior Temporal Lobe and What Does It Do?” *Journal of Neuroscience* 33 (10): 4213–15. https://doi.org/10.1523/JNEUROSCI.0041-13.2013.

Borch, Casey, Vincent Buskens, Steve E Clayman, Douglas D Heckathorn, Jeffrey A Houser, Edward J Lawler, John M Levine, *et al.* n.d. “LIST OF CONTRIBUTORS,” 257.

Borsboom, Denny. n.d. “The Philosophy of Psychometrics,” 85.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2004. “The Concept of Validity.” *Psychological Review* 111 (4): 1061–71. https://doi.org/10.1037/0033-295X.111.4.1061.

Bosson, Jennifer K., William B. Swann Jr., and James W. Pennebaker. 2000. “Stalking the Perfect Measure of Implicit Self-Esteem: The Blind Men and the Elephant Revisited?” *Journal of Personality and Social Psychology* 79 (4): 631–43. https://doi.org/10.1037/0022-3514.79.4.631.

Bostan, Andreea C., and Peter L. Strick. 2018. “The Basal Ganglia and the Cerebellum: Nodes in an Integrated Network.” *Nature Reviews Neuroscience* 19 (6): 338–50. https://doi.org/10.1038/s41583-018-0002-7.

Bouret, Sebastien, and Barry J. Richmond. 2010. “Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys.” *The Journal of Neuroscience* 30 (25): 8591–8601. https://doi.org/10.1523/JNEUROSCI.0049-10.2010.

Bourguignon, Nicolas J., Senne Braem, Egbert Hartstra, Jan De Houwer, and Marcel Brass. 2018. "Encoding of Novel Verbal Instructions for Prospective Action in the Lateral Prefrontal Cortex: Evidence from Univariate and Multivariate Functional Magnetic Resonance Imaging Analysis." *Journal of Cognitive Neuroscience* 30 (8): 1170–84. https://doi.org/10.1162/jocn_a_01270.

Bouton, M. E. 2004. "Context and Behavioral Processes in Extinction." *Learning & Memory* 11 (5): 485–94. https://doi.org/10.1101/lm.78804.

Bowman, Caitlin R., and Dagmar Zeithamova. 2018. "Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization." *The Journal of Neuroscience* 38 (10): 2605–14. https://doi.org/10.1523/JNEUROSCI.2811-17.2018.

Boyd, Richard. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61 (1–2): 127–48. https://doi.org/10.1007/BF00385837.

Breckler, Steven J. n.d. "Empirical Validation of Affect, Behavior, and Cognition as Distinct Components of Attitude," 15.

Brooks, Thom. 2010. "Guidelines on How to Referee." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1719043.

Brown-Iannuzzi, Jazmin L., Erin Cooley, Stephanie E. McKee, and Charly Hyden. 2019. "Wealthy Whites and Poor Blacks: Implicit Associations between Racial Groups and Wealth Predict Explicit Opposition toward Helping the Poor." *Journal of Experimental Social Psychology* 82 (May): 26–34. https://doi.org/10.1016/j.jesp.2018.11.006.

Brownstein, Michael. 2018a. *Caring, Implicit Attitudes, and the Self*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190633721.003.0004.

———. 2018b. *Deliberation and Spontaneity*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190633721.003.0006.

———. 2018c. *Reflection, Responsibility, and Fractured Selves*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190633721.003.0005.

———. 2018d. *The Implicit Mind*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190633721.001.0001.

Brownstein, Michael, and Alex Madva. 2012. "The Normativity of Automaticity: The Normativity of Automaticity." *Mind & Language* 27 (4): 410–34. https://doi.org/10.1111/j.1468-0017.2012.01450.x.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2019. "What Do Implicit Measures Measure?" *Wiley Interdisciplinary Reviews: Cognitive Science*, April, e1501. https://doi.org/10.1002/wcs.1501.

———. n.d. "Understanding Implicit Bias: How the Critics Miss the Point," 26.

Bublatzky, Florian, and Harald T. Schupp. 2012. "Pictures Cueing Threat: Brain Dynamics in Viewing Explicitly Instructed Danger Cues." *Social Cognitive and Affective Neuroscience* 7 (6): 611–22. https://doi.org/10.1093/scan/nsr032.

Buckholtz, Joshua W, and René Marois. 2012. "The Roots of Modern Justice: Cognitive and Neural Foundations of Social Norms and Their Enforcement." *Nature Neuroscience* 15 (5): 655–61. https://doi.org/10.1038/nn.3087.

Buckner, Cameron. 2015. "A Property Cluster Theory of Cognition." *Philosophical Psychology* 28 (3): 307–36. https://doi.org/10.1080/09515089.2013.843274.

———. n.d. "Two Approaches to the Distinction between Cognition and 'Mere Association,'" 36.

Cacciaglia, Raffaele, Sebastian T. Pohlack, Herta Flor, and Frauke Nees. 2015. "Dissociable Roles for Hippocampal and Amygdalar Volume in Human Fear Conditioning." *Brain Structure and Function* 220 (5): 2575–86. https://doi.org/10.1007/s00429-014-0807-8.

Calvo, Paco, and John Symons, eds. 2014. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, Massachusetts: The MIT Press.

Cameron, C. Daryl, Justin Reber, Victoria L. Spring, and Daniel Tranel. 2018. "Damage to the Ventromedial Prefrontal Cortex Is Associated with Impairments in Both Spontaneous and Deliberative Moral Judgments." *Neuropsychologia* 111 (March): 261–68. https://doi.org/10.1016/j.neuropsychologia.2018.01.038.

Cameron, Gemma, Bryan Roche, Michael W. Schlund, and Simon Dymond. 2016. "Learned, Instructed and Observed Pathways to Fear and Avoidance." *Journal of Behavior Therapy and Experimental Psychiatry* 50 (March): 106–12. https://doi.org/10.1016/j.jbtep.2015.06.003.

Carey, Susan. 1986. "Cognitive Science and Science Education." *American Psychologist*, 8.

Carlston, Donal E., ed. 2013. *The Oxford Handbook of Social Cognition*. Oxford ; New York: Oxford University Press.

Carrabine, Eamonn, ed. 2002. *Crime in Modern Britain*. Oxford Modern Britain. Oxford ; New York: Oxford University Press.

Carruthers, Peter. 2013. "On Knowing Your Own Beliefs: A Representationalist Account." In *New Essays on Belief*, edited by Nikolaj Nottelmann, 145–65. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137026521_8.

———. 2018a. "Episodic Memory Isn't Essentially Autonoetic." *Behavioral and Brain Sciences* 41: e6. https://doi.org/10.1017/S0140525X17001285.

———. 2018b. "Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures?" *Review of Philosophy and Psychology* 9 (1): 51–72. https://doi.org/10.1007/s13164-017-0354-3.

Cartoni, Emilio, Bernard Balleine, and Gianluca Baldassarre. 2016. "Appetitive Pavlovian-Instrumental Transfer: A Review." *Neuroscience & Biobehavioral Reviews* 71 (December): 829–48. https://doi.org/10.1016/j.neubiorev.2016.09.020.

Cassam, Quassim. 2017. "Diagnostic Error, Overconfidence and Self-Knowledge." *Palgrave Communications* 3 (1): 17025. https://doi.org/10.1057/palcomms.2017.25.

Cassidy, Brittany S., and Angela H. Gutchess. 2015. "Influences of Appearance-Behavior Congruity on Memory and Social Judgments." *Memory (Hove, England)* 23 (7): 1039–55. https://doi.org/10.1080/09658211.2014.951364.

Chaaya, Nicholas, Andrew R. Battle, and Luke R. Johnson. 2018. "An Update on Contextual Fear Memory Mechanisms: Transition between Amygdala and Hippocampus." *Neuroscience & Biobehavioral Reviews* 92 (September): 43–54. https://doi.org/10.1016/j.neubiorev.2018.05.013.

"Changing Our Minds: The Neural Bases of Dynamic Impression Updating | Elsevier Enhanced Reader." n.d. Accessed February 11, 2020. https://doi.org/10.1016/j.copsyc.2018.08.007.

Chekroud, Adam M., Jim A. C. Everett, Holly Bridge, and Miles Hewstone. 2014. "A Review of Neuroimaging Studies of Race-Related Prejudice: Does Amygdala Response Reflect Threat?" *Frontiers in Human Neuroscience* 8 (March). https://doi.org/10.3389/fnhum.2014.00179.

Chen, Chong, Yuki Omiya, and Si Yang. 2015. "Dissociating Contributions of Ventral and Dorsal Striatum to Reward Learning." *Journal of Neurophysiology* 114 (3): 1364–66. https://doi.org/10.1152/jn.00873.2014.

Chen, Patricia M. 2019. "Housing First and Single-Site Housing." *Social Sciences* 8 (4): 129. https://doi.org/10.3390/socsci8040129.

Chiao, Juan Y., ed. 2009. *Cultural Neuroscience: Cultural Influences on Brain Function*. 1. ed. Progress in Brain Research 178. New York, NY: Elsevier.

Chmielewski, Witold X., and Christian Beste. 2019. "Neurophysiological Mechanisms Underlying the Modulation of Cognitive Control by Simultaneous Conflicts." *Cortex* 115 (June): 216–30. https://doi.org/10.1016/j.cortex.2019.02.006.

Christiansen, Morten H. 2019. "Implicit Statistical Learning: A Tale of Two Literatures." *Topics in Cognitive Science* 11 (3): 468–81. https://doi.org/10.1111/tops.12332.

Chudek, Maciej, and Joseph Henrich. 2011. "Culture–Gene Coevolution, Norm-Psychology and the Emergence of Human Prosociality." *Trends in Cognitive Sciences* 15 (5): 218–26. https://doi.org/10.1016/j.tics.2011.03.003.

Chudek, Maciek, Wanying Zhao, and Joseph Henrich. 2013. "Culture-Gene Coevolution, Large-Scale Cooperation, and the Shaping of Human Social Psychology." In *Cooperation and Its Evolution*, edited by Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser, 425. MIT Press.

Claire, Mayor-Dubois, Deglise Sophie, Poloni Claudia, Maeder Philippe, and Roulet-Perez Eliane. 2016. "Verbal Emotional Memory in a Case with Left Amygdala Damage." *Neurocase* 22 (1): 45–54. https://doi.org/10.1080/13554794.2015.1037843.

Clark Barrett, H., Christopher D. Peterson, and Willem E. Frankenhuis. 2016. "Mapping the Cultural Learnability Landscape of Danger." *Child Development* 87 (3): 770–81. https://doi.org/10.1111/cdev.12495.

Clarke, Alex, and Lorraine K. Tyler. 2014. "Object-Specific Semantic Coding in Human Perirhinal Cortex." *The Journal of Neuroscience* 34 (14): 4766–75. https://doi.org/10.1523/JNEUROSCI.2828-13.2014.

"Climate Change and Individual Responsibility." n.d., 21.

Cohen, Anna-Lisa, and Jason L. Hicks. 2017. “Implementation Intentions.” In *Prospective Memory: Remembering to Remember, Remembering to Forget*, edited by Anna-Lisa Cohen and Jason L. Hicks, 81–97. SpringerBriefs in Psychology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68990-6_5.

Cole, Michael W., Patryk Laurent, and Andrea Stocco. 2013. “Rapid Instructed Task Learning: A New Window into the Human Brain’s Unique Capacity for Flexible Cognitive Control.” *Cognitive, Affective, & Behavioral Neuroscience* 13 (1): 1–22. https://doi.org/10.3758/s13415-012-0125-7.

Cole, Scott N., Catriona M. Morrison, Ohr Barak, Katalin Pauly-Takacs, and Martin A. Conway. 2016. “Amnesia and Future Thinking: Exploring the Role of Memory in the Quantity and Quality of Episodic Future Thoughts.” *British Journal of Clinical Psychology* 55 (2): 206–24. https://doi.org/10.1111/bjc.12094.

Collins-Chobanian, Shari, ed. 2017. *Being Ethical: Classic and New Voices on Contemporary Issues*. Peterborough, Ontario: Broadview Press.

Cone, Jeremy, and Melissa J. Ferguson. 2015. “He Did What? The Role of Diagnosticity in Revising Implicit Evaluations.” *Journal of Personality and Social Psychology* 108 (1): 37–57. https://doi.org/10.1037/pspa0000014.

Cone, Jeremy, Kathryn Flaharty, and Melissa J. Ferguson. 2019. “Believability of Evidence Matters for Correcting Social Impressions.” *Proceedings of the National Academy of Sciences* 116 (20): 9802–7. https://doi.org/10.1073/pnas.1903222116.

Cone, Jeremy, Thomas C. Mann, and Melissa J. Ferguson. 2017. “Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised.” In *Advances in Experimental Social Psychology*, 56:131–99. Elsevier. https://doi.org/10.1016/bs.aesp.2017.03.001.

Connor, David A., and Thomas J. Gould. 2016. “The Role of Working Memory and Declarative Memory in Trace Conditioning.” *Neurobiology of Learning and Memory* 134 (October): 193–209. https://doi.org/10.1016/j.nlm.2016.07.009.

Conrey, Frederica R., Jeffrey W. Sherman, Bertram Gawronski, Kurt Hugenberg, and Carla J. Groom. 2005. “Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance.” *Journal of Personality and Social Psychology* 89 (4): 469–87. https://doi.org/10.1037/0022-3514.89.4.469.

Conrey, Frederica R., and Eliot R. Smith. 2007. “Attitude Representation: Attitudes as Patterns in a Distributed, Connectionist Representational System.” *Social Cognition* 25 (5): 718–35. https://doi.org/10.1521/soco.2007.25.5.718.

Constantinescu, Alexandra O., Jill X. O’Reilly, and Timothy E. J. Behrens. 2016. “Organizing Conceptual Knowledge in Humans with a Grid-like Code.” *Science (New York, N.Y.)* 352 (6292): 1464–68. https://doi.org/10.1126/science.aaf0941.

Cooley, Erin, and B. Keith Payne. 2017. “Using Groups to Measure Intergroup Prejudice.” *Personality and Social Psychology Bulletin* 43 (1): 46–59. https://doi.org/10.1177/0146167216675331.

Cooley, Erin, B. Keith Payne, Chris Loersch, and Ryan Lei. 2015. “Who Owns Implicit Attitudes? Testing a Metacognitive Perspective.” *Personality and Social Psychology Bulletin* 41 (1): 103–15. https://doi.org/10.1177/0146167214559712.

Cooper, Elisa, Andrea Greve, and Richard N. Henson. 2019. “Investigating Fast Mapping Task Components: No Evidence for the Role of Semantic Referent nor Semantic Inference in Healthy Adults.” *Frontiers in Psychology* 10 (March): 394. https://doi.org/10.3389/fpsyg.2019.00394.

Corneille, Olivier, and Christoph Stahl. 2019. "Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models." *Personality and Social Psychology Review* 23 (2): 161–89. https://doi.org/10.1177/1088868318763261.

Correll, Joshua, Sean M. Hudson, Steffanie Guillermo, and Debbie S. Ma. 2014. "The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot: The Police Officer's Dilemma." *Social and Personality Psychology Compass* 8 (5): 201–13. https://doi.org/10.1111/spc3.12099.

Correll, Joshua, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. 2007. "The Influence of Stereotypes on Decisions to Shoot." *European Journal of Social Psychology* 37 (6): 1102–17. https://doi.org/10.1002/ejsp.450.

Costa, Vincent D., Margaret M. Bradley, and Peter J. Lang. 2015. "From Threat to Safety: Instructed Reversal of Defensive Reactions: Reversing Defensive Reactions." *Psychophysiology* 52 (3): 325–32. https://doi.org/10.1111/psyp.12359.

Crocker, Linda, and James Algina. 2008. *Introduction to Classical and Modern Test Theory*. Mason, Ohio: Cengage Learning.

Crockett, Molly J. 2013. "Models of Morality." *Trends in Cognitive Sciences* 17 (8): 363–66. https://doi.org/10.1016/j.tics.2013.06.005.

Crockett, Molly J, Jenifer Z Siegel, Zeb Kurth-Nelson, Peter Dayan, and Raymond J Dolan. 2017. "Moral Transgressions Corrupt Neural Representations of Value." *Nature Neuroscience* 20 (6): 879–85. https://doi.org/10.1038/nn.4557.

Cunningham, William A., Kristopher J. Preacher, and Mahzarin R. Banaji. 2001. "Implicit Attitude Measures: Consistency, Stability, and Convergent Validity." *Psychological Science* 12 (2): 163–70. https://doi.org/10.1111/1467-9280.00328.

Cunningham, William A., and Philip David Zelazo. 2007. "Attitudes and Evaluations: A Social Cognitive Neuroscience Perspective." *Trends in Cognitive Sciences* 11 (3): 97–104. https://doi.org/10.1016/j.tics.2006.12.005.

Dacey, Mike. 2018. "Simplicity and the Meaning of Mental Association." *Erkenntnis*, May. https://doi.org/10.1007/s10670-018-0005-9.

Das, Ramon. 2014. "Has Industrialization Benefited No One? Climate Change and the Non-Identity Problem." *Ethical Theory and Moral Practice* 17 (4): 747–59. https://doi.org/10.1007/s10677-013-9479-3.

Dasgupta, Nilanjana. 2013. "Implicit Attitudes and Beliefs Adapt to Situations." In *Advances in Experimental Social Psychology*, 47:233–79. Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00005-X.

Davis, M, and P J Whalen. 2001. "The Amygdala: Vigilance and Emotion." *Molecular Psychiatry* 6 (1): 13–34. https://doi.org/10.1038/sj.mp.4000812.

Davis, Steven. n.d. "Connectionism : Theory and Practice," 331.

Daw, Nathaniel D., and John P. O'Doherty. 2014. "Multiple Systems for Value Learning." In *Neuroeconomics*, 393–410. Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00021-8.

Dayan, Peter, and Kent C. Berridge. 2014. "Model-Based and Model-Free Pavlovian Reward Learning: Revaluation, Revision, and Revelation." *Cognitive, Affective, & Behavioral Neuroscience* 14 (2): 473–92. https://doi.org/10.3758/s13415-014-0277-8.

De Houwer, Jan. 2007. "A Conceptual and Theoretical Analysis of Evaluative Conditioning." *The Spanish Journal of Psychology* 10 (2): 230–41. https://doi.org/10.1017/S1138741600006491.

———. 2009. "The Propositional Approach to Associative Learning as an Alternative for Association Formation Models." *Learning & Behavior* 37 (1): 1–20. https://doi.org/10.3758/LB.37.1.1.

———. 2018. “Propositional Models of Evaluative Conditioning.” *Social Psychological Bulletin* 13 (3): e28046. https://doi.org/10.5964/spb.v13i3.28046.

De Houwer, Jan, Bertram Gawronski, and Dermot Barnes-Holmes. 2013. “A Functional-Cognitive Framework for Attitude Research.” *European Review of Social Psychology* 24 (1): 252–87. https://doi.org/10.1080/10463283.2014.892320.

De Houwer, Jan, and Sean Hughes. 2016. “Evaluative Conditioning as a Symbolic Phenomenon: On the Relation between Evaluative Conditioning, Evaluative Conditioning via Instructions, and Persuasion.” *Social Cognition* 34 (5): 480–94. https://doi.org/10.1521/soco.2016.34.5.480.

De Houwer, Jan, Pieter Van Dessel, and Tal Moran. 2020a. “Attitudes beyond Associations: On the Role of Propositional Representations in Stimulus Evaluation.” In *Advances in Experimental Social Psychology*, 61:127–83. Elsevier. https://doi.org/10.1016/bs.aesp.2019.09.004.

———. 2020b. “Chapter Three - Attitudes beyond Associations: On the Role of Propositional Representations in Stimulus Evaluation.” In *Advances in Experimental Social Psychology*, edited by Bertram Gawronski, 61:127–83. Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.004.

De Renzi, Ennio, Mario Liotti, and Paolo Nichelli. 1987. “Semantic Amnesia with Preservation of Autobiographic Memory. A Case Report.” *Cortex* 23 (4): 575–97. https://doi.org/10.1016/S0010-9452(87)80050-3.

Demanet, Jelle, Baptist Liefooghe, Egbert Hartstra, Dorit Wenke, Jan De Houwer, and Marcel Brass. 2016. “There Is More into ‘Doing’ than ‘Knowing’: The Function of the Right Inferior Frontal Sulcus Is Specific for Implementing versus Memorising Verbal Instructions.” *NeuroImage* 141 (November): 350–56. https://doi.org/10.1016/j.neuroimage.2016.07.059.

Descartes, René, John Cottingham, and Bernard Arthur Owen Williams. n.d. "Meditations on First Philosophy," 172.

Di Gregorio, Francesco, Martin E. Maier, and Marco Steinhauser. 2018. "Errors Can Elicit an Error Positivity in the Absence of an Error Negativity: Evidence for Independent Systems of Human Error Monitoring." *NeuroImage* 172 (May): 427–36. https://doi.org/10.1016/j.neuroimage.2018.01.081.

Dieciuc, Michael A., and Jonathan R. Folstein. 2019. "Typicality: Stable Structures and Flexible Functions." *Psychonomic Bulletin & Review* 26 (2): 491–505. https://doi.org/10.3758/s13423-018-1546-2.

Dienes, Zoltán, and Dianne Berry. 1997. "Implicit Learning: Below the Subjective Threshold." *Psychonomic Bulletin & Review* 4 (1): 3–23. https://doi.org/10.3758/BF03210769.

Dienes, Zoltan, and Anil K. Seth. 2010. "Measuring Any Conscious Content versus Measuring the Relevant Conscious Content: Comment on Sandberg *et al.*" *Consciousness and Cognition* 19 (4): 1079–80. https://doi.org/10.1016/j.concog.2010.03.009.

Dixon, John, Mark Levine, Steve Reicher, and Kevin Durrheim. 2012. "Beyond Prejudice: Are Negative Evaluations the Problem and Is Getting Us to like One Another More the Solution?" *Behavioral and Brain Sciences* 35 (6): 411–25. https://doi.org/10.1017/S0140525X11002214.

Dodell-Feder, David, Kerry J. Ressler, and Laura T. Germine. 2020. "Social Cognition or Social Class and Culture? On the Interpretation of Differences in Social Cognitive Performance." *Psychological Medicine* 50 (1): 133–45. https://doi.org/10.1017/S003329171800404X.

Dolcos, Florin, Yuta Katsumi, Matthew Moore, Nick Berggren, Beatrice de Gelder, Nazanin Derakshan, Alfons O. Hamm, *et al.* 2020. "Neural Correlates of Emotion-Attention Interactions: From Perception, Learning, and Memory to Social Cognition, Individual

Differences, and Training Interventions." *Neuroscience & Biobehavioral Reviews* 108 (January): 559–601. https://doi.org/10.1016/j.neubiorev.2019.08.017.

Doll, Bradley B., W. Jake Jacobs, Alan G. Sanfey, and Michael J. Frank. 2009. "Instructional Control of Reinforcement Learning: A Behavioral and Neurocomputational Investigation." *Brain Research* 1299 (November): 74–94. https://doi.org/10.1016/j.brainres.2009.07.007.

Don, Hilary J., Micah B. Goldwater, A. Ross Otto, and Evan J. Livesey. 2016. "Rule Abstraction, Model-Based Choice, and Cognitive Reflection." *Psychonomic Bulletin & Review* 23 (5): 1615–23. https://doi.org/10.3758/s13423-016-1012-y.

Doumas, Leonidas A. A., and John E. Hummel. 2012. *Computational Models of Higher Cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0005.

Dovidio, John F., and Samuel L. Gaertner. 2004. "Aversive Racism." In *Advances in Experimental Social Psychology*, 36:1–52. Elsevier. https://doi.org/10.1016/S0065-2601(04)36001-6.

Doya, Kenji, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. 2002. "Multiple Model-Based Reinforcement Learning." *Neural Computation* 14 (6): 1347–69. https://doi.org/10.1162/089976602753712972.

Duckworth, Kimberly L., John A. Bargh, Magda Garcia, and Shelly Chaiken. 2002. "The Automatic Evaluation of Novel Stimuli." *Psychological Science* 13 (6): 513–19. https://doi.org/10.1111/1467-9280.00490.

Duncan, John, Paul Burgess, and Hazel Emslie. 1995. "Fluid Intelligence after Frontal Lobe Lesions." *Neuropsychologia* 33 (3): 261–68. https://doi.org/10.1016/0028-3932(94)00124-8.

Duncan, Seth, and Lisa Feldman Barrett. 2007. "Affect Is a Form of Cognition: A Neurobiological Analysis." *Cognition & Emotion* 21 (6): 1184–1211. https://doi.org/10.1080/02699930701437931.

Dunne, Simon, Arun D'Souza, and John P. O'Doherty. 2016. "The Involvement of Model-Based but Not Model-Free Learning Signals during Observational Reward Learning in the Absence of Choice." *Journal of Neurophysiology* 115 (6): 3195–3203. https://doi.org/10.1152/jn.00046.2016.

Dunsmoor, Joseph E., Jennifer T. Kubota, Jian Li, Cesar A.O. Coelho, and Elizabeth A. Phelps. 2016. "Racial Stereotypes Impair Flexibility of Emotional Learning." *Social Cognitive and Affective Neuroscience* 11 (9): 1363–73. https://doi.org/10.1093/scan/nsw053.

Durante, Federica, and Susan T Fiske. 2017. "How Social-Class Stereotypes Maintain Inequality." *Current Opinion in Psychology* 18 (December): 43–48. https://doi.org/10.1016/j.copsyc.2017.07.033.

Duss, Simone B., Thomas P. Reber, Jürgen Hänggi, Simon Schwab, Roland Wiest, René M. Müri, Peter Brugger, Klemens Gutbrod, and Katharina Henke. 2014. "Unconscious Relational Encoding Depends on Hippocampus." *Brain* 137 (12): 3355–70. https://doi.org/10.1093/brain/awu270.

Dymond, Simon, Michael W. Schlund, Bryan Roche, Jan De Houwer, and Gary P. Freegard. 2012. "Safe From Harm: Learned, Instructed, and Symbolic Generalization Pathways of Human Threat-Avoidance." Edited by Reginald Frederick Westbrook. *PLoS ONE* 7 (10): e47539. https://doi.org/10.1371/journal.pone.0047539.

Dzieciol, Anna M., Jocelyne Bachevalier, Kadharbatcha S. Saleem, David G. Gadian, Richard Saunders, W.K. Kling Chong, Tina Banks, Mortimer Mishkin, and Faraneh Vargha-Khadem. 2017. "Hippocampal and Diencephalic Pathology in Developmental Amnesia." *Cortex* 86 (January): 33–44. https://doi.org/10.1016/j.cortex.2016.09.016.

Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies. 2004. "Seeing Black: Race, Crime, and Visual Processing." *Journal of Personality and Social Psychology* 87 (6): 876–93. https://doi.org/10.1037/0022-3514.87.6.876.

Economides, Marcos, Zeb Kurth-Nelson, Annika Lübbert, Marc Guitart-Masip, and Raymond J. Dolan. 2015. "Model-Based Reasoning in Humans Becomes Automatic with Training." Edited by Olaf Sporns. *PLOS Computational Biology* 11 (9): e1004463. https://doi.org/10.1371/journal.pcbi.1004463.

Eder, Andreas B., Hartmut Leuthold, Klaus Rothermund, and Stefan R. Schweinberger. 2012. "Automatic Response Activation in Sequential Affective Priming: An ERP Study." *Social Cognitive and Affective Neuroscience* 7 (4): 436–45. https://doi.org/10.1093/scan/nsr033.

Eichenbaum, Howard. 2017a. "Memory: Organization and Control." *Annual Review of Psychology* 68 (1): 19–45. https://doi.org/10.1146/annurev-psych-010416-044131.

———. 2017b. "Prefrontal–Hippocampal Interactions in Episodic Memory." *Nature Reviews Neuroscience* 18 (9): 547–58. https://doi.org/10.1038/nrn.2017.74.

Elward, Rachael L., Anna M. Dzieciol, and Faraneh Vargha-Khadem. 2019. "Little Evidence for Fast Mapping in Adults with Developmental Amnesia." *Cognitive Neuroscience* 10 (4): 215–17. https://doi.org/10.1080/17588928.2019.1593123.

Elward, Rachael L., and Faraneh Vargha-Khadem. 2018. "Semantic Memory in Developmental Amnesia." *Neuroscience Letters* 680 (July): 23–30. https://doi.org/10.1016/j.neulet.2018.04.040.

Eriksson, Johan, Edward K. Vogel, Anders Lansner, Fredrik Bergström, and Lars Nyberg. 2015. "Neurocognitive Architecture of Working Memory." *Neuron* 88 (1): 33–46. https://doi.org/10.1016/j.neuron.2015.09.020.

Etkin, Amit, and Tor D. Wager. 2007. "Functional Neuroimaging of Anxiety: A Meta-Analysis of Emotional Processing in PTSD, Social Anxiety Disorder, and Specific Phobia." *American Journal of Psychiatry* 164 (10): 1476–88. https://doi.org/10.1176/appi.ajp.2007.07030504.

Evans, Jonathan, and Keith Frankish, eds. 2009. *In Two Minds: Dual Processes and Beyond*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.001.0001.

Fazio, Russell H. 1990. "Multiple Processes by Which Attitudes Guide Behavior: The Mode Model as an Integrative Framework." In *Advances in Experimental Social Psychology*, 23:75–109. Elsevier. https://doi.org/10.1016/S0065-2601(08)60318-4.

———. 2001. "On the Automatic Activation of Associated Evaluations: An Overview." *Cognition & Emotion* 15 (2): 115–41. https://doi.org/10.1080/02699930125908.

Fazio, Russell H., and Michael A. Olson. 2003. "Implicit Measures in Social Cognition Research: Their Meaning and Use." *Annual Review of Psychology* 54 (1): 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225.

Fazio, Russell H., David M. Sanbonmatsu, Martha C. Powell, and Frank R. Kardes. 1986. "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology* 50 (2): 229–38. https://doi.org/10.1037/0022-3514.50.2.229.

Feczko, Eric, Oscar Miranda-Dominguez, Mollie Marr, Alice M. Graham, Joel T. Nigg, and Damien A. Fair. 2019. "The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes." *Trends in Cognitive Sciences* 23 (7): 584–601. https://doi.org/10.1016/j.tics.2019.03.009.

Ferbinteanu, J. 2019. "Memory Systems 2018 – Towards a New Paradigm." *Neurobiology of Learning and Memory* 157 (January): 61–78. https://doi.org/10.1016/j.nlm.2018.11.005.

Ferguson, Melissa J, and John A Bargh. n.d. "Beyond the Attitude Object:," 48.

Ferguson, Melissa J., Thomas C. Mann, Jeremy Cone, and Xi Shen. 2019. "When and How Implicit First Impressions Can Be Updated." *Current Directions in Psychological Science* 28 (4): 331–36. https://doi.org/10.1177/0963721419835206.

Fermin, Alan S. R., Takehiko Yoshida, Junichiro Yoshimoto, Makoto Ito, Saori C. Tanaka, and Kenji Doya. 2016. "Model-Based Action Planning Involves Cortico-Cerebellar and Basal Ganglia Networks." *Scientific Reports* 6 (1): 31378. https://doi.org/10.1038/srep31378.

Fernández, Rodrigo S., Soledad Picco, Fernando Messore, and María E. Pedreira. 2018. "Effects of Threat Conditioning on the Negative Valanced Systems and Cognitive Systems." *Scientific Reports* 8 (1): 11221. https://doi.org/10.1038/s41598-018-29603-3.

Fiedler, Klaus, Matthias Bluemke, and Christian Unkelbach. 2011. "On the Adaptive Flexibility of Evaluative Priming." *Memory & Cognition* 39 (4): 557–72. https://doi.org/10.3758/s13421-010-0056-x.

Floden, Darlene, and Donald T. Stuss. 2006. "Inhibitory Control Is Slowed in Patients with Right Superior Medial Frontal Damage." *Journal of Cognitive Neuroscience* 18 (11): 1843–49. https://doi.org/10.1162/jocn.2006.18.11.1843.

Fodor, J. A. 1974. "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)." *Synthese* 28 (2): 97–115.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass: MIT Press.

Folstein, Jonathan R., and Michael A. Dieciuc. 2019. "The Cognitive Neuroscience of Stable and Flexible Semantic Typicality." *Frontiers in Psychology* 10 (May): 1265. https://doi.org/10.3389/fpsyg.2019.01265.

Forscher, Patrick S., Calvin K. Lai, Jordan R. Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine, and Brian A. Nosek. 2019. "A Meta-Analysis of Procedures to Change Implicit

Measures." *Journal of Personality and Social Psychology* 117 (3): 522–59. https://doi.org/10.1037/pspa0000160.

Frank, Michael J., Jerry W. Rudy, William B. Levy, and Randall C. O'Reilly. 2005. "When Logic Fails: Implicit Transitive Inference in Humans." *Memory & Cognition* 33 (4): 742–50. https://doi.org/10.3758/BF03195340.

Frank, Michael J., Jerry W. Rudy, and Randall C. O'Reilly. 2003. "Transitivity, Flexibility, Conjunctive Representations, and the Hippocampus. II. A Computational Analysis." *Hippocampus* 13 (3): 341–54. https://doi.org/10.1002/hipo.10084.

Freedberg, Michael, Andrew C. Toader, Eric M. Wassermann, and Joel L. Voss. 2020. "Competitive and Cooperative Interactions between Medial Temporal and Striatal Learning Systems." *Neuropsychologia* 136 (January): 107257. https://doi.org/10.1016/j.neuropsychologia.2019.107257.

Friese, Malte, Wilhelm Hofmann, and Manfred Schmitt. 2008. "When and Why Do Implicit Measures Predict Behaviour? Empirical Evidence for the Moderating Role of Opportunity, Motivation, and Process Reliance." *European Review of Social Psychology* 19 (1): 285–338. https://doi.org/10.1080/10463280802556958.

Friese, Malte, Colin Tucker Smith, Thomas Plischke, Matthias Bluemke, and Brian A. Nosek. 2012. "Do Implicit Attitudes Predict Actual Voting Behavior Particularly for Undecided Voters?" Edited by Frank Krueger. *PLoS ONE* 7 (8): e44130. https://doi.org/10.1371/journal.pone.0044130.

Frymer, Paul, and Jacob M. Grumbach. 2021. "Labor Unions and White Racial Politics." *American Journal of Political Science* 65 (1): 225–40. https://doi.org/10.1111/ajps.12537.

Fulcher, Eamon P., and Roger P. Cocks. 1997. "Dissociative Storage Systems in Human Evaluative Conditioning." *Behaviour Research and Therapy* 35 (1): 1–10. https://doi.org/10.1016/S0005-7967(96)00081-2.

Gabrieli, John D.E., Debra A. Fleischman, Margaret M. Keane, Sheryl L. Reminger, and Frank Morrell. 1995. "Double Dissociation Between Memory Systems Underlying Explicit and Implicit Memory in the Human Brain." *Psychological Science* 6 (2): 76–82. https://doi.org/10.1111/j.1467-9280.1995.tb00310.x.

Gao, Andrew F., Julia L. Keith, Fu-qiang Gao, Sandra E. Black, Morris Moscovitch, and R. Shayna Rosenbaum. 2020. "Neuropathology of a Remarkable Case of Memory Impairment Informs Human Memory." *Neuropsychologia* 140 (March): 107342. https://doi.org/10.1016/j.neuropsychologia.2020.107342.

Gardiner, John M., Karen R. Brandt, Alan D. Baddeley, Faraneh Vargha-Khadem, and Mortimer Mishkin. 2008. "Charting the Acquisition of Semantic Knowledge in a Case of Developmental Amnesia." *Neuropsychologia* 46 (11): 2865–68. https://doi.org/10.1016/j.neuropsychologia.2008.05.021.

Garolera, Maite, Richard Coppola, Karen E. Muñoz, Brita Elvevåg, Frederick W. Carver, Daniel R. Weinberger, and Terry E. Goldberg. 2007. "Amygdala Activation in Affective Priming: A Magnetoencephalogram Study." *NeuroReport* 18 (14): 1449–1453. https://doi.org/10.1097/WNR.0b013e3282efa253.

Garolera, Maite, Richard Coppola, Daniel R Weinberger, and Terry E Goldberg. n.d. "Amygdala Activation in A¡ective Priming: A Magnetoencephalogram Study," 5.

Gast, Anne. 2018. "A Declarative Memory Model of Evaluative Conditioning." *Social Psychological Bulletin* 13 (3): e28590. https://doi.org/10.5964/spb.v13i3.28590.

Gast, Anne, and Jan De Houwer. 2013. "The Influence of Extinction and Counterconditioning Instructions on Evaluative Conditioning Effects." *Learning and Motivation* 44 (4): 312–25. https://doi.org/10.1016/j.lmot.2013.03.003.

Gast, Anne, and Klaus Rothermund. 2011. "I like It Because I Said That I like It: Evaluative Conditioning Effects Can Be Based on Stimulus-Response Learning." *Journal of Experimental Psychology: Animal Behavior Processes* 37 (4): 466–76. https://doi.org/10.1037/a0023077.

Gawronski, Bertram. 2019. "Six Lessons for a Cogent Science of Implicit Bias and Its Criticism." *Perspectives on Psychological Science* 14 (4): 574–95. https://doi.org/10.1177/1745691619826015.

Gawronski, Bertram, and Galen Bodenhausen. 2007. "What Do We Know About Implicit Measures and What Do We Have to Learn?" In *Implicit Measures of Attitudes*, edited by Bernd Wittenbrink and Norbert Schwarz. New York: Guilford Press.

Gawronski, Bertram, and Galen V. Bodenhausen. 2005. "Accessibility Effects on Implicit Social Cognition: The Role of Knowledge Activation and Retrieval Experiences." *Journal of Personality and Social Psychology* 89 (5): 672–85. https://doi.org/10.1037/0022-3514.89.5.672.

———. 2006. "Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change." *Psychological Bulletin* 132 (5): 692–731. https://doi.org/10.1037/0033-2909.132.5.692.

———. 2011. "The Associative–Propositional Evaluation Model." In *Advances in Experimental Social Psychology*, 44:59–127. Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00002-0.

———. 2018. "Evaluative Conditioning From the Perspective of the Associative-Propositional Evaluation Model." *Social Psychological Bulletin* 13 (3): e28024. https://doi.org/10.5964/spb.v13i3.28024.

Gawronski, Bertram, Skylar M Brannon, and Galen V Bodenhausen. n.d. "THE ASSOCIATIVE-PROPOSITIONAL DUALITY IN THE REPRESENTATION, FORMATION, AND EXPRESSION OF ATTITUDES," 9.

Gawronski, Bertram, and Jan De Houwer. 2013. "Implicit Measures in Social and Personality Psychology." In *Handbook of Research Methods in Social and Personality Psychology*, edited by Harry T. Reis and Charles M. Judd, 2nd ed., 283–310. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.016.

Gawronski, Bertram, Roland Deutsch, Sawsan Mbirkou, Beate Seibt, and Fritz Strack. 2008. "When 'Just Say No' Is Not Enough: Affirmation versus Negation Training and the Reduction of Automatic Stereotype Activation." *Journal of Experimental Social Psychology* 44 (2): 370–77. https://doi.org/10.1016/j.jesp.2006.12.004.

Gawronski, Bertram, Anne Gast, and Jan De Houwer. 2015. "Is Evaluative Conditioning Really Resistant to Extinction? Evidence for Changes in Evaluative Judgements without Changes in Evaluative Representations." *Cognition and Emotion* 29 (5): 816–30. https://doi.org/10.1080/02699931.2014.947919.

Gawronski, Bertram, Daniel Geschke, and Rainer Banse. 2003. "Implicit Bias in Impression Formation: Associations Influence the Construal of Individuating Information." *European Journal of Social Psychology* 33 (5): 573–89. https://doi.org/10.1002/ejsp.166.

Gawronski, Bertram, and Adam Hahn. n.d. "Procedures, Use, and Interpretation," 27.

Gawronski, Bertram, and Derek G. V. Mitchell. 2014. "Simultaneous Conditioning of Valence and Arousal." *Cognition and Emotion* 28 (4): 577–95. https://doi.org/10.1080/02699931.2013.843506.

Gawronski, Bertram, Mike Morrison, Curtis E Phills, and Silvia Galdi. n.d. “Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis.” *Personality and Social Psychology Bulletin*, 14.

Gawronski, Bertram, and Eva Walther. 2012. “What Do Memory Data Tell Us about the Role of Contingency Awareness in Evaluative Conditioning?” *Journal of Experimental Social Psychology* 48 (3): 617–23. https://doi.org/10.1016/j.jesp.2012.01.002.

Gawronski, Bertram, Eva Walther, and Hartmut Blank. 2005. “Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information.” *Journal of Experimental Social Psychology* 41 (6): 618–26. https://doi.org/10.1016/j.jesp.2004.10.005.

Gazzaniga, Michael S., ed. 2009. *The Cognitive Neurosciences*. 4th ed. Cambridge, Mass: MIT Press.

Gemar, Michael C., Zindel V. Segal, Sandra Sagrati, and Sidney J. Kennedy. 2001. “Mood-Induced Changes on the Implicit Association Test in Recovered Depressed Patients.” *Journal of Abnormal Psychology* 110 (2): 282–89. https://doi.org/10.1037/0021-843X.110.2.282.

Gendler, Tamar Szabó. 2008. “Alief in Action (and Reaction).” *Mind & Language* 23 (5): 552–85. https://doi.org/10.1111/j.1468-0017.2008.00352.x.

Germar, Markus, and Andreas Mojzisch. 2019. “Learning of Social Norms Can Lead to a Persistent Perceptual Bias: A Diffusion Model Approach.” *Journal of Experimental Social Psychology* 84 (September): 103801. https://doi.org/10.1016/j.jesp.2019.03.012.

Ghosh, Vanessa E., and Asaf Gilboa. 2014. “What Is a Memory Schema? A Historical Perspective on Current Neuroscience Literature.” *Neuropsychologia* 53 (January): 104–14. https://doi.org/10.1016/j.neuropsychologia.2013.11.010.

Ghuman, Avniel Singh, and Alex Martin. 2019. “Dynamic Neural Representations: An Inferential Challenge for FMRI.” *Trends in Cognitive Sciences* 23 (7): 534–36. https://doi.org/10.1016/j.tics.2019.04.004.

Gilbert, Sam J., Jillian K. Swencionis, and David M. Amodio. 2012. “Evaluative vs. Trait Representation in Intergroup Social Judgments: Distinct Roles of Anterior Temporal Lobe and Prefrontal Cortex.” *Neuropsychologia* 50 (14): 3600–3611. https://doi.org/10.1016/j.neuropsychologia.2012.09.002.

Gilboa, Asaf, and Hannah Marlatte. 2017. “Neurobiology of Schemas and Schema-Mediated Memory.” *Trends in Cognitive Sciences* 21 (8): 618–31. https://doi.org/10.1016/j.tics.2017.04.013.

Gilboa, Asaf, and Morris Moscovitch. 2017. “Ventromedial Prefrontal Cortex Generates Pre-Stimulus Theta Coherence Desynchronization: A Schema Instantiation Hypothesis.” *Cortex* 87 (February): 16–30. https://doi.org/10.1016/j.cortex.2016.10.008.

Gladwell, Malcolm, Richard Nisbett, and Lee Ross. n.d. “The Person and the Situation,” 691.

Gladwin, Thomas E., Martin Möbius, and Eni S. Becker. 2019. “Predictive Attentional Bias Modification Induces Stimulus-Evoked Attentional Bias for Threat.” *Europe’s Journal of Psychology* 15 (3): 479–90. https://doi.org/10.5964/ejop.v15i3.1633.

Glaser, Tina, Marcella L. Woud, Michael Labib Iskander, Vera Schmalenstroth, and Thuy My Vo. 2018. “Positive, Negative, or All Relative? Evaluative Conditioning of Ambivalence.” *Acta Psychologica* 185 (April): 155–65. https://doi.org/10.1016/j.actpsy.2018.02.006.

Glautier, Steven, Jan De Houwer, and Edward Redhead. n.d. “Evaluative Conditioning and Affective Priming: Awareness,” 54.

Goel, Vinod. 2007. “Anatomy of Deductive Reasoning.” *Trends in Cognitive Sciences* 11 (10): 435–441.

Goel, Vinod, Milan Makale, and Jordan Grafman. 2004. "The Hippocampal System Mediates Logical Reasoning about Familiar Spatial Environments." *Journal of Cognitive Neuroscience* 16 (4): 654–664.

Goldfield, Michael. 1993. "Race and the CIO: The Possibilities for Racial Egalitarianism During the 1930s and 1940s." *International Labor and Working-Class History* 44: 1–32. https://doi.org/10.1017/S0147547900012187.

Goodman, Jarid, Christa McIntyre, and Mark G. Packard. 2017. "Amygdala and Emotional Modulation of Multiple Memory Systems." In *The Amygdala - Where Emotions Shape Perception, Learning and Memories*, edited by Barbara Ferry. InTech. https://doi.org/10.5772/intechopen.69109.

Gopnik, Alison, and Laura Schulz. 2007. *Causal Learning*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176803.001.0001.

Gordon Hayman, C. A., Carol A. Macdonald, and Endel Tulving. 1993. "The Role of Repetition and Associative Interference in New Semantic Learning in Amnesia: A Case Experiment." *Journal of Cognitive Neuroscience* 5 (4): 375–89. https://doi.org/10.1162/jocn.1993.5.4.375.

Gore, Felicity, Edmund C. Schwartz, Baylor C. Brangers, Stanley Aladi, Joseph M. Stujenske, Ekaterina Likhtik, Marco J. Russo, Joshua A. Gordon, C. Daniel Salzman, and Richard Axel. 2015. "Neural Representations of Unconditioned Stimuli in Basolateral Amygdala Mediate Innate and Learned Responses." *Cell* 162 (1): 134–45. https://doi.org/10.1016/j.cell.2015.06.027.

Greenberg, Daniel L., and Mieke Verfaellie. 2010. "Interdependence of Episodic and Semantic Memory: Evidence from Neuropsychology." *Journal of the International Neuropsychological Society* 16 (5): 748–53. https://doi.org/10.1017/S1355617710000676.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review* 102 (1): 4–27. https://doi.org/10.1037/0033-295X.102.1.4.

Greenwald, Anthony G, and Shelly D Farnham. n.d. "Using the Implicit Association Test to Measure Self-Esteem and Self-Concept," 17.

Gregg, Aiden P., Beate Seibt, and Mahzarin R. Banaji. 2006. "Easier Done than Undone: Asymmetry in the Malleability of Implicit Preferences." *Journal of Personality and Social Psychology* 90 (1): 1–20. https://doi.org/10.1037/0022-3514.90.1.1.

Greifeneder, Rainer, Herbert Bless, Klaus Fiedler, and Herbert Bless. 2018. *Social Cognition: How Individuals Construct Social Reality*. Second Edition. New York: Routledge.

Griffiths, Paul E. 2004. "Emotions as Natural and Normative Kinds." *Philosophy of Science* 71 (5): 901–11. https://doi.org/10.1086/425944.

Gupta, Rashmi, Young-Jin Hur, and Nilli Lavie. 2016. "Distracted by Pleasure: Effects of Positive versus Negative Valence on Emotional Capture under Load." *Emotion* 16 (3): 328–37. https://doi.org/10.1037/emo0000112.

Hackel, Leor M., Jeffrey J. Berg, Björn R. Lindström, and David M. Amodio. 2019a. "Model-Based and Model-Free Social Cognition: Investigating the Role of Habit in Social Attitude Formation and Choice." *Frontiers in Psychology* 10 (November): 2592. https://doi.org/10.3389/fpsyg.2019.02592.

Hackel, Leor M, Jeffrey Jordan Berg, Björn Lindström, and David Amodio. 2019b. "Model-Based and Model-Free Social Cognition." Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/ue6j2.

Hackel, Leor M, Bradley B Doll, and David M Amodio. 2015. "Instrumental Learning of Traits versus Rewards: Dissociable Neural Correlates and Effects on Choice." *Nature Neuroscience* 18 (9): 1233–35. https://doi.org/10.1038/nn.4080.

Haidt, Jonathan, and Craig Joseph. n.d. "19 The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules," 31.

Hall, John F. 1984. "Backward Conditioning in Pavlovian Type Studies" 19 (4): 6.

Hall, N. 2007. "Structural Equations and Causation." *Philosophical Studies* 132 (1): 109–36. https://doi.org/10.1007/s11098-006-9057-9.

Hamann, Stephan, and Hui Mao. 2002. "Positive and Negative Emotional Verbal Stimuli Elicit Activity in the Left Amygdala:" *Neuroreport* 13 (1): 15–19. https://doi.org/10.1097/00001756-200201210-00008.

Hammack, Phillip L. 2017. *Social Psychology and Social Justice: Critical Principles and Perspectives for the Twenty-First Century*. Edited by Phillip L. Hammack. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199938735.013.1.

Hampshire, Adam, Richard E. Daws, Ines Das Neves, Eyal Soreq, Stefano Sandrone, and Ines R. Violante. 2019. "Probing Cortical and Sub-Cortical Contributions to Instruction-Based Learning: Regional Specialisation and Global Network Dynamics." *NeuroImage* 192 (May): 88–100. https://doi.org/10.1016/j.neuroimage.2019.03.002.

Han, H. Anna, Sandor Czellar, Michael A. Olson, and Russell H. Fazio. 2010. "Malleability of Attitudes or Malleability of the IAT?" *Journal of Experimental Social Psychology* 46 (2): 286–98. https://doi.org/10.1016/j.jesp.2009.11.011.

Hannula, Deborah E., and Anthony J. Greene. 2012. "The Hippocampus Reevaluated in Unconscious Learning and Memory: At a Tipping Point?" *Frontiers in Human Neuroscience* 6. https://doi.org/10.3389/fnhum.2012.00080.

Hannula, Deborah E., Jennifer D. Ryan, Daniel Tranel, and Neal J. Cohen. 2007. "Rapid Onset Relational Memory Effects Are Evident in Eye Movement Behavior, but Not in Hippocampal Amnesia." *Journal of Cognitive Neuroscience* 19 (10): 1690–1705. https://doi.org/10.1162/jocn.2007.19.10.1690.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science, New Series* 162 (3859): 1243–48.

Harenski, Carla L., and Stephan Hamann. 2006. "Neural Correlates of Regulating Negative Emotions Related to Moral Violations." *NeuroImage* 30 (1): 313–24. https://doi.org/10.1016/j.neuroimage.2005.09.034.

Hariri, Ahmad R., Alessandro Tessitore, Venkata S. Mattay, Francesco Fera, and Daniel R. Weinberger. 2002. "The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes." *NeuroImage* 17 (1): 317–23. https://doi.org/10.1006/nimg.2002.1179.

Harman, Gilbert. 2000. *Explaining Value*. Oxford University Press. https://doi.org/10.1093/0198238045.001.0001.

Hartley, Tom, and Neil Burgess. 2005. "Complementary Memory Systems: Competition, Cooperation and Compensation." *Trends in Neurosciences* 28 (4): 169–70. https://doi.org/10.1016/j.tins.2005.02.004.

Haslanger, Sally Anne. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.

Hausman, Alan, Howard Kahane, and Paul Tidman. 2010. *Logic and Philosophy: A Modern Introduction*. 11th ed. Australia ; Boston, MA: Thomson Wadsworth/Cengage learning.

Hausman, Daniel Murray. 2005. "Causal Relata: Tokens, Types, or Variables?" *Erkenntnis* 63 (1): 33–54. https://doi.org/10.1007/s10670-005-0562-6.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2–3): 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henson, Richard N., Doris Eckstein, Florian Waszak, Christian Frings, and Aidan J. Horner. 2014. "Stimulus–Response Bindings in Priming." *Trends in Cognitive Sciences* 18 (7): 376–84. https://doi.org/10.1016/j.tics.2014.03.004.

Henwood, Benjamin F., Taylor Harris, Darlene Woo, Hailey Winetrobe, Harmony Rhoades, and Suzanne L. Wenzel. 2018. "Availability of Comprehensive Services in Permanent Supportive Housing in Los Angeles." *Health & Social Care in the Community* 26 (2): 207–13. https://doi.org/10.1111/hsc.12510.

Hermans, Dirk, Debora Vansteenwegen, Geert Crombez, Frank Baeyens, and Paul Eelen. 2002. "Expectancy-Learning and Evaluative Learning in Human Classical Conditioning: Affective Priming as an Indirect and Unobtrusive Measure of Conditioned Stimulus Valence." *Behaviour Research and Therapy* 40 (3): 217–34. https://doi.org/10.1016/S0005-7967(01)00006-7.

Hermans, Erno J., Jonathan W. Kanen, Arielle Tambini, Guillén Fernández, Lila Davachi, and Elizabeth A. Phelps. 2017. "Persistence of Amygdala–Hippocampal Connectivity and Multi-Voxel Correlation Structures During Awake Rest After Fear Learning Predicts Long-Term Expression of Fear." *Cerebral Cortex* 27 (5): 3028–41. https://doi.org/10.1093/cercor/bhw145.

Herring, David R., Katherine R. White, Linsa N. Jabeen, Michelle Hinojos, Gabriela Terrazas, Stephanie M. Reyes, Jennifer H. Taylor, and Stephen L. Crites. 2013. "On the Automatic Activation of Attitudes: A Quarter Century of Evaluative Priming Research." *Psychological Bulletin* 139 (5): 1062–89. https://doi.org/10.1037/a0031309.

Heycke, Tobias, and Bertram Gawronski. 2020. "Co-Occurrence and Relational Information in Evaluative Learning: A Multinomial Modeling Approach." *Journal of Experimental Psychology: General* 149 (1): 104–24. https://doi.org/10.1037/xge0000620.

Heyman, Tom, Keith A. Hutchison, and Gert Storms. 2016. "Is Semantic Priming (Ir)Rational? Insights from the Speeded Word Fragment Completion Task." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42 (10): 1657–63. https://doi.org/10.1037/xlm0000260.

Hilverman, Caitlin, Sharice A. Clough, Melissa C. Duff, and Susan Wagner Cook. 2018. "Patients with Hippocampal Amnesia Successfully Integrate Gesture and Speech." *Neuropsychologia* 117 (August): 332–38. https://doi.org/10.1016/j.neuropsychologia.2018.06.012.

Hinman, Lawrence M. 2013. *Contemporary Moral Issues: Diversity and Consensus*. 4th ed. Boston: Pearson.

Hodges, John R., and Kim S. Graham. 2001. "Episodic Memory: Insights from Semantic Dementia." Edited by A. Baddeley, M. Conway, and J. Aggleton. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356 (1413): 1423–34. https://doi.org/10.1098/rstb.2001.0943.

Hofmann, Wilhelm, Jan De Houwer, Marco Perugini, Frank Baeyens, and Geert Crombez. 2010. "Evaluative Conditioning in Humans: A Meta-Analysis." *Psychological Bulletin* 136 (3): 390–421. https://doi.org/10.1037/a0018916.

"Holroyd and Sweetman 2016 - Google Search." n.d. Accessed September 30, 2019. https://www.google.com/search?q=Holroyd+and+sweetman+2016&rlz=1C5CHFA_enUS775US775&oq=Holroyd+and+sweetman+2016&aqs=chrome..69i57.6265j1j7&sourceid=chrome&ie=UTF-8.

Holroyd, Jules. 2015. "Implicit Bias, Awareness and Imperfect Cognitions." *Consciousness and Cognition* 33 (May): 511–23. https://doi.org/10.1016/j.concog.2014.08.024.

Holroyd, Jules, Robin Scaife, and Tom Stafford. 2017. "What Is Implicit Bias?" *Philosophy Compass* 12 (10): e12437. https://doi.org/10.1111/phc3.12437.

Houwer, Jan De. 2014. "A Propositional Model of Implicit Evaluation: Implicit Evaluation." *Social and Personality Psychology Compass* 8 (7): 342–53. https://doi.org/10.1111/spc3.12111.

Hoyle, Rick. 2011. *Structural Equation Modeling for Social and Personality Psychology*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781446287965.

Hu, Xiaoqing, Bertram Gawronski, and Robert Balas. 2017. "Propositional Versus Dual-Process Accounts of Evaluative Conditioning: II. The Effectiveness of Counter-Conditioning and Counter-Instructions in Changing Implicit and Explicit Evaluations." *Social Psychological and Personality Science* 8 (8): 858–66. https://doi.org/10.1177/1948550617691094.

———. n.d. "Propositional Versus Dual-Process Accounts of Evaluative Conditioning." *Personality and Social Psychology Bulletin*, 16.

Huang, Julie Y., and John A. Bargh. 2014. "The Selfish Goal: Autonomously Operating Motivational Structures as the Proximate Cause of Human Judgment and Behavior." *Behavioral and Brain Sciences* 37 (2): 121–35. https://doi.org/10.1017/S0140525X13000290.

Huebner, Bryce. 2016. "Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition." In *Implicit Bias and Philosophy, Volume 1*, edited by Michael Brownstein and Jennifer Saul, 47–79. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198713241.003.0003.

Hughes, Sean, Yang Ye, Pieter Van Dessel, and Jan De Houwer. 2019. "When People Co-Occur With Good or Bad Events: Graded Effects of Relational Qualifiers on Evaluative Conditioning." *Personality and Social Psychology Bulletin* 45 (2): 196–208. https://doi.org/10.1177/0146167218781340.

Hurley, Niamh C., Eleanor A. Maguire, and Faraneh Vargha-Khadem. 2011. "Patient HC with Developmental Amnesia Can Construct Future Scenarios." *Neuropsychologia* 49 (13): 3620–28. https://doi.org/10.1016/j.neuropsychologia.2011.09.015.

Huster, René J., Stefanie Enriquez-Geppert, Christina F. Lavallee, Michael Falkenstein, and Christoph S. Herrmann. 2013. "Electroencephalography of Response Inhibition Tasks: Functional Networks and Cognitive Contributions." *International Journal of Psychophysiology* 87 (3): 217–33. https://doi.org/10.1016/j.ijpsycho.2012.08.001.

Hutchinson, J. Benjamin, and Nicholas B. Turk-Browne. 2012. "Memory-Guided Attention: Control from Multiple Memory Systems." *Trends in Cognitive Sciences* 16 (12): 576–79. https://doi.org/10.1016/j.tics.2012.10.003.

Hütter, Mandy, and Jan De Houwer. 2017. "Examining the Contributions of Memory-Dependent and Memory-Independent Components to Evaluative Conditioning via Instructions." *Journal of Experimental Social Psychology* 71 (July): 49–58. https://doi.org/10.1016/j.jesp.2017.02.007.

Hütter, Mandy, and Steven Sweldens. 2013. "Implicit Misattribution of Evaluative Responses: Contingency-Unaware Evaluative Conditioning Requires Simultaneous Stimulus Presentations." *Journal of Experimental Psychology. General* 142 (3): 638–43. https://doi.org/10.1037/a0029989.

Iacoboni, Marco, Matthew D Lieberman, Barbara J Knowlton, Istvan Molnar-Szakacs, Mark Moritz, C.Jason Throop, and Alan Page Fiske. 2004. "Watching Social Interactions Produces Dorsomedial Prefrontal and Medial Parietal BOLD FMRI Signal Increases Compared to a Resting Baseline." *NeuroImage* 21 (3): 1167–73. https://doi.org/10.1016/j.neuroimage.2003.11.013.

Ito, Tiffany A., Naomi P. Friedman, Bruce D. Bartholow, Joshua Correll, Chris Loersch, Lee J. Altamirano, and Akira Miyake. 2015. "Toward a Comprehensive Understanding of Executive Cognitive Function in Implicit Racial Bias." *Journal of Personality and Social Psychology* 108 (2): 187–218. https://doi.org/10.1037/a0038557.

Ito, Tiffany A., and Silvia Tomelleri. 2017. “Seeing Is Not Stereotyping: The Functional Independence of Categorization and Stereotype Activation.” *Social Cognitive and Affective Neuroscience* 12 (5): 758–64. https://doi.org/10.1093/scan/nsx009.

Izuma, Keise, and Ralph Adolphs. 2013. “Social Manipulation of Preference in the Human Brain.” *Neuron* 78 (3): 563–73. https://doi.org/10.1016/j.neuron.2013.03.023.

Jacquette, Dale, ed. 2018. *The Bloomsbury Companion to the Philosophy of Consciousness*. Bloomsbury Academic. https://doi.org/10.5040/9781474229043.

Jarvers, Christian, Tobias Brosch, André Brechmann, Marie L. Woldeit, Andreas L. Schulz, Frank W. Ohl, Marcel Lommerzheim, and Heiko Neumann. 2016. “Reversal Learning in Humans and Gerbils: Dynamic Control Network Facilitates Learning.” *Frontiers in Neuroscience* 10 (November). https://doi.org/10.3389/fnins.2016.00535.

Jenkin, Zoe, and Susanna Siegel. 2015. “Cognitive Penetrability: Modularity, Epistemology, and Ethics.” *Review of Philosophy and Psychology* 6 (4): 531–45. https://doi.org/10.1007/s13164-015-0252-5.

Jenkins, Adrianna C. 2019. “Rethinking Cognitive Load: A Default-Mode Network Perspective.” *Trends in Cognitive Sciences* 23 (7): 531–33. https://doi.org/10.1016/j.tics.2019.04.008.

Jin, Jingji, and Stephen Maren. 2015. “Prefrontal-Hippocampal Interactions in Memory and Emotion.” *Frontiers in Systems Neuroscience* 9 (December). https://doi.org/10.3389/fnsys.2015.00170.

Johnen, Ann-Kathrin, and Neil R. Harrison. 2019. “The Effects of Valid and Invalid Expectations about Stimulus Valence on Behavioural and Electrophysiological Responses to Emotional Pictures.” *International Journal of Psychophysiology* 144 (October): 47–55. https://doi.org/10.1016/j.ijpsycho.2019.08.002.

Johnson-Laird, P.N., Sangeet S. Khemlani, and Geoffrey P. Goodwin. 2015. "Logic, Probability, and Human Reasoning." *Trends in Cognitive Sciences* 19 (4): 201–14. https://doi.org/10.1016/j.tics.2015.02.006.

Joiner, Jessica, Matthew Piva, Courtney Turrin, and Steve W. C. Chang. 2017. "Social Learning through Prediction Error in the Brain." *Npj Science of Learning* 2 (1): 8. https://doi.org/10.1038/s41539-017-0009-2.

Jones, Christopher R., Michael A. Olson, and Russell H. Fazio. 2010. "Evaluative Conditioning." In *Advances in Experimental Social Psychology*, 43:205–55. Elsevier. https://doi.org/10.1016/S0065-2601(10)43005-1.

Jones, Michael N., Jon Willits, and Simon Dennis. 2015. *Models of Semantic Memory*. Edited by Jerome R. Busemeyer, Zheng Wang, James T. Townsend, and Ami Eidels. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.11.

Jonin, Pierre-Yves, Gabriel Besson, Renaud La Joie, Jérémie Pariente, Serge Belliard, Christian Barillot, and Emmanuel J. Barbeau. 2018. "Superior Explicit Memory despite Severe Developmental Amnesia: In-Depth Case Study and Neural Correlates." *Hippocampus* 28 (12): 867–85. https://doi.org/10.1002/hipo.23010.

Josselyn, Sheena A., and Paul W. Frankland. 2018. "Memory Allocation: Mechanisms and Function." *Annual Review of Neuroscience* 41 (1): 389–413. https://doi.org/10.1146/annurev-neuro-080317-061956.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin. 2009. "The Existence of Implicit Bias Is beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore." *Research in Organizational Behavior* 29: 39–69. https://doi.org/10.1016/j.riob.2009.10.001.

Judd, Charles M., Irene V. Blair, and Kristine M. Chapleau. 2004. "Automatic Stereotypes vs. Automatic Prejudice: Sorting out the Possibilities in the Weapon Paradigm." *Journal of Experimental Social Psychology* 40 (1): 75–81. https://doi.org/10.1016/S0022-1031(03)00063-5.

Karpinski, Andrew, Jessie C. Briggs, and Miriam Yale. 2019. "A Direct Replication: Unconscious Arithmetic Processing." *European Journal of Social Psychology* 49 (3): 637–44. https://doi.org/10.1002/ejsp.2390.

Kashima, Yoshihisa, Simon M. Laham, Jennifer Dix, Bianca Levis, Darlene Wong, and Melissa Wheeler. 2015. "Social Transmission of Cultural Practices and Implicit Attitudes." *Organizational Behavior and Human Decision Processes* 129 (July): 113–25. https://doi.org/10.1016/j.obhdp.2014.05.005.

Kawakami, K., D.M. Amodio, and K. Hugenberg. 2017. "Intergroup Perception and Cognition." In *Advances in Experimental Social Psychology*, 55:1–80. Elsevier. https://doi.org/10.1016/bs.aesp.2016.10.001.

Kawakami, Kerry, Jasper Moll, Sander Hermsen, John F Dovidio, and Abby Russin. n.d. "Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation," 18.

Kay, Kenneth, and Loren M. Frank. 2019. "Three Brain States in the Hippocampus and Cortex." *Hippocampus* 29 (3): 184–238. https://doi.org/10.1002/hipo.22956.

Kelly, Daniel, and Taylor Davis. 2018. "SOCIAL NORMS AND HUMAN NORMATIVE PSYCHOLOGY." *Social Philosophy and Policy* 35 (1): 54–76. https://doi.org/10.1017/S0265052518000122.

Kelly, Daniel, and Erica Roedder. 2008. "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3 (3): 522–40. https://doi.org/10.1111/j.1747-9991.2008.00138.x.

Kensinger, Elizabeth A, and Kelly S Giovanello. n.d. "The Status of Semantic and Episodic Memory in Amnesia," 15.

Kenward, Ben, Markus Karlsson, and Joanna Persson. 2011. "Over-Imitation Is Better Explained by Norm Learning than by Distorted Causal Learning." *Proceedings of the Royal Society B: Biological Sciences* 278 (1709): 1239–46. https://doi.org/10.1098/rspb.2010.1399.

Kerman, Nick, John Sylvestre, Tim Aubry, and Jino Distasio. 2018. "The Effects of Housing Stability on Service Use among Homeless Adults with Mental Illness in a Randomized Controlled Trial of Housing First." *BMC Health Services Research* 18 (1): 190. https://doi.org/10.1186/s12913-018-3028-7.

Kesteren, Marlieke T.R. van, Dirk J. Ruiter, Guillén Fernández, and Richard N. Henson. 2012. "How Schema and Novelty Augment Memory Formation." *Trends in Neurosciences* 35 (4): 211–19. https://doi.org/10.1016/j.tins.2012.02.001.

Khalidi, Muhammad Ali. n.d. "Natural Categories and Human Kinds," 268.

Khemlani, Sangeet S., Anthony M. Harrison, and J. Gregory Trafton. 2015. "Episodes, Events, and Models." *Frontiers in Human Neuroscience* 9 (October). https://doi.org/10.3389/fnhum.2015.00590.

Kidder, Ciara K., Katherine R. White, Michelle R. Hinojos, Mayra Sandoval, and Stephen L. Crites. 2018. "Sequential Stereotype Priming: A Meta-Analysis." *Personality and Social Psychology Review* 22 (3): 199–227. https://doi.org/10.1177/1088868317723532.

Kiefer, Markus, Nathalie Liegel, Monika Zovko, and Dirk Wentura. 2017. "Mechanisms of Masked Evaluative Priming: Task Sets Modulate Behavioral and Electrophysiological Priming for Picture and Words Differentially." *Social Cognitive and Affective Neuroscience* 12 (4): 596–608. https://doi.org/10.1093/scan/nsw167.

Kim, Hongkeun. 2019. "Neural Correlates of Explicit and Implicit Memory at Encoding and Retrieval: A Unified Framework and Meta-Analysis of Functional Neuroimaging Studies." *Biological Psychology* 145 (July): 96–111. https://doi.org/10.1016/j.biopsycho.2019.04.006.

Kim, Jeehye Christine, Steven Sweldens, and Mandy Hütter. 2016. "The Symmetric Nature of Evaluative Memory Associations: Equal Effectiveness of Forward Versus Backward Evaluative Conditioning." *Social Psychological and Personality Science* 7 (1): 61–68. https://doi.org/10.1177/1948550615599237.

Kim, John, Chris T. Allen, and Frank R. Kardes. 1996. "An Investigation of the Mediational Mechanisms Underlying Attitudinal Conditioning." *Journal of Marketing Research* 33 (3): 318. https://doi.org/10.2307/3152128.

Kindt, Merel, and Bert Hoekzema. n.d. "Cognitive Bias for Pictorial and Linguistic Threat Cues in Children," 19.

Kitayama, Shinobu, Anthony King, Ming Hsu, Israel Liberzon, and Carolyn Yoon. 2016. "Dopamine-System Genes and Cultural Acquisition: The Norm Sensitivity Hypothesis." *Current Opinion in Psychology* 8 (April): 167–74. https://doi.org/10.1016/j.copsyc.2015.11.006.

Klein, Stanley B. 2016. "Autonoetic Consciousness: Reconsidering the Role of Episodic Memory in Future-Oriented Self-Projection." *Quarterly Journal of Experimental Psychology* 69 (2): 381–401. https://doi.org/10.1080/17470218.2015.1007150.

Klinger, Mark R., Philip C. Burton, and G. Shane Pitts. 2000. "Mechanisms of Unconscious Priming: I. Response Competition, Not Spreading Activation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26 (2): 441–55. https://doi.org/10.1037/0278-7393.26.2.441.

Klingler, J., and P. Gloor. 1960. "The Connections of the Amygdala and of the Anterior Temporal Cortex in the Human Brain: TEMPORAL LOBE CONNECTIONS IN MAN." *Journal of Comparative Neurology* 115 (3): 333–69. https://doi.org/10.1002/cne.901150305.

Klucken, Tim, Katharina Tabbert, Jan Schweckendiek, Christian Josef Merz, Sabine Kagerer, Dieter Vaitl, and Rudolf Stark. 2009. "Contingency Learning in Human Fear Conditioning Involves the Ventral Striatum." *Human Brain Mapping* 30 (11): 3636–44. https://doi.org/10.1002/hbm.20791.

Knowlton, Barbara J., Robert G. Morrison, John E. Hummel, and Keith J. Holyoak. 2012. "A Neurocomputational System for Relational Reasoning." *Trends in Cognitive Sciences* 16 (7): 373–81. https://doi.org/10.1016/j.tics.2012.06.002.

Koban, Leonie, Marieke Jepma, Stephan Geuter, and Tor D. Wager. 2017. "What's in a Word? How Instructions, Suggestions, and Social Information Change Pain and Emotion." *Neuroscience & Biobehavioral Reviews* 81 (October): 29–42. https://doi.org/10.1016/j.neubiorev.2017.02.014.

Koban, Leonie, Marieke Jepma, Marina López-Solà, and Tor D. Wager. 2019. "Different Brain Networks Mediate the Effects of Social and Conditioned Expectations on Pain." *Nature Communications* 10 (1): 1–13. https://doi.org/10.1038/s41467-019-11934-y.

Koban, Leonie, Daniel Kusko, and Tor D. Wager. 2018. "Generalization of Learned Pain Modulation Depends on Explicit Learning." *Acta Psychologica* 184 (March): 75–84. https://doi.org/10.1016/j.actpsy.2017.09.009.

Kok, Peter, Lindsay I. Rait, and Nicholas B. Turk-Browne. 2020. "Content-Based Dissociation of Hippocampal Involvement in Prediction." *Journal of Cognitive Neuroscience* 32 (3): 527–45. https://doi.org/10.1162/jocn_a_01509.

Koppehele-Gossel, Judith, Lisa Hoffmann, Rainer Banse, and Bertram Gawronski. 2020. "Evaluative Priming as an Implicit Measure of Evaluation: An Examination of Outlier-

Treatments for Evaluative Priming Scores." *Journal of Experimental Social Psychology* 87 (March): 103905. https://doi.org/10.1016/j.jesp.2019.103905.

Krawczyk, Daniel C. 2012. "The Cognition and Neuroscience of Relational Reasoning." *Brain Research* 1428 (January): 13–23. https://doi.org/10.1016/j.brainres.2010.11.080.

Kubota, Jennifer T, Mahzarin R Banaji, and Elizabeth A Phelps. 2012. "The Neuroscience of Race." *Nature Neuroscience* 15 (7): 940–48. https://doi.org/10.1038/nn.3136.

Kuperman, Victor, Zachary Estes, Marc Brysbaert, and Amy Beth Warriner. 2014. "Emotion and Language: Valence and Arousal Affect Word Recognition." *Journal of Experimental Psychology: General* 143 (3): 1065–81. https://doi.org/10.1037/a0035669.

Kurdi, Benedek. n.d. "Explicit and Implicit Attitudes towards Norm-Conforming and Norm-Breaking Gay Men in the United States," 94.

Kurdi, Benedek, and Mahzarin R. Banaji. 2017. "Repeated Evaluative Pairings and Evaluative Statements: How Effectively Do They Shift Implicit Attitudes?" *Journal of Experimental Psychology: General* 146 (2): 194–213. https://doi.org/10.1037/xge0000239.

Kurdi, Benedek, Samuel J. Gershman, and Mahzarin R. Banaji. 2019. "Model-Free and Model-Based Learning Processes in the Updating of Explicit and Implicit Evaluations." *Proceedings of the National Academy of Sciences* 116 (13): 6035–44. https://doi.org/10.1073/pnas.1820238116.

Kurdi, Benedek, Thomas C. Mann, Tessa E. S. Charlesworth, and Mahzarin R. Banaji. 2019. "The Relationship between Implicit Intergroup Attitudes and Beliefs." *Proceedings of the National Academy of Sciences* 116 (13): 5862–71. https://doi.org/10.1073/pnas.1820240116.

Kurdi, Benedek, Adam Morris, and Fiery Andrews Cushman. 2020. "The Role of Causal Structure in Implicit Cognition." Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/r7cfa.

Kurdi, Benedek, Allison E. Seitchik, Jordan R. Axt, Timothy J. Carroll, Arpi Karapetyan, Neela Kaushik, Diana Tomezsko, Anthony G. Greenwald, and Mahzarin R. Banaji. 2019.

"Relationship between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis." *American Psychologist* 74 (5): 569–86. https://doi.org/10.1037/amp0000364.

Kwan, D., N. Carson, D.R. Addis, and R.S. Rosenbaum. 2010. "Deficits in Past Remembering Extend to Future Imagining in a Case of Developmental Amnesia." *Neuropsychologia* 48 (11): 3179–86. https://doi.org/10.1016/j.neuropsychologia.2010.06.011.

LaBar, K. S., J. E. LeDoux, D. D. Spencer, and E. A. Phelps. 1995. "Impaired Fear Conditioning Following Unilateral Temporal Lobectomy in Humans." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 15 (10): 6846–55.

Laeger, Inga, Christian Dobel, Udo Dannlowski, Harald Kugel, Dominik Grotegerd, Johanna Kissler, Katharina Keuper, Annuschka Eden, Pienie Zwitserlood, and Peter Zwanzger. 2012. "Amygdala Responsiveness to Emotional Words Is Modulated by Subclinical Anxiety and Depression." *Behavioural Brain Research* 233 (2): 508–16. https://doi.org/10.1016/j.bbr.2012.05.036.

Laisney, Mickaël, Bénédicte Giffard, Serge Belliard, Vincent de la Sayette, Béatrice Desgranges, and Francis Eustache. 2011. "When the Zebra Loses Its Stripes: Semantic Priming in Early Alzheimer's Disease and Semantic Dementia." *Cortex* 47 (1): 35–46. https://doi.org/10.1016/j.cortex.2009.11.001.

Landis, Theodor. 2006. "Emotional Words: What's so Different from Just Words?" *Cortex* 42 (6): 823–30. https://doi.org/10.1016/S0010-9452(08)70424-6.

Lane, Kristin A, Mahzarin R Banaji, Brian A Nosek, and Anthony G Greenwald. n.d. "The Implicit Association Test: IV," 23.

Lane, Richard D., Lee Ryan, Lynn Nadel, and Leslie Greenberg. 2015. "Memory Reconsolidation, Emotional Arousal, and the Process of Change in Psychotherapy: New Insights from Brain Science." *Behavioral and Brain Sciences* 38: e1. https://doi.org/10.1017/S0140525X14000041.

Larson, Christine L., Hillary S. Schaefer, Greg J. Siegle, Cory A.B. Jackson, Michael J. Anderle, and Richard J. Davidson. 2006. "Fear Is Fast in Phobic Individuals: Amygdala Activation in Response to Fear-Relevant Stimuli." *Biological Psychiatry* 60 (4): 410–17. https://doi.org/10.1016/j.biopsych.2006.03.079.

Lascelles, Kristy R. R., and Graham C. L. Davey. 2006. "Successful Differential Evaluative Conditioning Using Simultaneous and Trace Conditioning Procedures in the Picture–Picture Paradigm." *Quarterly Journal of Experimental Psychology* 59 (3): 482–92. https://doi.org/10.1080/02724990444000140.

Lau, Ellen F., Alexandre Gramfort, Matti S. Hämäläinen, and Gina R. Kuperberg. 2013. "Automatic Semantic Facilitation in Anterior Temporal Cortex Revealed through Multimodal Neuroimaging." *The Journal of Neuroscience* 33 (43): 17174–81. https://doi.org/10.1523/JNEUROSCI.1018-13.2013.

Laurita, Anne C., and R. Nathan Spreng. 2017. "The Hippocampus and Social Cognition." In *The Hippocampus from Cells to Systems*, edited by Deborah E. Hannula and Melissa C. Duff, 537–58. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_17.

LeDoux, Joseph, and Nathaniel D. Daw. 2018. "Surviving Threats: Neural Circuit and Computational Implications of a New Taxonomy of Defensive Behaviour." *Nature Reviews Neuroscience* 19 (5): 269–82. https://doi.org/10.1038/nrn.2018.22.

LeDoux, Joseph E. n.d. "Emotion Circuits in the Brain," 31.

Lee, Anne H., Cindy L. Brandon, Jean Wang, and William N. Frost. 2018. "An Argument for Amphetamine-Induced Hallucinations in an Invertebrate." *Frontiers in Physiology* 9 (June): 730. https://doi.org/10.3389/fphys.2018.00730.

Lenaert, Bert, Yannick Boddez, James W. Griffith, Bram Vervliet, Koen Schruers, and Dirk Hermans. 2014. "Aversive Learning and Generalization Predict Subclinical Levels of

Anxiety: A Six-Month Longitudinal Study." *Journal of Anxiety Disorders* 28 (8): 747–53. https://doi.org/10.1016/j.janxdis.2014.09.006.

Levari, David E., Daniel T. Gilbert, Timothy D. Wilson, Beau Sievers, David M. Amodio, and Thalia Wheatley. 2018. "Prevalence-Induced Concept Change in Human Judgment." *Science* 360 (6396): 1465–67. https://doi.org/10.1126/science.aap8731.

Levey, A.B., and Irene Martin. 1975. "Classical Conditioning of Human 'Evaluative' Responses." *Behaviour Research and Therapy* 13 (4): 221–26. https://doi.org/10.1016/0005-7967(75)90026-1.

Levy, Neil. 2015. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements: Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs* 49 (4): 800–823. https://doi.org/10.1111/nous.12074.

Li, Jian, Daniela Schiller, Geoffrey Schoenbaum, Elizabeth A Phelps, and Nathaniel D Daw. 2011. "Differential Roles of Human Striatum and Amygdala in Associative Learning." *Nature Neuroscience* 14 (10): 1250–52. https://doi.org/10.1038/nn.2904.

Li, Tianyi, Carlos Cardenas-Iniguez, Joshua Correll, and Jasmin Cloutier. 2016. "The Impact of Motivation on Race-Based Impression Formation." *NeuroImage* 124 (January): 1–7. https://doi.org/10.1016/j.neuroimage.2015.08.035.

Liefooghe, Baptist, and Jan De Houwer. 2018. "Automatic Effects of Instructions Do Not Require the Intention to Execute These Instructions." *Journal of Cognitive Psychology* 30 (1): 108–21. https://doi.org/10.1080/20445911.2017.1365871.

Lira, Marilia, Julia H. Egito, Patricia A. Dall'Agnol, David M. Amodio, Óscar F. Gonçalves, and Paulo S. Boggio. 2017. "The Influence of Skin Colour on the Experience of Ownership in the Rubber Hand Illusion." *Scientific Reports* 7 (1): 15745. https://doi.org/10.1038/s41598-017-16137-3.

Little, Todd D., William A. Cunningham, Golan Shahar, and Keith F. Widaman. 2002. "To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits." *Structural Equation Modeling: A Multidisciplinary Journal* 9 (2): 151–73. https://doi.org/10.1207/S15328007SEM0902_1.

Loftus, Elizabeth F. 2007. "Elizabeth F. Loftus." In *A History of Psychology in Autobiography, Vol. IX.*, edited by Gardner Lindzey and William M. Runyan, 199–227. Washington: American Psychological Association. https://doi.org/10.1037/11571-006.

Lonsdorf, Tina B., Mareike M. Menz, Marta Andreatta, Miguel A. Fullana, Armita Golkar, Jan Haaker, Ivo Heitland, *et al.* 2017. "Don't Fear 'Fear Conditioning': Methodological Considerations for the Design and Analysis of Studies on Human Fear Acquisition, Extinction, and Return of Fear." *Neuroscience & Biobehavioral Reviews* 77 (June): 247–85. https://doi.org/10.1016/j.neubiorev.2017.02.026.

Lopresto, Dora, Pieter Schipper, and Judith R. Homberg. 2016. "Neural Circuits and Mechanisms Involved in Fear Generalization: Implications for the Pathophysiology and Treatment of Posttraumatic Stress Disorder." *Neuroscience & Biobehavioral Reviews* 60 (January): 31–42. https://doi.org/10.1016/j.neubiorev.2015.10.009.

Lott, Bernice. 2014. "Social Class Myopia: The Case of Psychology and Labor Unions: Psychology and Labor Unions." *Analyses of Social Issues and Public Policy* 14 (1): 261–80. https://doi.org/10.1111/asap.12029.

Lovibond, Peter F. n.d. "Cognitive Processes in Extinction," 7.

Lovibond, Peter F., and David R. Shanks. 2002. "The Role of Awareness in Pavlovian Conditioning: Empirical Evidence and Theoretical Implications." *Journal of Experimental Psychology: Animal Behavior Processes* 28 (1): 3–26. https://doi.org/10.1037/0097-7403.28.1.3.

Lucas, Heather D., Jessica D. Creery, Xiaoqing Hu, and Ken A. Paller. 2019. "Grappling With Implicit Social Bias: A Perspective From Memory Research." *Neuroscience* 406 (May): 684–97. https://doi.org/10.1016/j.neuroscience.2019.01.037.

Luck, Camilla C., and Ottmar V. Lipp. 2017. "Startle Modulation and Explicit Valence Evaluations Dissociate during Backward Fear Conditioning: Dissociation between Startle and CS Valence." *Psychophysiology* 54 (5): 673–83. https://doi.org/10.1111/psyp.12834.

Maas, Andrew L, and Charles Kemp. n.d. "One-Shot Learning with Bayesian Networks," 6.

Mace, John H., Megan L. McQueen, Kamille E. Hayslett, Bobbie Jo A. Staley, and Talia J. Welch. 2019. "Semantic Memories Prime Autobiographical Memories: General Implications and Implications for Everyday Autobiographical Remembering." *Memory & Cognition* 47 (2): 299–312. https://doi.org/10.3758/s13421-018-0866-9.

Machery, Edouard. 2005. "Concepts Are Not a Natural Kind*." *Philosophy of Science* 72 (3): 444–67. https://doi.org/10.1086/498473.

———. 2010. "Pre´ Cis of Doing without Concepts." *BEHAVIORAL AND BRAIN SCIENCES*, 50.

———. 2017. "Do Indirect Measures of Biases Measure Traits or Situations?" *Psychological Inquiry* 28 (4): 288–91. https://doi.org/10.1080/1047840X.2017.1373557.

Madan, Christopher R., Esther Fujiwara, Jeremy B. Caplan, and Tobias Sommer. 2017. "Emotional Arousal Impairs Association-Memory: Roles of Amygdala and Hippocampus." *NeuroImage* 156 (August): 14–28. https://doi.org/10.1016/j.neuroimage.2017.04.065.

Madva, Alex. 2016. "Why Implicit Attitudes Are (Probably) Not Beliefs." *Synthese* 193 (8): 2659–84. https://doi.org/10.1007/s11229-015-0874-2.

Madva, Alex, and Michael Brownstein. 2018. "Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind [1]: Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind." *Noûs* 52 (3): 611–44. https://doi.org/10.1111/nous.12182.

Mahr, Johannes, and Gergely Csibra. 2017. "Why Do We Remember? The Communicative Function of Episodic Memory." *The Behavioral and Brain Sciences*, January, 1–93. https://doi.org/10.1017/S0140525X17000012.

Maister, Lara, Mel Slater, Maria V. Sanchez-Vives, and Manos Tsakiris. 2015. "Changing Bodies Changes Minds: Owning Another Body Affects Social Cognition." *Trends in Cognitive Sciences* 19 (1): 6–12. https://doi.org/10.1016/j.tics.2014.11.001.

Mallan, Kimberley M., James Sax, and Ottmar V. Lipp. 2009. "Verbal Instruction Abolishes Fear Conditioned to Racial Out-Group Faces." *Journal of Experimental Social Psychology* 45 (6): 1303–7. https://doi.org/10.1016/j.jesp.2009.08.001.

Mallon, Ron. 2006. "'Race': Normative, Not Metaphysical or Semantic." *Ethics* 116 (3): 525–51. https://doi.org/10.1086/500495.

Mandalaywala, Tara M., Gabrielle Ranger-Murdock, David M. Amodio, and Marjorie Rhodes. 2019. "The Nature and Consequences of Essentialist Beliefs About Race in Early Childhood." *Child Development* 90 (4): e437–53. https://doi.org/10.1111/cdev.13008.

Mandelbaum, Eric. 2014. "Thinking Is Believing." *Inquiry* 57 (1): 55–96. https://doi.org/10.1080/0020174X.2014.858417.

Mann, Thomas C, and Jeremy Cone. n.d. "Updating Implicit Impressions: New Evidence on Intentionality and the Affect Misattribution Procedure," 26.

Mann, Thomas C., Benedek Kurdi, and Mahzarin R. Banaji. 2019. "How Effectively Can Implicit Evaluations Be Updated? Using Evaluative Statements after Aversive Repeated Evaluative

Pairings." *Journal of Experimental Psychology: General*, October. https://doi.org/10.1037/xge0000701.

Manns, Joseph R., Robert E. Clark, and Larry R. Squire. 2000. "Parallel Acquisition of Awareness and Trace Eyeblink Classical Conditioning." *Learning & Memory* 7 (5): 267–72.

March, David S., Michael A. Olson, and Lowell Gaertner. 2020. "Lions, and Tigers, and Implicit Measures, Oh My! Implicit Assessment and the Valence vs. Threat Distinction." *Social Cognition* 38 (Supplement): s154–64. https://doi.org/10.1521/soco.2020.38.supp.s154.

March, David S., Micheal A. Olson, and Russell H. Fazio. 2018. "The Implicit Misattribution Model of Evaluative Conditioning." *Social Psychological Bulletin* 13 (3): e27574. https://doi.org/10.5964/spb.v13i3.27574.

Maren, Stephen. 2016. "Parsing Reward and Aversion in the Amygdala." *Neuron* 90 (2): 209–11. https://doi.org/10.1016/j.neuron.2016.04.011.

Maren, Stephen, K. Luan Phan, and Israel Liberzon. 2013. "The Contextual Brain: Implications for Fear Conditioning, Extinction and Psychopathology." *Nature Reviews Neuroscience* 14 (6): 417–28. https://doi.org/10.1038/nrn3492.

Marini, Maddalena, Mahzarin R. Banaji, and Alvaro Pascual-Leone. 2018. "Studying Implicit Social Cognition with Noninvasive Brain Stimulation." *Trends in Cognitive Sciences* 22 (11): 1050–66. https://doi.org/10.1016/j.tics.2018.07.014.

Martin, I. n.d. "EVALUATIVE CONDITIONING," 45.

Martínez, Manolo. 2015. "Informationally-Connected Property Clusters, and Polymorphism." *Biology & Philosophy* 30 (1): 99–117. https://doi.org/10.1007/s10539-014-9443-1.

Mather, Mara, David Clewett, Michiko Sakaki, and Carolyn W. Harley. 2016. "Norepinephrine Ignites Local Hotspots of Neuronal Excitation: How Arousal Amplifies Selectivity in

Perception and Memory." *Behavioral and Brain Sciences* 39: e200. https://doi.org/10.1017/S0140525X15000667.

Mattan, Bradley D., Jennifer T. Kubota, Tianyi Li, Samuel A. Venezia, and Jasmin Cloutier. 2019. "Implicit Evaluative Biases Toward Targets Varying in Race and Socioeconomic Status." *Personality and Social Psychology Bulletin* 45 (10): 1512–27. https://doi.org/10.1177/0146167219835230.

Mattan, Bradley D, Kevin Y Wei, Jasmin Cloutier, and Jennifer T Kubota. 2018. "The Social Neuroscience of Race-Based and Status-Based Prejudice." *Current Opinion in Psychology* 24 (December): 27–34. https://doi.org/10.1016/j.copsyc.2018.04.010.

McClelland, James L. 2013. "Incorporating Rapid Neocortical Learning of New Schema-Consistent Information into Complementary Learning Systems Theory." *Journal of Experimental Psychology. General* 142 (4): 1190–1210. https://doi.org/10.1037/a0033812.

McClelland, James L, and Randall C O'Reilly. n.d. "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex:InsightsFrom the Successesand Failuresof Connectionist Models of Learning and Memory," 39.

McCormick, Cornelia, Elisa Ciaramelli, Flavia De Luca, and Eleanor A. Maguire. 2018. "Comparing and Contrasting the Cognitive Effects of Hippocampal and Ventromedial Prefrontal Cortex Damage: A Review of Human Lesion Studies." *Neuroscience* 374 (March): 295–318. https://doi.org/10.1016/j.neuroscience.2017.07.066.

McCullough, Michael E., Robert Kurzban, and Benjamin A. Tabak. 2013. "Cognitive Systems for Revenge and Forgiveness." *Behavioral and Brain Sciences* 36 (1): 1–15. https://doi.org/10.1017/S0140525X11002160.

McDannald, Michael A., Yuji K. Takahashi, Nina Lopatina, Brad W. Pietras, Josh L. Jones, and Geoffrey Schoenbaum. 2012. "Model-Based Learning and the Contribution of the

Orbitofrontal Cortex to the Model-Free World: OFC in Model-Based Learning." *European Journal of Neuroscience* 35 (7): 991–96. https://doi.org/10.1111/j.1460-9568.2011.07982.x.

Mechias, Marie-Luise, Amit Etkin, and Raffael Kalisch. 2010. "A Meta-Analysis of Instructed Fear Studies: Implications for Conscious Appraisal of Threat." *NeuroImage* 49 (2): 1760–68. https://doi.org/10.1016/j.neuroimage.2009.09.040.

"Medial Prefrontal Cortex and Pavlovian Conditioning: Trace versus Delay Conditioning. - PsycNET." n.d. Accessed January 28, 2020. https://psycnet.apa.org/record/2002-10110-004.

Meer, Matthijs A.A. van der, Adam Johnson, Neil C. Schmitzer-Torbert, and A. David Redish. 2010. "Triple Dissociation of Information Processing in Dorsal Striatum, Ventral Striatum, and Hippocampus on a Learned Spatial Decision Task." *Neuron* 67 (1): 25–32. https://doi.org/10.1016/j.neuron.2010.06.023.

Meiran, Nachshon, Maayan Pereg, Yoav Kessler, Michael W. Cole, and Todd S. Braver. 2015. "The Power of Instructions: Proactive Configuration of Stimulus–Response Translation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41 (3): 768–86. https://doi.org/10.1037/xlm0000063.

Melnikoff, David E., and John A. Bargh. 2018. "The Mythical Number Two." *Trends in Cognitive Sciences* 22 (4): 280–93. https://doi.org/10.1016/j.tics.2018.02.001.

"Memory Consolidation in Both Trace and Delay Fear Conditioning Is Disrupted by Intra-Amygdala Infusion of the Protein Synthesis Inhibitor Anisomycin." n.d. Accessed January 28, 2020. http://learnmem.cshlp.org/content/18/11/728.full.

Mende-Siedlecki, Peter. 2018. "Changing Our Minds: The Neural Bases of Dynamic Impression Updating." *Current Opinion in Psychology*, Social Neuroscience, 24 (December): 72–76. https://doi.org/10.1016/j.copsyc.2018.08.007.

Mende-Siedlecki, Peter, Yang Cai, and Alexander Todorov. 2013. "The Neural Dynamics of Updating Person Impressions." *Social Cognitive and Affective Neuroscience* 8 (6): 623–31. https://doi.org/10.1093/scan/nss040.

Merck, Catherine, Pierre-Yves Jonin, Mickaël Laisney, Hélène Vichard, and Serge Belliard. 2014. "When the Zebra Loses Its Stripes but Is Still in the Savannah: Results from a Semantic Priming Paradigm in Semantic Dementia." *Neuropsychologia* 53 (January): 221–32. https://doi.org/10.1016/j.neuropsychologia.2013.11.024.

Mertens, Gaëtan, Yannick Boddez, Angelos-Miltiadis Krypotos, and Iris M. Engelhard. 2021. "Human Fear Conditioning Is Moderated by Stimulus Contingency Instructions." *Biological Psychology* 158 (January): 107994. https://doi.org/10.1016/j.biopsycho.2020.107994.

Mertens, Gaëtan, Yannick Boddez, Dieuwke Sevenster, Iris M. Engelhard, and Jan De Houwer. 2018. "A Review on the Effects of Verbal Instructions in Human Fear Conditioning: Empirical Findings, Theoretical Considerations, and Future Directions." *Biological Psychology* 137 (September): 49–64. https://doi.org/10.1016/j.biopsycho.2018.07.002.

Mertens, Gaëtan, and Iris M. Engelhard. 2020. "A Systematic Review and Meta-Analysis of the Evidence for Unaware Fear Conditioning." *Neuroscience & Biobehavioral Reviews* 108 (January): 254–68. https://doi.org/10.1016/j.neubiorev.2019.11.012.

Mesulam, M.-Marsel, ed. 2000. *Principles of Behavioral and Cognitive Neurology*. 2nd ed. Oxford ; New York: Oxford University Press.

Meulders, Ann. 2019. "From Fear of Movement-Related Pain and Avoidance to Chronic Pain Disability: A State-of-the-Art Review." *Current Opinion in Behavioral Sciences* 26 (April): 130–36. https://doi.org/10.1016/j.cobeha.2018.12.007.

Meusburger, Peter. 2016. *Knowledge and Action*. New York, NY: Springer Berlin Heidelberg.

Meyer, Meghan L. 2019. "Social by Default: Characterizing the Social Functions of the Resting Brain." *Current Directions in Psychological Science* 28 (4): 380–86. https://doi.org/10.1177/0963721419857759.

Mierop, Adrien, Mikael Molet, and Olivier Corneille. 2019. "Response Production during Extinction Training Is Not Sufficient for Extinction of Evaluative Conditioning." *Cognition and Emotion* 33 (6): 1181–95. https://doi.org/10.1080/02699931.2018.1545633.

Miloyan, Beyon, Kimberley A. McFarlane, and Thomas Suddendorf. 2019. "Measuring Mental Time Travel: Is the Hippocampus Really Critical for Episodic Memory and Episodic Foresight?" *Cortex* 117 (August): 371–84. https://doi.org/10.1016/j.cortex.2019.01.020.

Mine, Chisato, and Jun Saiki. 2018. "Pavlovian Reward Learning Elicits Attentional Capture by Reward-Associated Stimuli." *Attention, Perception, & Psychophysics* 80 (5): 1083–95. https://doi.org/10.3758/s13414-018-1502-2.

Mioni, Giovanna, Peter G. Rendell, Gill Terrett, and Franca Stablum. 2015. "Prospective Memory Performance in Traumatic Brain Injury Patients: A Study of Implementation Intentions." *Journal of the International Neuropsychological Society* 21 (4): 305–13. https://doi.org/10.1017/S1355617715000211.

Mitchell, Chris J., Jan De Houwer, and Peter F. Lovibond. 2009. "The Propositional Nature of Human Associative Learning." *Behavioral and Brain Sciences* 32 (2): 183–98. https://doi.org/10.1017/S0140525X09000855.

Mneimne, Malek, Alice S. Powers, Kate E. Walton, David S. Kosson, Samantha Fonda, and Jessica Simonetti. 2010. "Emotional Valence and Arousal Effects on Memory and Hemispheric Asymmetries." *Brain and Cognition* 74 (1): 10–17. https://doi.org/10.1016/j.bandc.2010.05.011.

Mody, Shilpa, and Susan Carey. 2016. "The Emergence of Reasoning by the Disjunctive Syllogism in Early Childhood." *Cognition* 154 (September): 40–48. https://doi.org/10.1016/j.cognition.2016.05.012.

Monroe, Andrew E., and Bertram F. Malle. 2019. "People Systematically Update Moral Judgments of Blame." *Journal of Personality and Social Psychology* 116 (2): 215–36. https://doi.org/10.1037/pspa0000137.

Montagrin, Alison, Catarina Saiote, and Daniela Schiller. 2018. "The Social Hippocampus: M ontagrin et al ." *Hippocampus* 28 (9): 672–79. https://doi.org/10.1002/hipo.22797.

Moran, Tal, and Yoav Bar-Anan. 2013. "The Effect of Object–Valence Relations on Automatic Evaluation." *Cognition & Emotion* 27 (4): 743–52. https://doi.org/10.1080/02699931.2012.732040.

———. 2020. "The Effect of Co-Occurrence and Relational Information on Speeded Evaluation." *Cognition and Emotion* 34 (1): 144–55. https://doi.org/10.1080/02699931.2019.1604321.

Moran, Tal, Yoav Bar-Anan, Bar-Anan Lab, and Tzipora Dror. 2020. "Testing the Judgment-Related Account for the Extinction of Evaluative Conditioning." https://doi.org/10.17605/OSF.IO/XUHPT.

Moran, Tal, Yoav Bar-Anan, and Brian A. Nosek. 2015. "Processing Goals Moderate the Effect of Co-Occurrence on Automatic Evaluation." *Journal of Experimental Social Psychology* 60 (September): 157–62. https://doi.org/10.1016/j.jesp.2015.05.009.

———. 2017. "The Effect of the Validity of Co-Occurrence on Automatic and Deliberate Evaluations: Validity and Automatic Evaluations." *European Journal of Social Psychology* 47 (6): 708–23. https://doi.org/10.1002/ejsp.2266.

Moran, Tal, Yoav Bar-Anan, and Brian A Nosek. n.d. "The Assimilative Effect of Co-Occurrence on Evaluation Above and Beyond the Effect of Relational Qualifiers," 28.

Morishima, Yosuke, Daniel Schunk, Adrian Bruhin, Christian C. Ruff, and Ernst Fehr. 2012. "Linking Brain Structure and Activation in Temporoparietal Junction to Explain the Neurobiology of Human Altruism." *Neuron* 75 (1): 73–79. https://doi.org/10.1016/j.neuron.2012.05.021.

Morris, Adam, and Fiery Cushman. 2018. "A COMMON FRAMEWORK FOR THEORIES OF NORM COMPLIANCE." *Social Philosophy and Policy* 35 (1): 101–27. https://doi.org/10.1017/S0265052518000134.

Moscovitch, Morris. 2008. "The Hippocampus as a 'Stupid,' Domain-Specific Module: Implications for Theories of Recent and Remote Memory, and of Imagination." *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 62 (1): 62–79. https://doi.org/10.1037/1196-1961.62.1.62.

Moscovitch, Morris, Roberto Cabeza, Gordon Winocur, and Lynn Nadel. 2016. "Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation." *Annual Review of Psychology* 67 (1): 105–34. https://doi.org/10.1146/annurev-psych-113011-143733.

Moutoussis, K., and S. Zeki. 2002. "The Relationship between Cortical Activation and Perception Investigated with Invisible Stimuli." *Proceedings of the National Academy of Sciences* 99 (14): 9527–32. https://doi.org/10.1073/pnas.142305699.

Mudrik, Liad, Nathan Faivre, and Christof Koch. 2014. "Information Integration without Awareness." *Trends in Cognitive Sciences* 18 (9): 488–96. https://doi.org/10.1016/j.tics.2014.04.009.

———. 2016. "Information Integration without Awareness." *Trends in Cognitive Sciences* 20 (7): 559. https://doi.org/10.1016/j.tics.2016.05.005.

Mullally, Sinéad L., Faraneh Vargha-Khadem, and Eleanor A. Maguire. 2014. "Scene Construction in Developmental Amnesia: An FMRI Study." *Neuropsychologia* 52 (January): 1–10. https://doi.org/10.1016/j.neuropsychologia.2013.11.001.

Murphy, Dominic, and Stephen Stich. n.d. "GRIFFITHS, ELIMINATION & PSYCHOPATHOLOGY," 10.

Murray, Elisabeth A. 2007. "The Amygdala, Reward and Emotion." *Trends in Cognitive Sciences* 11 (11): 489–97. https://doi.org/10.1016/j.tics.2007.08.013.

Murray, Elisabeth A., Steven P. Wise, and Kim S. Graham. 2017. *The Evolution of Memory Systems: Ancestors, Anatomy, and Adaptations*. First Edition. Oxford, United Kingdom ; New York, NY: Oxford University Press.

Murray, Hugh T. 1967. "The NAACP versus the Communist Party: The Scottsboro Rape Cases, 1931-1932." *Phylon (1960-)* 28 (3): 276. https://doi.org/10.2307/273666.

Nadel, Lynn, and Oliver Hardt. 2011. "Update on Memory Systems and Processes." *Neuropsychopharmacology* 36 (1): 251–73. https://doi.org/10.1038/npp.2010.169.

Nakamura, Kimihiro, Tomoe Inomata, and Akira Uno. 2020. "Left Amygdala Regulates the Cerebral Reading Network During Fast Emotion Word Processing." *Frontiers in Psychology* 11 (January): 1. https://doi.org/10.3389/fpsyg.2020.00001.

Navarrete, Carlos David, Andreas Olsson, Arnold K. Ho, Wendy Berry Mendes, Lotte Thomsen, and James Sidanius. 2009. "Fear Extinction to an Out-Group Face: The Role of Target Gender." *Psychological Science* 20 (2): 155–58. https://doi.org/10.1111/j.1467-9280.2009.02273.x.

Neumann, Roland, and Ljubica Lozo. 2012. "Priming the Activation of Fear and Disgust: Evidence for Semantic Processing." *Emotion* 12 (2): 223–28. https://doi.org/10.1037/a0026500.

“Neural Substrates Underlying Human Delay and Trace Eyeblink Conditioning | PNAS.” n.d. Accessed January 28, 2020. https://www.pnas.org/content/105/23/8108.

Newman, Ian R., Maia Gibb, and Valerie A. Thompson. 2017. “Rule-Based Reasoning Is Fast and Belief-Based Reasoning Can Be Slow: Challenging Current Explanations of Belief-Bias and Base-Rate Neglect.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43 (7): 1154–70. https://doi.org/10.1037/xlm0000372.

Niedenthal, Paula M., Lawrence W. Barsalou, Piotr Winkielman, Silvia Krauth-Gruber, and François Ric. 2005. “Embodiment in Attitudes, Social Perception, and Emotion.” *Personality and Social Psychology Review* 9 (3): 184–211. https://doi.org/10.1207/s15327957pspr0903_1.

Norcross, Alastair. 2004. “Puppies, Pigs, and People: Eating Meat and Marginal Cases.” *Philosophical Perspectives* 18 (1): 229–45. https://doi.org/10.1111/j.1520-8583.2004.00027.x.

Norman, Kenneth A., and Randall C. O’Reilly. 2003. “Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach.” *Psychological Review* 110 (4): 611–46. https://doi.org/10.1037/0033-295X.110.4.611.

Nosek, Brian A. n.d. “Implicit and Explicit Attitudes Are Related but Distinct Constructs,” 36.

Nosek, Brian A, and Mahzarin R Banaji. n.d. “THE GO/NO-GO ASSOCIATION TASK,” 22.

Nosek, Brian A, Anthony G Greenwald, and Mahzarin R Banaji. n.d. “The Implicit Association Test at Age 7: A Methodological and Conceptual Review,” 28.

O’Callaghan, Claire, Kestutis Kveraga, James M. Shine, Reginald B. Adams, and Moshe Bar. 2017. “Predictions Penetrate Perception: Converging Insights from Brain, Behaviour and Disorder.” *Consciousness and Cognition* 47 (January): 63–74. https://doi.org/10.1016/j.concog.2016.05.003.

Olson, Ingrid R., David McCoy, Elizabeth Klobusicky, and Lars A. Ross. 2013. "Social Cognition and the Anterior Temporal Lobes: A Review and Theoretical Framework." *Social Cognitive and Affective Neuroscience* 8 (2): 123–33. https://doi.org/10.1093/scan/nss119.

Olsson, A. 2005. "The Role of Social Groups in the Persistence of Learned Fear." *Science* 309 (5735): 785–87. https://doi.org/10.1126/science.1113551.

Olsson, Andreas, Katherine I. Nearing, and Elizabeth A. Phelps. 2007. "Learning Fears by Observing Others: The Neural Systems of Social Fear Transmission." *Social Cognitive and Affective Neuroscience* 2 (1): 3–11. https://doi.org/10.1093/scan/nsm005.

Olsson, Andreas, and Elizabeth A. Phelps. 2004. "Learned Fear of 'Unseen' Faces after Pavlovian, Observational, and Instructed Fear." *Psychological Science* 15 (12): 822–28. https://doi.org/10.1111/j.0956-7976.2004.00762.x.

Operskalski, Joachim T., and Aron K. Barbey. 2017. *Cognitive Neuroscience of Causal Reasoning*. Edited by Michael R. Waldmann. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.16.

O'Reilly, Randall C., Rajan Bhattacharyya, Michael D. Howard, and Nicholas Ketz. 2014. "Complementary Learning Systems." *Cognitive Science* 38 (6): 1229–48. https://doi.org/10.1111/j.1551-6709.2011.01214.x.

O'Reilly, Randall C., Thomas E. Hazy, and Seth A. Herd. 2017a. "The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win!" In *The Oxford Handbook of Cognitive Science*, 91–115. New York, NY, US: Oxford University Press.

———. 2017b. "The Leabra Cognitive Architecture." *The Oxford Handbook of Cognitive Science*, October. https://doi.org/10.1093/oxfordhb/9780199842193.013.8.

O'Reilly, Randall C, Thomas E Hazy, and Seth A Herd. n.d. "The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win!," 31.

O'Reilly, Randall C, and Kenneth A Norman. 2002. "Hippocampal and Neocortical Contributions to Memory: Advances in the Complementary Learning Systems Framework." *Trends in Cognitive Sciences* 6 (12): 505–10. https://doi.org/10.1016/S1364-6613(02)02005-3.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard, and Philip E. Tetlock. 2013. "Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies." *Journal of Personality and Social Psychology* 105 (2): 171–92. https://doi.org/10.1037/a0032734.

Otto, A. R., C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw. 2013. "Working-Memory Capacity Protects Model-Based Learning from Stress." *Proceedings of the National Academy of Sciences* 110 (52): 20941–46. https://doi.org/10.1073/pnas.1312011110.

Otto, A. Ross, Anya Skatova, Seth Madlon-Kay, and Nathaniel D. Daw. 2015. "Cognitive Control Predicts Use of Model-Based Reinforcement Learning." *Journal of Cognitive Neuroscience* 27 (2): 319–33. https://doi.org/10.1162/jocn_a_00709.

Parfit, Derek. 2017. "Future People, the Non-Identity Problem, and Person-Affecting Principles: Future People, the Non-Identity Problem, and Person-Affecting Principles." *Philosophy & Public Affairs* 45 (2): 118–57. https://doi.org/10.1111/papa.12088.

Parikh, Natasha, Luka Ruzic, Gregory W. Stewart, R. Nathan Spreng, and Felipe De Brigard. 2018. "What If? Neural Activity Underlying Semantic and Episodic Counterfactual Thinking." *NeuroImage* 178 (September): 332–45. https://doi.org/10.1016/j.neuroimage.2018.05.053.

Paul, L. A., and Edward J. Hall. 2013. *Causation: A User's Guide*. First edition. Oxford: Oxford University Press.

Payne, B. Keith, Clara Michelle Cheng, Olesya Govorun, and Brandon D. Stewart. 2005. "An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement." *Journal of Personality and Social Psychology* 89 (3): 277–93. https://doi.org/10.1037/0022-3514.89.3.277.

Payne, B. Keith, Olesya Govorun, and Nathan L. Arbuckle. 2008. "Automatic Attitudes and Alcohol: Does Implicit Liking Predict Drinking?" *Cognition and Emotion* 22 (2): 238–71. https://doi.org/10.1080/02699930701357394.

Payne, B. Keith, Heidi A. Vuletich, and Kristjen B. Lundberg. 2017. "The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice." *Psychological Inquiry* 28 (4): 233–48. https://doi.org/10.1080/1047840X.2017.1335568.

Payne, Keith, and Kristjen Lundberg. 2014. "The Affect Misattribution Procedure: Ten Years of Evidence on Reliability, Validity, and Mechanisms: Affect Misattribution Procedure." *Social and Personality Psychology Compass* 8 (12): 672–86. https://doi.org/10.1111/spc3.12148.

"(PDF) Incorporating Rapid Neocortical Learning of New Schema-Consistent Information Into Complementary Learning Systems Theory." n.d. Accessed October 26, 2019. https://www.researchgate.net/publication/256117997_Incorporating_Rapid_Neocortical_Learning_of_New_Schema-Consistent_Information_Into_Complementary_Learning_Systems_Theory.

"(PDF) The Heterogeneity of Implicit Bias." n.d. Accessed September 30, 2019. https://www.researchgate.net/publication/289505154_The_Heterogeneity_of_Implicit_Bias.

Pearson, Carol, Ann Elizabeth Montgomery, and Gretchen Locke. 2009. "Housing Stability among Homeless Individuals with Serious Mental Illness Participating in Housing First Programs." *Journal of Community Psychology* 37 (3): 404–17. https://doi.org/10.1002/jcop.20303.

Peira, Nathalie, Mats Fredrikson, and Gilles Pourtois. 2014. "Controlling the Emotional Heart: Heart Rate Biofeedback Improves Cardiac Control during Emotional Reactions." *International Journal of Psychophysiology* 91 (3): 225–31. https://doi.org/10.1016/j.ijpsycho.2013.12.008.

Pennycook, Gordon, Wim De Neys, Jonathan St. B.T. Evans, Keith E. Stanovich, and Valerie A. Thompson. 2018. "The Mythical Dual-Process Typology." *Trends in Cognitive Sciences* 22 (8): 667–68. https://doi.org/10.1016/j.tics.2018.04.008.

Persuh, Marjan, Eric LaRock, and Jacob Berger. 2018. "Working Memory and Consciousness: The Current State of Play." *Frontiers in Human Neuroscience* 12 (March): 78. https://doi.org/10.3389/fnhum.2018.00078.

Peters, Kurt R., and Bertram Gawronski. 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): 557–69. https://doi.org/10.1177/0146167211400423.

Petty, Richard E. 2012. *Attitudes: Insights from the New Implicit Measures*. 1st ed. Psychology Press. https://doi.org/10.4324/9780203809884.

Pezzulo, Giovanni, Francesco Rigoli, and Karl J. Friston. 2018. "Hierarchical Active Inference: A Theory of Motivated Control." *Trends in Cognitive Sciences* 22 (4): 294–306. https://doi.org/10.1016/j.tics.2018.01.009.

Pfeiffer, Brad E., and David J. Foster. 2013. "Hippocampal Place-Cell Sequences Depict Future Paths to Remembered Goals." *Nature* 497 (7447): 74–79. https://doi.org/10.1038/nature12112.

Phelps, Elizabeth A. 2004. "Human Emotion and Memory: Interactions of the Amygdala and Hippocampal Complex." *Current Opinion in Neurobiology* 14 (2): 198–202. https://doi.org/10.1016/j.conb.2004.03.015.

Phelps, Elizabeth A. 2006. "Emotion and Cognition: Insights from Studies of the Human Amygdala." *Annual Review of Psychology* 57 (1): 27–53. https://doi.org/10.1146/annurev.psych.56.091103.070234.

Phelps, Elizabeth A, Christopher J Cannistraci, and William A Cunningham. 2003. "Intact Performance on an Indirect Measure of Race Bias Following Amygdala Damage." *Neuropsychologia* 41 (2): 203–8. https://doi.org/10.1016/S0028-3932(02)00150-1.

Phelps, Elizabeth A, Kevin J O'Connor, William A Cunningham, E Sumie Funayama, J Christopher Gatenby, John C Gore, and Mahzarin R Banaji. n.d. "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation" 12 (5): 31.

Philips, Michael. 1984. "Racist Acts and Racist Humor." *Canadian Journal of Philosophy* 14 (1): 75–96.

Picard, Laurence, Claire Mayor-Dubois, Philippe Maeder, Sandrine Kalenzaga, Maria Abram, Céline Duval, Francis Eustache, Eliane Roulet-Perez, and Pascale Piolino. 2013. "Functional Independence within the Self-Memory System: New Insights from Two Cases of Developmental Amnesia." *Cortex* 49 (6): 1463–81. https://doi.org/10.1016/j.cortex.2012.10.003.

Pichon, Swann, Sebastian W. Rieger, and Patrik Vuilleumier. 2012. "Persistent Affective Biases in Human Amygdala Response Following Implicit Priming with Negative Emotion Concepts." *NeuroImage* 62 (3): 1610–21. https://doi.org/10.1016/j.neuroimage.2012.06.004.

Pine, Alex, Noa Sadeh, Aya Ben-Yakov, Yadin Dudai, and Avi Mendelsohn. 2018. "Knowledge Acquisition Is Governed by Striatal Prediction Errors." *Nature Communications* 9 (1): 1673. https://doi.org/10.1038/s41467-018-03992-5.

Plotka, Irina, and Nina Blumenau. 2015. "Implicit Methods for Studying Attitudes: Modern Approach to Research in Social Sciences." *SOCIETY, INTEGRATION, EDUCATION. Proceedings of the International Scientific Conference* 1 (May): 303. https://doi.org/10.17770/sie2015vol1.311.

Pohlack, Sebastian T., Frauke Nees, Michaela Ruttorf, Lothar R. Schad, and Herta Flor. 2012. "Activation of the Ventral Striatum during Aversive Contextual Conditioning in Humans." *Biological Psychology* 91 (1): 74–80. https://doi.org/10.1016/j.biopsycho.2012.04.004.

Pojman, Louis P., ed. 2000. *The Moral Life: An Introductory Reader in Ethics and Literature*. New York: Oxford University Press.

Poldrack, Russell A, and Mark G Packard. 2003. "Competition among Multiple Memory Systems: Converging Evidence from Animal and Human Brain Studies." *Neuropsychologia* 41 (3): 245–51. https://doi.org/10.1016/S0028-3932(02)00157-4.

Preckel, Katrin, Fynn-Mathis Trautwein, Frieder M. Paulus, Peter Kirsch, Sören Krach, Tania Singer, and Philipp Kanske. 2019. "Neural Mechanisms of Affective Matching across Faces and Scenes." *Scientific Reports* 9 (1): 1–10. https://doi.org/10.1038/s41598-018-37163-9.

Prével, Arthur, Vinca Rivière, Jean-Claude Darcheville, Gonzalo P Urcelay, and Ralph R Miller. 2019. "Excitatory Second-Order Conditioning Using a Backward First-Order Conditioned Stimulus: A Challenge for *Prediction* Error Reduction." *Quarterly Journal of Experimental Psychology* 72 (6): 1453–65. https://doi.org/10.1177/1747021818793376.

Prévost, Charlotte, Daniel McNamee, Ryan K. Jessup, Peter Bossaerts, and John P. O'Doherty. 2013. "Evidence for Model-Based Computations in the Human Amygdala during Pavlovian Conditioning." Edited by Olaf Sporns. *PLoS Computational Biology* 9 (2): e1002918. https://doi.org/10.1371/journal.pcbi.1002918.

Price, Mason H., and Jeffrey D. Johnson. 2017. "Failure to Reactivate Salient Episodic Information during Indirect and Direct Tests of Memory Retrieval." Preprint. Neuroscience. https://doi.org/10.1101/189522.

Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. Philosophy of Mind Series. Oxford ; New York: Oxford University Press.

Quadflieg, Susanne, Francesco Gentile, and Bruno Rossion. 2015. "The Neural Basis of Perceiving Person Interactions." *Cortex* 70 (September): 5–20. https://doi.org/10.1016/j.cortex.2014.12.020.

Quamme, Joel R., Andrew P. Yonelinas, and Kenneth A. Norman. 2007. "Effect of Unitization on Associative Recognition in Amnesia." *Hippocampus* 17 (3): 192–200. https://doi.org/10.1002/hipo.20257.

Rabin, Jennifer S., Anna Braverman, Asaf Gilboa, Donald T. Stuss, and R. Shayna Rosenbaum. 2012. "Theory of Mind Development Can Withstand Compromised Episodic Memory Development." *Neuropsychologia* 50 (14): 3781–85. https://doi.org/10.1016/j.neuropsychologia.2012.10.016.

Race, Elizabeth, Keely Burke, and Mieke Verfaellie. 2019. "Repetition Priming in Amnesia: Distinguishing Associative Learning at Different Levels of Abstraction." *Neuropsychologia* 122 (January): 98–104. https://doi.org/10.1016/j.neuropsychologia.2018.11.007.

Raposo, A., H.E. Moss, E.A. Stamatakis, and L.K. Tyler. 2006. "Repetition Suppression and Semantic Enhancement: An Investigation of the Neural Correlates of Priming." *Neuropsychologia* 44 (12): 2284–95. https://doi.org/10.1016/j.neuropsychologia.2006.05.017.

Ravdin, Lisa D., and Heather L. Katzen, eds. 2019. *Handbook on the Neuropsychology of Aging and Dementia*. Clinical Handbooks in Neuropsychology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93497-6.

Reber, Paul J. 2013. "The Neural Basis of Implicit Learning and Memory: A Review of Neuropsychological and Neuroimaging Research." *Neuropsychologia* 51 (10): 2026–42. https://doi.org/10.1016/j.neuropsychologia.2013.06.019.

Reber, T. P., R. Luechinger, P. Boesiger, and K. Henke. 2012. "Unconscious Relational Inference Recruits the Hippocampus." *Journal of Neuroscience* 32 (18): 6138–48. https://doi.org/10.1523/JNEUROSCI.5639-11.2012.

Reder, Lynne M., Heekyeong Park, and Paul D. Kieffaber. 2009. "Memory Systems Do Not Divide on Consciousness: Reinterpreting Memory in Terms of Activation and Binding." *Psychological Bulletin* 135 (1): 23–49. https://doi.org/10.1037/a0013974.

Renoult, Louis, Patrick S.R. Davidson, Daniela J. Palombo, Morris Moscovitch, and Brian Levine. 2012. "Personal Semantics: At the Crossroads of Semantic and Episodic Memory." *Trends in Cognitive Sciences* 16 (11): 550–58. https://doi.org/10.1016/j.tics.2012.09.003.

Renoult, Louis, Muireann Irish, Morris Moscovitch, and Michael D. Rugg. 2019. "From Knowing to Remembering: The Semantic–Episodic Distinction." *Trends in Cognitive Sciences* 23 (12): 1041–57. https://doi.org/10.1016/j.tics.2019.09.008.

Rezaei, Ali R. 2011. "Validity and Reliability of the IAT: Measuring Gender and Ethnic Stereotypes." *Computers in Human Behavior* 27 (5): 1937–41. https://doi.org/10.1016/j.chb.2011.04.018.

Richardson, Mark P, Bryan A Strange, and Raymond J Dolan. 2004. "Encoding of Emotional Memories Depends on Amygdala and Hippocampus and Their Interactions." *Nature Neuroscience* 7 (3): 278–85. https://doi.org/10.1038/nn1190.

Rim, SoYon, James S. Uleman, and Yaacov Trope. 2009. "Spontaneous Trait Inference and Construal Level Theory: Psychological Distance Increases Nonconscious Trait Thinking." *Journal of Experimental Social Psychology* 45 (5): 1088–97. https://doi.org/10.1016/j.jesp.2009.06.015.

Ritchey, Maureen, Shao-Fang Wang, Andrew P. Yonelinas, and Charan Ranganath. 2019. "Dissociable Medial Temporal Pathways for Encoding Emotional Item and Context

Information." *Neuropsychologia* 124 (February): 66–78. https://doi.org/10.1016/j.neuropsychologia.2018.12.015.

Rose, J D, R Arlinghaus, S J Cooke, B K Diggles, W Sawynok, E D Stevens, and C D L Wynne. 2014. "Can Fish Really Feel Pain?" *Fish and Fisheries* 15 (1): 97–133. https://doi.org/10.1111/faf.12010.

Rosen, Gideon A., Alex Byrne, Joshua Cohen, and Seana Valentine Shiffrin, eds. 2015. *The Norton Introduction to Philosophy*. First Edition. New York: W.W. Norton & Company.

Rothkirch, Marcus, and Guido Hesselmann. 2017. "What We Talk about When We Talk about Unconscious Processing – A Plea for Best Practices." *Frontiers in Psychology* 8. https://doi.org/10.3389/fpsyg.2017.00835.

Rothkirch, Marcus, Morten Overgaard, and Guido Hesselmann. 2018. "Editorial: Transitions between Consciousness and Unconsciousness." *Frontiers in Psychology* 9. https://doi.org/10.3389/fpsyg.2018.00020.

Roy, Mathieu, Daphna Shohamy, and Tor D. Wager. 2012. "Ventromedial Prefrontal-Subcortical Systems and the Generation of Affective Meaning." *Trends in Cognitive Sciences* 16 (3): 147–56. https://doi.org/10.1016/j.tics.2012.01.005.

Rubin, Rachael D., Patrick D. Watson, Melissa C. Duff, and Neal J. Cohen. 2014. "The Role of the Hippocampus in Flexible Cognition and Social Behavior." *Frontiers in Human Neuroscience* 8 (September). https://doi.org/10.3389/fnhum.2014.00742.

Runyan, Jason D., Anthony N. Moore, and Pramod K. Dash. 2019. "Coordinating What We've Learned about Memory Consolidation: Revisiting a Unified Theory." *Neuroscience & Biobehavioral Reviews* 100 (May): 77–84. https://doi.org/10.1016/j.neubiorev.2019.02.010.

Rupert, Robert D. 2009. *Cognitive Systems and the Extended Mind*. Philosophy of Mind. Oxford ; New York: Oxford University Press.

———. 2013. "'Memory, Natural Kinds, and Cognitive Extension; or, Martians Don't Remember, and Cognitive Science Is Not about Cognition.'" *Review of Philosophy and Psychology* 4 (1): 25–47. https://doi.org/10.1007/s13164-012-0129-9.

———. 2018. "Representation and Mental Representation." *Philosophical Explorations* 21 (2): 204–25. https://doi.org/10.1080/13869795.2018.1477979.

Rupert, Robert D., and University of Arkansas Press. 2011. "Embodiment, Consciousness, and the Massively Representational Mind:" *Philosophical Topics* 39 (1): 99–120. https://doi.org/10.5840/philtopics201139116.

Ryan, Jennifer D., Maria C. D'Angelo, Arber Kacollja, Sandra Gardner, and R. Shayna Rosenbaum. 2020. "Gradual Learning and Inflexible Strategy Use in Amnesia: Evidence from Case H.C." *Neuropsychologia* 137 (February): 107280. https://doi.org/10.1016/j.neuropsychologia.2019.107280.

Rydell, Robert J., and Allen R. McConnell. 2006. "Understanding Implicit and Explicit Attitude Change: A Systems of Reasoning Analysis." *Journal of Personality and Social Psychology* 91 (6): 995–1008. https://doi.org/10.1037/0022-3514.91.6.995.

Rydell, Robert J., Allen R. McConnell, Laura M. Strain, Heather M. Claypool, and Kurt Hugenberg. 2007. "Implicit and Explicit Attitudes Respond Differently to Increasing Amounts of Counterattitudinal Information." *European Journal of Social Psychology* 37 (5): 867–78. https://doi.org/10.1002/ejsp.393.

Samuels, Richard, Stephen Stich, and Michael Bishop. 2002. "Ending the Rationality Wars How to Make Disputes about Human Rationality Disappear." In *Common Sense, Reasoning, and Rationality*, edited by Renee Elio, 236–68. Oxford University Press. https://doi.org/10.1093/0195147669.003.0011.

Satpute, Ajay B., and Matthew D. Lieberman. 2006. "Integrating Automatic and Controlled Processes into Neurocognitive Models of Social Cognition." *Brain Research*, Multiple Perspectives on the Psychological and Neural Bases of Understanding Other People's Behavior, 1079 (1): 86–97. https://doi.org/10.1016/j.brainres.2006.01.005.

Saul, Jennifer. 2018. "(How) Should We Tell Implicit Bias Stories?" *Disputatio* 10 (50): 217–44. https://doi.org/10.2478/disp-2018-0014.

Savage, Hannah S., Christopher G. Davey, Miquel A. Fullana, and Ben J. Harrison. 2020. "Clarifying the Neural Substrates of Threat and Safety Reversal Learning in Humans." *NeuroImage* 207 (February): 116427. https://doi.org/10.1016/j.neuroimage.2019.116427.

Savic, Branislav, Dario Cazzoli, René Müri, and Beat Meier. 2017. "No Effects of Transcranial DLPFC Stimulation on Implicit Task Sequence Learning and Consolidation." *Scientific Reports* 7 (1): 9649. https://doi.org/10.1038/s41598-017-10128-0.

Schacter, Daniel L. 1992. "Priming and Multiple Memory Systems: Perceptual Mechanisms of Implicit Memory." *Journal of Cognitive Neuroscience* 4 (3): 244–56. https://doi.org/10.1162/jocn.1992.4.3.244.

Schacter, Daniel L. n.d. "Implicit Memory: History and Current Status," 18.

Schacter, Daniel L., and Endel Tulving, eds. 1994a. *Memory Systems 1994*. Cambridge, Mass: MIT Press.

———. 1994b. "What Are the Memory Systems of 1994?" In *Memory Systems 1994*, edited by Daniel L Schacter and Endel Tulving, 2–38. Cambridge, Mass.: MIT Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1744.

Schapira, Marilyn M., Ann B. Nattinger, and Colleen A. McHorney. 2001. "Frequency or Probability? A Qualitative Study of Risk Communication Formats Used in Health Care." *Medical Decision Making* 21 (6): 459–67. https://doi.org/10.1177/0272989X0102100604.

Schiller, Daniela, and Mauricio R. Delgado. 2010. "Overlapping Neural Systems Mediating Extinction, Reversal and Regulation of Fear." *Trends in Cognitive Sciences* 14 (6): 268–76. https://doi.org/10.1016/j.tics.2010.04.002.

Schmidt, James R., Jan De Houwer, and Klaus Rothermund. 2016. "The Parallel Episodic Processing (PEP) Model 2.0: A Single Computational Model of Stimulus-Response Binding, Contingency Learning, Power Curves, and Mixing Costs." *Cognitive Psychology* 91 (December): 82–108. https://doi.org/10.1016/j.cogpsych.2016.10.004.

Schneider, Walter, and Richard M Shiffrin. n.d. "Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention," 66.

Schuck, Nicolas W., Ming Bo Cai, Robert C. Wilson, and Yael Niv. 2016. "Human Orbitofrontal Cortex Represents a Cognitive Map of State Space." *Neuron* 91 (6): 1402–12. https://doi.org/10.1016/j.neuron.2016.08.019.

Schwabe, Lars, Karim Nader, and Jens C. Pruessner. 2014. "Reconsolidation of Human Memory: Brain Mechanisms and Clinical Relevance." *Biological Psychiatry*, Neurobiological Moderators of Stress Response, 76 (4): 274–80. https://doi.org/10.1016/j.biopsych.2014.03.008.

Schwager, Susanne, and Klaus Rothermund. 2013. "Counter-Regulation Triggered by Emotions: Positive/Negative Affective States Elicit Opposite Valence Biases in Affective Processing." *Cognition and Emotion* 27 (5): 839–55. https://doi.org/10.1080/02699931.2012.750599.

Schwarz, Norbert. 2007. "Attitude Construction: Evaluation in Context." *Social Cognition* 25 (5): 638–56. https://doi.org/10.1521/soco.2007.25.5.638.

Scott, Ryan B., Jason Samaha, Ron Chrisley, and Zoltan Dienes. 2018. “Prevailing Theories of Consciousness Are Challenged by Novel Cross-Modal Associations Acquired between Subliminal Stimuli.” *Cognition* 175 (June): 169–85. https://doi.org/10.1016/j.cognition.2018.02.008.

Sechman, Michael. n.d. “Instructional Appointment Offer,” 8.

———. n.d. “Why Our Children Don’t Think There Are Moral Facts - The New York Times.” *The New York Times*, 5.

Sechman, Michael Scott. n.d. “UNIVERSITY OF COLORADO 1800 GRANT STREET SUITE 400 DENVER CO 80203-1187,” 1.

Seib-Pfeifer, Laura-Effi, and Henning Gibbons. 2019. “Independent ERP Predictors of Affective Priming Underline the Importance of Depth of Prime and Target Processing and Implicit Affect Misattribution.” *Brain and Cognition* 136 (November): 103595. https://doi.org/10.1016/j.bandc.2019.103595.

Senholzi, Keith B., Brendan E. Depue, Joshua Correll, Marie T. Banich, and Tiffany A. Ito. 2015. “Brain Activation Underlying Threat Detection to Targets of Different Races.” *Social Neuroscience* 10 (6): 651–62. https://doi.org/10.1080/17470919.2015.1091380.

Seo, D.-o., M. A. Carillo, S. Chih-Hsiung Lim, K. F. Tanaka, and M. R. Drew. 2015. “Adult Hippocampal Neurogenesis Modulates Fear Learning through Associative and Nonassociative Mechanisms.” *Journal of Neuroscience* 35 (32): 11330–45. https://doi.org/10.1523/JNEUROSCI.0483-15.2015.

Shanks, David R., and Mark F. St. John. 1994. “Characteristics of Dissociable Human Learning Systems.” *Behavioral and Brain Sciences* 17 (3): 367–95. https://doi.org/10.1017/S0140525X00035032.

Sherman, Jeffrey. 2009. "Controlled Influences on Implicit Measures: Confronting the Myth of Process-Purity and Taming the Cognitive Monster." *Attitudes: Insights from the New Wave of Implicit Measures*, January, 391–426.

Sherman, Jeffrey W. 2006. "AUTHORS' RESPONSES: Clearing Up Some Misconceptions About the Quad Model." *Psychological Inquiry* 17 (3): 269–76. https://doi.org/10.1207/s15327965pli1703_7.

———. 2008. "Controlled Influences on Implicit Measures: Confronting the Myth of Process-Purity and Taming the Cognitive Monster." In *Attitudes: Insights from the New Implicit Measures*, 391–426. New York, NY, US: Psychology Press.

Sherman, Jeffrey W., Bertram Gawronski, and Yaacov Trope, eds. 2014. *Dual-Process Theories of the Social Mind*. New York: The Guilford Press.

Sherry, David F., and Daniel L. Schacter. 1987. "The Evolution of Multiple Memory Systems." *Psychological Review* 94 (4): 439–54. https://doi.org/10.1037/0033-295X.94.4.439.

Shimamura, Arthur P. 1986. "Priming Effects in Amnesia: Evidence for a Dissociable Memory Function." *The Quarterly Journal of Experimental Psychology Section A* 38 (4): 619–44. https://doi.org/10.1080/14640748608401617.

Siegel, Shepard, and Lorraine G. Allan. 1996. "The Widespread Influence of the Rescorla-Wagner Model." *Psychonomic Bulletin & Review* 3 (3): 314–21. https://doi.org/10.3758/BF03210755.

Sigurdsson, Torfi, and Sevil Duvarci. 2016. "Hippocampal-Prefrontal Interactions in Cognition, Behavior and Psychiatric Disease." *Frontiers in Systems Neuroscience* 9 (January). https://doi.org/10.3389/fnsys.2015.00190.

Silva, Francisco J. 2018. "The Puzzling Persistence of 'Neutral' Conditioned Stimuli." *Behavioural Processes* 157 (December): 80–90. https://doi.org/10.1016/j.beproc.2018.07.004.

Smedley, Joseph W, and James A Bayton. n.d. “Evaluative Race-Class Stereotypes by Race and Perceived Class of Subjects,” 6.

Smith, Eliot R. 1996. “What Do Connectionism and Social Psychology Offer Each Other?” *Journal of Personality and Social Psychology* 70 (5): 893–912. https://doi.org/10.1037/0022-3514.70.5.893.

———. 2009. “Distributed Connectionist Models in Social Psychology.” *Social and Personality Psychology Compass* 3 (1): 64–76. https://doi.org/10.1111/j.1751-9004.2008.00160.x.

Smith, Eliot R., and Jamie DeCoster. 2000. “Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems.” *Personality and Social Psychology Review* 4 (2): 108–31. https://doi.org/10.1207/S15327957PSPR0402_01.

Smith, J. David, and Barbara A. Church. 2018. “Dissociable Learning Processes in Comparative Psychology.” *Psychonomic Bulletin & Review* 25 (5): 1565–84. https://doi.org/10.3758/s13423-017-1353-1.

Smolensky, Paul. 1988. “On the Proper Treatment of Connectionism.” *BEHAVIORAL AND BRAIN SCIENCES*, 74.

Spitzer, Manfred, Urs Fischbacher, Bärbel Herrnberger, Georg Grön, and Ernst Fehr. 2007. “The Neural Signature of Social Norm Compliance.” *Neuron* 56 (1): 185–96. https://doi.org/10.1016/j.neuron.2007.09.011.

Spreng, R. Nathan. 2013. “Examining the Role of Memory in Social Cognition.” *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00437.

Spruyt, Adriaan, Jan De Houwer, Dirk Hermans, and Paul Eelen. 2007. “Affective Priming of Nonaffective Semantic Categorization Responses.” *Experimental Psychology* 54 (1): 44–53. https://doi.org/10.1027/1618-3169.54.1.44.

Squire, L. R. 1992. "Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory." *Journal of Cognitive Neuroscience* 4 (3): 232–43. https://doi.org/10.1162/jocn.1992.4.3.232.

Squire, L. R., and S. Zola-Morgan. 1991. "The Medial Temporal Lobe Memory System." *Science (New York, N.Y.)* 253 (5026): 1380–86. https://doi.org/10.1126/science.1896849.

Squire, Larry R. 2004. "Memory Systems of the Brain: A Brief History and Current Perspective." *Neurobiology of Learning and Memory* 82 (3): 171–77. https://doi.org/10.1016/j.nlm.2004.06.005.

Squire, Larry R, and Pablo Alvarez. 1995. "Retrograde Amnesia and Memory Consolidation: A Neurobiological Perspective." *Current Opinion in Neurobiology* 5 (2): 169–77. https://doi.org/10.1016/0959-4388(95)80023-9.

Squire, Larry R, and Stuart M Zola. n.d. "Episodic Memory, Semantic Memory, and Amnesia," 7.

Staats, Arthur W., and Carolyn K. Staats. 1958. "Attitudes Established by Classical Conditioning." *The Journal of Abnormal and Social Psychology* 57 (1): 37–40. https://doi.org/10.1037/h0042782.

Staniloiu, Angelica, Andreas Kordon, and Hans J. Markowitsch. 2020. "Quo Vadis 'Episodic Memory'? – Past, Present, and Perspective." *Neuropsychologia* 141 (April): 107362. https://doi.org/10.1016/j.neuropsychologia.2020.107362.

Stanley, D. A., P. Sokol-Hessner, M. R. Banaji, and E. A. Phelps. 2011. "Implicit Race Attitudes Predict Trustworthiness Judgments and Economic Trust Decisions." *Proceedings of the National Academy of Sciences* 108 (19): 7710–15. https://doi.org/10.1073/pnas.1014345108.

Stanley, Damian, Elizabeth Phelps, and Mahzarin Banaji. 2008. "The Neural Basis of Implicit Attitudes." *Current Directions in Psychological Science* 17 (2): 164–70. https://doi.org/10.1111/j.1467-8721.2008.00568.x.

Stein, Timo, Daniel Kaiser, and Guido Hesselmann. 2016. “Can Working Memory Be Non-Conscious?” *Neuroscience of Consciousness* 2016 (1): niv011. https://doi.org/10.1093/nc/niv011.

Sterelny, Kim, Richard Joyce, Brett Calcott, and Ben Fraser, eds. 2013. *Cooperation and Its Evolution*. Life and Mind: Philosophical Issues in Biology and Psychology. Cambridge, Mass. ; London, Eng: MIT Press.

Stevens, W.D., G.S. Wig, and D.L. Schacter. 2008. “Implicit Memory and Priming.” In *Learning and Memory: A Comprehensive Reference*, 623–44. Elsevier. https://doi.org/10.1016/B978-012370509-9.00150-9.

Stich, Stephen. 2012. *Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199733477.001.0001.

Stich, Stephen P. n.d. “Beliefs and Subdoxastic States,” 20.

———. n.d. “The Flight to Reference, or How Not to Make Progress in the Philosophy of Science*,” 17.

Stich, Stephen, and Kevin P. Tobia. 2016. “Experimental Philosophy and the Philosophical Tradition.” In *A Companion to Experimental Philosophy*, edited by Justin Sytsma and Wesley Buckwalter, 3–21. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch1.

Stoianov, Ivilin Peev, Cyriel M. A. Pennartz, Carien S. Lansink, and Giovani Pezzulo. 2018. “Model-Based Spatial Navigation in the Hippocampus-Ventral Striatum Circuit: A Computational Analysis.” Edited by Jean Daunizeau. *PLOS Computational Biology* 14 (9): e1006316. https://doi.org/10.1371/journal.pcbi.1006316.

Storbeck, Justin, and Gerald L. Clore. 2008. “The Affective Regulation of Cognitive Priming.” *Emotion* 8 (2): 208–15. https://doi.org/10.1037/1528-3542.8.2.208.

Strack, Fritz, and Roland Deutsch. n.d. "Reflective and Impulsive Determinants of Social Behavior," 28.

"Studying Implicit Social Cognition with Noninvasive Brain Stimulation - ScienceDirect." n.d. Accessed February 11, 2020. https://www.sciencedirect.com/science/article/pii/S1364661318301761.

Stussi, Yoann, Aude Ferrero, Gilles Pourtois, and David Sander. 2019. "Achievement Motivation Modulates Pavlovian Aversive Conditioning to Goal-Relevant Stimuli." *Npj Science of Learning* 4 (1): 4. https://doi.org/10.1038/s41539-019-0043-3.

Stussi, Yoann, Gilles Pourtois, Andreas Olsson, and David Sander. 2020. "Learning Biases to Angry and Happy Faces during Pavlovian Aversive Conditioning." *Emotion*, March. https://doi.org/10.1037/emo0000733.

Stussi, Yoann, Gilles Pourtois, and David Sander. 2018. "Enhanced Pavlovian Aversive Conditioning to Positive Emotional Stimuli." *Journal of Experimental Psychology: General* 147 (6): 905–23. https://doi.org/10.1037/xge0000424.

Suresh, Abhijit, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. "Automating Analysis and Feedback to Improve Mathematics Teachers' Classroom Discourse." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July): 9721–28. https://doi.org/10.1609/aaai.v33i01.33019721.

Suslow, Thomas, Harald Kugel, Patricia Ohrmann, Anja Stuhrmann, Dominik Grotegerd, Ronny Redlich, Jochen Bauer, and Udo Dannlowski. 2013. "Neural Correlates of Affective Priming Effects Based on Masked Facial Emotion: An FMRI Study." *Psychiatry Research: Neuroimaging* 211 (3): 239–45. https://doi.org/10.1016/j.pscychresns.2012.09.008.

Swannell, Ellen R., Christopher A. Brown, Anthony K.P. Jones, and Richard J. Brown. 2016. "Some Words Hurt More Than Others: Semantic Activation of Pain Concepts in Memory and

Subsequent Experiences of Pain." *The Journal of Pain* 17 (3): 336–49. https://doi.org/10.1016/j.jpain.2015.11.004.

Sweldens, Steven. 2018. "Putting Operating Principles (What) Before Conditions (How and When) to Improve Theorizing: S - S versus S - R Learning in Evaluative Conditioning." Preprint. Open Science Framework. https://doi.org/10.31219/osf.io/bu4v5.

Sweldens, Steven, Olivier Corneille, and Vincent Yzerbyt. 2014. "The Role of Awareness in Attitude Formation Through Evaluative Conditioning." *Personality and Social Psychology Review* 18 (2): 187–209. https://doi.org/10.1177/1088868314527832.

Talmi, Deborah, Martina Slapkova, and Matthias J. Wieser. 2018. "Testing the Possibility of Model-Based Pavlovian Control of Attention to Threat." *Journal of Cognitive Neuroscience* 31 (1): 36–48. https://doi.org/10.1162/jocn_a_01329.

Taubenfeld, Assaf, Michael C. Anderson, and Daniel A. Levy. 2019. "The Impact of Retrieval Suppression on Conceptual Implicit Memory." *Memory* 27 (5): 686–97. https://doi.org/10.1080/09658211.2018.1554079.

The SAGE Handbook of Prejudice, Stereotyping and Discrimination. 2010. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781446200919.

Theiner, Georg. 2016. "Muhammad Ali Khalidi: Natural Categories and Human Kinds. Classification in the Natural and Social Sciences: Cambridge University Press, Cambridge, 2013, 288 Pp, $90.00, ISBN: 978-1-107-01274-5." *Journal for General Philosophy of Science* 47 (1): 247–55. https://doi.org/10.1007/s10838-015-9309-5.

Tobin, Emma. 2013. "Are Natural Kinds and Natural Properties Distinct?" In *Metaphysics and Science*, edited by Stephen Mumford and Matthew Tugby, 164–82. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199674527.003.0008.

Tottenham, Nim, Mor Shapiro, Eva H. Telzer, and Kathryn L. Humphreys. 2012. "Amygdala Response to Mother." *Developmental Science* 15 (3): 307–19. https://doi.org/10.1111/j.1467-7687.2011.01128.x.

"Trace Conditioning and the Hippocampus: The Importance of Contiguity | Journal of Neuroscience." n.d. Accessed January 28, 2020. https://www.jneurosci.org/content/26/34/8702.

Trübutschek, Darinka, Sébastien Marti, and Stanislas Dehaene. 2019. "Temporal-Order Information Can Be Maintained in Non-Conscious Working Memory." *Scientific Reports* 9 (1): 6484. https://doi.org/10.1038/s41598-019-42942-z.

Tulving, E, and D. Schacter. 1990. "Priming and Human Memory Systems." *Science* 247 (4940): 301–6. https://doi.org/10.1126/science.2296719.

Tulving, Endel. 1983. *Elements of Episodic Memory*. Oxford Psychology Series, no. 2. Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press.

———. 1995. "Organization of Memory: Quo Vadis?" In *The Cognitive Neurosciences*, 839–47. The Cognitive Neurosciences. Cambridge, MA, US: The MIT Press.

Tylén, Kristian, Ethan Weed, Mikkel Wallentin, Andreas Roepstorff, and Chris D. Frith. 2010. "Language as a Tool for Interacting Minds." *Mind & Language* 25 (1): 3–29. https://doi.org/10.1111/j.1468-0017.2009.01379.x.

Tyler, Lorraine K, and Helen E Moss. 1998. "Going, Going, Gone . . . ? Implicit and Explicit Tests of Conceptual Knowledge in a Longitudinal Study of Semantic Dementia." *Neuropsychologia* 36 (12): 1313–23. https://doi.org/10.1016/S0028-3932(98)00029-3.

Unkelbach, Christian, and Sabine Förderer. 2018. "A Model of Attribute Conditioning." *Social Psychological Bulletin* 13 (3): e28568. https://doi.org/10.5964/spb.v13i3.28568.

Vallila-Rohter, Sofia, and Swathi Kiran. 2013. "Non-Linguistic Learning and Aphasia: Evidence from a Paired Associate and Feedback-Based Task." *Neuropsychologia* 51 (1): 79–90. https://doi.org/10.1016/j.neuropsychologia.2012.10.024.

Van Dessel, Pieter, and Jan De Houwer. 2019. "Hypnotic Suggestions Can Induce Rapid Change in Implicit Attitudes." *Psychological Science* 30 (9): 1362–70. https://doi.org/10.1177/0956797619865183.

Van Dessel, Pieter, Jan De Houwer, Anne Gast, Colin Tucker Smith, and Maarten De Schryver. 2016. "Instructing Implicit Processes: When Instructions to Approach or Avoid Influence Implicit but Not Explicit Evaluation." *Journal of Experimental Social Psychology* 63 (March): 1–9. https://doi.org/10.1016/j.jesp.2015.11.002.

Van Dessel, Pieter, Jan De Houwer, Anne Gast, and Colin Tucker Smith. 2015. "Instruction-Based Approach-Avoidance Effects: Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli." *Experimental Psychology* 62 (3): 161–69. https://doi.org/10.1027/1618-3169/a000282.

Van Dessel, Pieter, Bertram Gawronski, and Jan De Houwer. 2019. "Does Explaining Social Behavior Require Multiple Memory Systems?" *Trends in Cognitive Sciences* 23 (5): 368–69. https://doi.org/10.1016/j.tics.2019.02.001.

Van Dessel, Pieter, Gaëtan Mertens, Colin Tucker Smith, and Jan De Houwer. 2017. "The Mere Exposure Instruction Effect: Mere Exposure Instructions Influence Liking." *Experimental Psychology* 64 (5): 299–314. https://doi.org/10.1027/1618-3169/a000376.

———. 2019. "Mere Exposure Effects on Implicit Stimulus Evaluation: The Moderating Role of Evaluation Task, Number of Stimulus Presentations, and Memory for Presentation Frequency." *Personality and Social Psychology Bulletin* 45 (3): 447–60. https://doi.org/10.1177/0146167218789065.

Van Dessel, Pieter, Yang Ye, and Jan De Houwer. 2019. “Changing Deep-Rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi.” *Social Psychological and Personality Science* 10 (2): 266–73. https://doi.org/10.1177/1948550617752064.

Van Hoeck, Nicole. 2015. “Cognitive Neuroscience of Human Counterfactual Reasoning.” *Frontiers in Human Neuroscience* 9. https://doi.org/10.3389/fnhum.2015.00420.

Vaughn, Lewis. n.d. “Effective Reasoning About Ordinary and Extraordinary Claims,” 521.

Wade, Mark, Heather Prime, Jennifer M. Jenkins, Keith O. Yeates, Tricia Williams, and Kang Lee. 2018. “On the Relation between Theory of Mind and Executive Functioning: A Developmental Cognitive Neuroscience Perspective.” *Psychonomic Bulletin & Review* 25 (6): 2119–40. https://doi.org/10.3758/s13423-018-1459-0.

Walther, Eva, Bertram Gawronski, Hartmut Blank, and Tina Langer. 2009. “Changing Likes and Dislikes through the Back Door: The US-Revaluation Effect.” *Cognition & Emotion* 23 (5): 889–917. https://doi.org/10.1080/02699930802212423.

Wang, Fang, Geoffrey Schoenbaum, and Thorsten Kahnt. 2020. “Interactions between Human Orbitofrontal Cortex and Hippocampus Support Model-Based Inference.” Edited by Matthew F. S. Rushworth. *PLOS Biology* 18 (1): e3000578. https://doi.org/10.1371/journal.pbio.3000578.

Wang, Wei-Chun, Michele M. Lazzara, Charan Ranganath, Robert T. Knight, and Andrew P. Yonelinas. 2010. “The Medial Temporal Lobe Supports Conceptual Implicit Memory.” *Neuron* 68 (5): 835–42. https://doi.org/10.1016/j.neuron.2010.11.009.

Wang, Yin, Thomas W Schubert, and Susanne Quadflieg. 2019. “Behavioral and Neural Evidence for an Evaluative Bias against Other People’s Mundane Interracial Encounters.” *Social Cognitive and Affective Neuroscience* 14 (12): 1329–39. https://doi.org/10.1093/scan/nsaa005.

Wang, Yinan, and Qin Zhang. 2016. “Affective Priming by Simple Geometric Shapes: Evidence from Event-Related Brain Potentials.” *Frontiers in Psychology* 7 (June). https://doi.org/10.3389/fpsyg.2016.00917.

Wang, Yingying, Andrea Luppi, Jonathan Fawcett, and Michael C. Anderson. 2019. “Reconsidering Unconscious Persistence: Suppressing Unwanted Memories Reduces Their Indirect Expression in Later Thoughts.” *Cognition* 187 (June): 78–94. https://doi.org/10.1016/j.cognition.2019.02.016.

Webber, Emily S., David E. Mankin, and Howard C. Cromwell. 2016. “Striatal Activity and Reward Relativity: Neural Signals Encoding Dynamic Outcome Valuation.” *Eneuro* 3 (5): ENEURO.0022-16.2016. https://doi.org/10.1523/ENEURO.0022-16.2016.

Weidemann, Christoph T., James E. Kragel, Bradley C. Lega, Gregory A. Worrell, Michael R. Sperling, Ashwini D. Sharan, Barbara C. Jobst, *et al.* 2019. “Neural Activity Reveals Interactions between Episodic and Semantic Memory Systems during Retrieval.” *Journal of Experimental Psychology: General* 148 (1): 1–12. https://doi.org/10.1037/xge0000480.

Weissengruber, Sebastian, Sang Wan Lee, John P. O’Doherty, and Christian C. Ruff. 2019. “Neurostimulation Reveals Context-Dependent Arbitration Between Model-Based and Model-Free Reinforcement Learning.” *Cerebral Cortex* 29 (11): 4850–62. https://doi.org/10.1093/cercor/bhz019.

Wheeler, S. Christian, and Kenneth G. DeMarree. 2009. “Multiple Mechanisms of Prime-to-Behavior Effects.” *Social and Personality Psychology Compass* 3 (4): 566–81. https://doi.org/10.1111/j.1751-9004.2009.00187.x.

“Where Is the Trace in Trace Conditioning? - ScienceDirect.” n.d. Accessed February 11, 2020. https://www.sciencedirect.com/science/article/abs/pii/S0166223607003013.

Whisman, Mark A., and Gary H. McClelland. 2005. "Designing, Testing, and Interpreting Interactions and Moderator Effects in Family Research." *Journal of Family Psychology* 19 (1): 111–20. https://doi.org/10.1037/0893-3200.19.1.111.

White, Norman M., Mark G. Packard, and Robert J. McDonald. 2013. "Dissociation of Memory Systems: The Story Unfolds." *Behavioral Neuroscience* 127 (6): 813–34. https://doi.org/10.1037/a0034859.

White, Stuart F., Joel L. Voss, Jessica J. Chiang, Lei Wang, Katie A. McLaughlin, and Gregory E. Miller. 2019. "Exposure to Violence and Low Family Income Are Associated with Heightened Amygdala Responsiveness to Threat among Adolescents." *Developmental Cognitive Neuroscience* 40 (December): 100709. https://doi.org/10.1016/j.dcn.2019.100709.

Whittlesea, Bruce W. A., and John R. Price. 2001. "Implicit /Explicit Memory versus Analytic/Nonanalytic Processing: Rethinking the Mere Exposure Effect." *Memory & Cognition* 29 (2): 234–46. https://doi.org/10.3758/BF03194917.

"Why Trace and Delay Conditioning Are Sometimes (but Not Always) Hippocampal Dependent: A Computational Model. - PsycNET." n.d. Accessed January 28, 2020. https://psycnet.apa.org/record/2012-33597-001.

Williams, M. A. 2004. "Amygdala Responses to Fearful and Happy Facial Expressions under Conditions of Binocular Suppression." *Journal of Neuroscience* 24 (12): 2898–2904. https://doi.org/10.1523/JNEUROSCI.4977-03.2004.

Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, Mass: Belknap Press of Harvard University Press.

Wolsiefer, Katie, Jacob Westfall, and Charles M. Judd. 2017. "Modeling Stimulus Variation in Three Common Implicit Attitude Tasks." *Behavior Research Methods* 49 (4): 1193–1209. https://doi.org/10.3758/s13428-016-0779-0.

Wuethrich, Sergej, Deborah E. Hannula, Fred W. Mast, and Katharina Henke. 2018. "Subliminal Encoding and Flexible Retrieval of Objects in Scenes." *Hippocampus* 28 (9): 633–43. https://doi.org/10.1002/hipo.22957.

Yan, Hao, Randi C. Martin, and L. Robert Slevc. 2018. "Lexical Overlap Increases Syntactic Priming in Aphasia Independently of Short-Term Memory Abilities: Evidence against the Explicit Memory Account of the Lexical Boost." *Journal of Neurolinguistics*, Short-term and working memory deficits in aphasia: Current issues in theory, evidence, and treatment, 48 (November): 76–89. https://doi.org/10.1016/j.jneuroling.2017.12.005.

Yang, Guangyu Robert, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. 2019. "Task Representations in Neural Networks Trained to Perform Many Cognitive Tasks." *Nature Neuroscience* 22 (2): 297–306. https://doi.org/10.1038/s41593-018-0310-2.

Yang, J., Z. Cao, X. Xu, and G. Chen. 2012. "The Amygdala Is Involved in Affective Priming Effect for Fearful Faces." *Brain and Cognition* 80 (1): 15–22. https://doi.org/10.1016/j.bandc.2012.04.005.

Yau, Joanna Oi-Yue, and Gavan P McNally. 2019. "Rules for Aversive Learning and Decision-Making." *Current Opinion in Behavioral Sciences* 26 (April): 1–8. https://doi.org/10.1016/j.cobeha.2018.08.006.

Yonelinas, Andrew P., and Larry L. Jacoby. 2012. "The Process-Dissociation Approach Two Decades Later: Convergence, Boundary Conditions, and New Directions." *Memory & Cognition* 40 (5): 663–80. https://doi.org/10.3758/s13421-012-0205-5.

Yonelinas, Andrew P., Charan Ranganath, Arne D. Ekstrom, and Brian J. Wiltgen. 2019. "A Contextual Binding Theory of Episodic Memory: Systems Consolidation Reconsidered." *Nature Reviews Neuroscience* 20 (6): 364–75. https://doi.org/10.1038/s41583-019-0150-4.

Young, Alex. 2018. "Will the Real Specification Please Stand Up? A Comment on Andrew Bird and Stephen Karolyi" 15 (1): 14.

Zanon, Riccardo, Jan De Houwer, Anne Gast, and Colin Tucker Smith. 2014. "When Does Relational Information Influence Evaluative Conditioning?" *Quarterly Journal of Experimental Psychology* 67 (11): 2105–22. https://doi.org/10.1080/17470218.2014.907324.

Zeithamova, Dagmar, Michael L. Mack, Kurt Braunlich, Tyler Davis, Carol A. Seger, Marlieke T.R. van Kesteren, and Andreas Wutz. 2019. "Brain Mechanisms of Concept Learning." *The Journal of Neuroscience* 39 (42): 8259–66. https://doi.org/10.1523/JNEUROSCI.1166-19.2019.

Zeng, An, Zhesi Shen, Jianlin Zhou, Jinshan Wu, Ying Fan, Yougui Wang, and H. Eugene Stanley. 2017. "The Science of Science: From the Perspective of Complex Systems." *Physics Reports* 714–715 (November): 1–73. https://doi.org/10.1016/j.physrep.2017.10.001.

Zhang, Qin, Adam Lawson, Chunyan Guo, and Yang Jiang. 2006. "Electrophysiological Correlates of Visual Affective Priming." *Brain Research Bulletin* 71 (1–3): 316–23. https://doi.org/10.1016/j.brainresbull.2006.09.023.

Zhang, Qin, Xiaohua Li, Brian T. Gold, and Yang Jiang. 2010. "Neural Correlates of Cross-Domain Affective Priming." *Brain Research* 1329 (May): 142–51. https://doi.org/10.1016/j.brainres.2010.03.021.

Zhang, Suyi, Hiroaki Mano, Gowrishankar Ganesh, Trevor Robbins, and Ben Seymour. 2016. "Dissociable Learning Processes Underlie Human Pain Conditioning." *Current Biology* 26 (1): 52–58. https://doi.org/10.1016/j.cub.2015.10.066.

Zhang, Yicheng, Shengdong Chen, Zhongyan Deng, Jiemin Yang, and Jiajin Yuan. 2020. "Benefits of Implicit Regulation of Instructed Fear: Evidence From Neuroimaging and Functional Connectivity." *Frontiers in Neuroscience* 14 (March): 201. https://doi.org/10.3389/fnins.2020.00201.

Ziaei, Maryam, Mansoureh Togha, Elham Rahimian, and Jonas Persson. 2018. "The Causal Role of Right Frontopolar Cortex in Moral Judgment, Negative Emotion Induction, and Executive Control." *Basic and Clinical Neuroscience Journal*, October. https://doi.org/10.32598/bcn.9.10.225.

Züst, Marc Alain, Patrizio Colella, Thomas Peter Reber, Patrik Vuilleumier, Martinus Hauf, Simon Ruch, and Katharina Henke. 2015. "Hippocampus Is Place of Interaction between Unconscious and Conscious Memories." Edited by Angela Sirigu. *PLOS ONE* 10 (3): e0122459. https://doi.org/10.1371/journal.pone.0122459.