BOSTON COLLEGE

Securing the Reliability of Episodic Memory

A Scholar of the College Thesis

Submitted to

The Morrissey College of Arts and Sciences

Department of Psychology

and

Morrissey College of Arts and Sciences Honors Program

by

Maria Khoudary

Department of Psychology

Department of Philosophy

May 2019

Abstract

An abundance of psychological and neuroscientific evidence of false and distorted memories suggests that episodic memory operates via active reconstruction of previously instantiated patterns of neural activity. The phenomenology of episodic memory retrieval, however, has systematically misled philosophers of mind and epistemologists into believing that memory functions to preserve past experience, and that it does so by retrieving a record from the storehouse and pressing "play." More recent research suggests that memory is properly situated as a subprocess of a larger system that functions to generate episodic hypothetical thoughts, which naturally raises questions about the veridicality of memory experience. The current thesis aims to identify the factors contributing to the reliability of constructive memory and use these to lay the groundwork for an epistemology of episodic memory, the neurocognitive (sub)system supporting re-experiencing of the personal past. I argue that metacognitive monitoring processes play a critical role in allowing agents to form justified beliefs about past experiences and report the results of an experiment designed to test how this ability is constrained by conditions during the initial experience. Ultimately, I suggest that a naturalistic epistemology of episodic memory requires taking a hybrid internal-external approach to justification and identifying episodic simulations as a basic and generative epistemic source.

Acknowledgements

This thesis would not be one fraction of what it has become without the patient guidance of Dr. Rose Cooper, Dr. Richard Atkins, and Dr. Maureen Ritchey. Thank you for nurturing this project throughout every phase of its creation: for giving me the freedom to pursue the kind of interdisciplinary work I have always wanted to do, for helping me turn dozens of half-baked ideas into substantive research questions, and for having faith in me when I struggled to do so myself. Thanks are also due to Dr. Cherie McGill, Dr. Scott Slotnick, Dr. Chris Constas, and Fr. Arthur Madigan. Without their initial and continued support, I would not have had enough courage in myself or my interests to undertake a project of this scope. I would also like to thank Dr. Daniel Dennett for showing me what can be gained by putting philosophy, psychology, and neuroscience in conversation with each other, and how to do that well. The heavy hand he has had in directing my intellectual evolution should be obvious in what follows.

To other full-time members of the Memory Modulation Lab: Helen Schmidt, Rosalie Samide, and Kyle Kurkela, I extend endless gratitude for ears lent and sanity checked. For sharing every moment of joy, fear, and frustration with me as I entered into uncharted intellectual territory, and for accompanying me on a parallel journey into the unknown, I thank the inimitable Emily Iannazzi. The friendship we have forged over the course of this year is something I will cherish for several years to come. And to Mary Nanna, the star that completes our constellation, I am deeply grateful for intellectual, emotional, and physical nourishment.

I am indebted to all of my friends whose unrelenting encouragement and affirmation buoyed me through the depths of imposter syndrome, and whose excellence inspires me to be and do better. Sarya Baladi, Amelia Culp, Ciara Bauwens, and Beckett Pulis's embarkment with me into the realm of senior thesis-making especially poised them to support me in this way. Rayan Habbab, Sean Kane, Harry Hoy, Louise Nessralla, Emma Arcos, and Aine McAlinden have been profound sources of domain-general support, and I owe much of who I am today to the pieces of their selves they have shared with me. And the entry of Emily Jennings into my orbit supplied me with precisely the motivation I needed to bring this project to life. Finally, I extend my deepest and fullest thanks to my parents, without whom none of this would be possible. Thank you for fostering my creativity and always making sure I had everything I needed to succeed. This thesis is only mine inasmuch as I am yours.

Table of Contents

Preface

> Out of the same storehouse, with these past impressions, I can construct now this, now that, image of things that I either have experienced or have believed on the basis of experience — and from these I can further construct future actions, events, and hopes; and I can meditate on all these things as if they were present.
>
> — Augustine, *Confessions*

Memory is central to our mental lives. It is the reason for the nostalgia you feel upon getting a whiff of your mother's perfume, the comfort you experience when listening to one of your favorite songs, and the fact that a *you* exists at all. The ability to retain, retrieve, and utilize information about past experience was critical for sustaining life on the plains of the Serengeti and in the deep sea, and, as such, has played an indispensable role in the development of human consciousness. Millions of years of evolutionary design have resulted in a nervous system that is incredibly adept at transducing, representing, and storing information from the environment in a manner that facilitates retrieval of that information at a later date. Despite this fact, even a brief bout of introspection should bring to mind a number of occasions on which your memory has performed less-than-optimally. How many times have you gone to the grocery store only to return without the item that motivated the trip? Even though the evolutionary significance of food could not be more obvious, these instances of forgetting are remarkably common.

Such "failures of omission" lend themselves nicely to our commonsense conception of memory as some kind of "filing cabinet," or "storehouse," that preserves information to be retrieved at a later date. Because this is our intuition about the nature of memory, instances of forgetting are especially salient marks of memory malfunction. However, an overwhelming

amount of empirical evidence demonstrates that memory can err in the opposite direction — i.e., storing and/or retrieving information which was not contained in an initial experience. This generation is not nearly as noticeable as forgetting, as our implicit belief in the preservative nature of memory leads us to accept its contents without much question. Clearly, these false memories challenge the storehouse model of human memory. Since Bartlett's (1932) initial demonstration of the phenomena, the psychological community has distinguished between *reproductive* and *reconstructive* memory. An abundance of evidence shows that the majority of memory processes are reconstructive in nature, yet much of the philosophical work on memory has operated on the preservative/reproductive model of human memory.

The current project aims to contribute to the small (but growing) literature that revisits philosophical questions of memory in light of up-to-date empirical knowledge. Specifically, I am interested in laying the groundwork for an epistemology of episodic memory. Because knowledge is a unique kind of belief, and because all beliefs can be expressed propositionally, nearly all epistemological work on memory has focused exclusively on the semantic memory system. However, a great deal of our beliefs are formed on the basis of past experience. Episodic memory is the cognitive process responsible for processing and transferring information about experience, and thus plays a central role in knowledge of the personal past. The fact that it operates via reconstruction rather than reproduction naturally raises questions about reliability, so the goal of this project is to identify some key factors contributing to its reliability and using these to begin developing an account of the role episodic memory plays in human knowledge.

I begin by addressing the metaphysics of episodic memory in Chapter 1. As scientific realism is a central commitment of this project, I review the methods experimental psychology uses to measure episodic memory and the insights those methods have borne. The focus is on

evidence of false or distorted memories and the conditions under which they arise, as these have been critical for understanding the constructive nature of human memory. In considering *why* memory should function via reconstruction rather than reproduction, a number of theorists have suggested that it plays a critical role in allowing us to imagine the future. Specifically, De Brigard (2014) suggests that remembering is a subprocess of a larger cognitive system that functions to generate episodic hypothetical thoughts. I conclude Chapter 1 by reviewing the development of this theory and showing how it offers the best explanation for misremembering.

The "episodic" dimension of the episodic hypothetical thought system stems from the fact that it generates simulations of possible events. If remembering is (1) reconstructive and (2) a subprocess of imagination, then two questions naturally arise: how does memory accurately reconstruct past experiences, and how is memory distinguished from imagination? Chapter 1 addresses some processes whereby memory generates accurate reconstructions, and Chapter 2 expands on these to give an account of how agents assess the accuracy of a reconstruction and discriminate between reconstructions and imagination. Because assessing the accuracy of a reconstruction requires first recognizing that a simulation is a reconstruction and not an imagination, the first half of Chapter 2 presents a revised version of Michaelian's (2016) process monitoring framework that explains how agents solve this problem on the episodic hypothetical thought theory. The second half uses Johnson et al.'s (1993) source monitoring framework to give a more thorough account for how agents assess the accuracy of reconstructions.

Chapter 3 details the methods and results of an experiment testing the effects of divided attention on item and source memory. I designed this experiment to better understand the encoding conditions that drive picture misattributions, a well-documented phenomenon where agents misremember seeing words as pictures. Johnson and collaborators (1980, 1981, 1993)

have argued that this effect represents a source monitoring error where participants mistake their memory for mental images generated in response to studying word stimuli as memory for having seen that item presented as an image. Next, I report data on the relationship between task performance and metacognitive judgments, metamemory beliefs and response biases, and self-reported methods which agents use to remember the medium through which a memory was formed. These data constitute the experimental philosophy portion of the project, which aims to understand how beliefs about cognition (1) are related to objective cognitive performance and (2) inform how agents use their cognitive capacities to solve problems.

Chapter 4 integrates all of these considerations into a preliminary epistemology of episodic memory. I argue that episodic memory is a basic and generative epistemic source, similar to perception. That is, we form beliefs about our personal past on the basis of episodic simulations, much like we form beliefs about our current environment on the basis of perception. The majority of the chapter is devoted to developing a theory of why agents are justified in forming beliefs about their pasts in this way. In no way is this intended to reflect a comprehensive or exhaustive account of episodic memorial justification. Rather, the goal is to lay down some basic claims about episodic memory knowledge that can serve as foundations for developing an epistemology of this cognitive capacity.

Chapter One: Metaphysics of Episodic Memory

1.1 Scientific realism about memory

In their respective pursuits of truth, both scientists and philosophers make use of experiments. Scientists do so by systematically controlling and manipulating energy, biology, or behavior to test hypotheses about how systems function. Because the goal is empirical demonstration, the results of any single experiment will be necessarily (1) limited in scope and (2) contingent on the conditions of the natural world. Philosophers, on the other hand, construct (thought) experiments to test the strengths and weaknesses of a philosophical thesis. Usually the goal is to demonstrate the insufficiency of a thesis in accounting for some logical possibilities, but sometimes the goal is simply to create a unique problem space within which different ideas and intuitions are explored.

In this vein, a popular style of argumentation in analytic philosophy is to pose a theory, construct a counterexample, and then refine the original theory. This happens both within and between pieces of writing. Because these counterexamples are constructed to probe the strength of a theory in accounting for any number of related phenomena, the only constraint on the structure of the thought experiment is logical consistency. A classic example is Putnam's (1973) "Twin Earth" thought experiment for semantic externalism ("meanings' just ain't in the head"):

> Suppose there is a nearby possible world, "Twin Earth." It is a molecule-for-molecule duplicate of this earth, up to and including the neural constitution of our identical populations. The only difference is that their analog of H_2O has a much more complicated chemical structure, abbreviated XYZ. Regardless, both us and our twins call this substance "water." The question is: when my twin and I say "water," do we mean the same thing?

A scientist might already be skeptical about the validity of such a thought experiment: how identical could two worlds be if such a fundamental substance is chemically different? But the setup is logically possible, and a number of philosophers frequently make use of similar "possible worlds" counterexamples.

Like Michaelian (2016), the goal of the current project is not to develop a theory of remembering that is immune to any and all possible counterexamples. Rather, I aim to defend a theory of remembering that is grounded in what is known about how human memory works in this world. This is the scientific realism commitment of the project, which logically entails a commitment to naturalism. Thus, evolution will be the conceptual bottleneck through which any theory must pass (Ginsburg and Jablonka, 2019), and any thought experiments have to speak to circumstances that are consistent with what is known about the natural world. I believe that these metaphysical constraints are necessary if we are to develop a serious account of memory knowledge as it is known to exist, and that this is a more pressing question than understanding all of the possible ways memory knowledge *could* exist.

1.2 Human memory: processes, components, and structures

Human memory is a complex capacity comprised of a number of unique component processes. Atkinson and Shiffrin's (1968) modal model has proven quite useful in conceptualizing the kinds of different information-processing undertaken by memory, evidenced by its persistence as an organizational scheme to this day. The model divides memory into three components: sensory memory, short-term memory (STM), and long-term memory (LTM). These distinctions are made on the basis of storage duration and capacity. Sensory memory is the smallest on both of these dimensions: its duration is on the order of milliseconds, and it stores just whatever the previous sensory configuration of the system was. Short-term memory, on

Atkinson and Shiffrin's (1968) is roughly equivalent to working memory. This component allows for maintenance and manipulation of a small (between 7 and 10 pieces; Miller, 1956) amount of information for a brief amount of time.

Most theories of memory maintain that processing by working memory is a necessary precursor to storage by long-term memory (LTM). On Atkinson and Shiffrin's (1968) model, this component stores information spanning from the beginning of one's life to the 30 seconds preceding the present, thus having the longest duration and largest storage of all of the components. Neuropsychological dissociations have helped researchers distinguish the different systems comprising long-term memory, each of which supports unique cognitive or behavioral processes (see Figure 1.1). LTM is first divided into explicit/declarative (Eichenbaum, 1997) and implicit/non-declarative systems (Schacter, 1987), named so because the former system supports memory for information on which we can report and the latter supports memory for unconscious but behaviorally relevant information. For example, procedural memory, a type of implicit memory, allows you to do things like ride a bike and drink from a straw without needing to think twice; it is *knowledge-how*.

In what follows, I am concerned exclusively with the explicit memory system. The primary focus is on the episodic (sub)system, which supports memory for previous experience. This system has been most heavily treated by philosophers interested in the metaphysics of memory because it encompasses our ability for "mental time travel" (Tulving, 1985). This raises questions about how memory facilitates our experience of past events (e.g., are traces necessary for remembering?) and the kind of relationship it allows us to have with the content it transfers (e.g., are we directly related to the objects represented by memory or directly related to the representations of those objects?) The semantic (sub)system, supporting memory for facts and

information, has been the sole target of epistemologists of memory. This is likely because epistemology deals with beliefs and beliefs are expressed as propositions. However, a great deal of our beliefs are dependent upon contents produced by episodic memory, so it surprising that episodic memory has been so overlooked in epistemological literature. Because fully understanding how episodic memory factors into knowledge requires understanding the role that episodic memory plays as a cognitive process, I turn next to a brief review of the scientific literature on episodic memory.

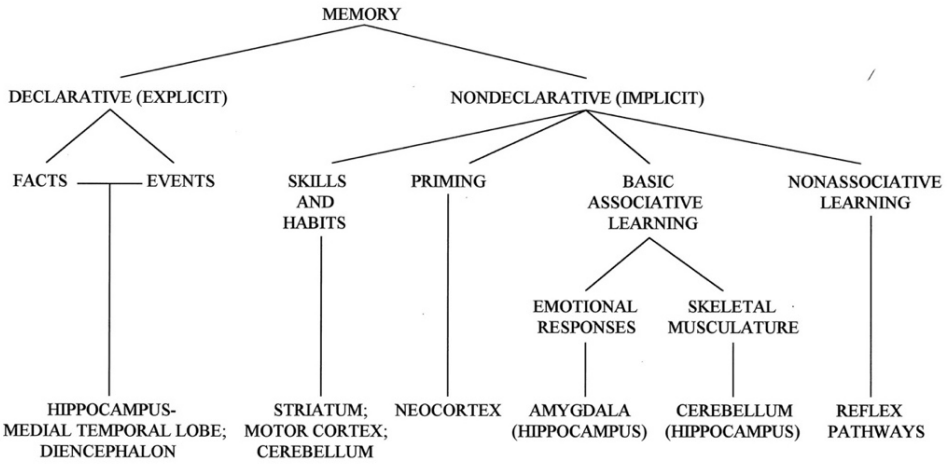


Figure 1.1: Squire and Zola's (1996) taxonomy of long-term memory

1.3 Measuring memory

1.3.1 Insights from neuropsychology

As noted earlier, the development of the standard taxonomy of long-term memory relied heavily on neuropsychological dissociations. This is true not only for the cognitive taxonomy (the third branched level), but also for its grounding in neuroanatomy (the branch terminals). Finding that bilateral hippocampal lesions selectively impaired episodic memory capacities while

keeping other memory capacities intact, as in the famous case of H.M., allows researchers to establish the closest thing to causality possible in neuroscience (Yoshihara and Yoshihara, 2018). For this reason, neuropsychology was the dominant methodology for studying the neuroscience of human memory for most of the twentieth century. In fact, Tulving's (1985) famous taxonomy of memory and consciousness was presented alongside a case study of a patient with selective phenomenological deficits.

Tulving reports a series of observations about patient N.N., who has suffered amnesia as a result of a car accident. N.N. demonstrates no linguistic deficits, is capable of describing the daily schedule of a college student, can identify and draw things like the North American continent and the Statue of liberty, and is well aware of the standard metrics used to measure time. Further, he can tell you what year his family moved into the house where they lived at the time, the names of schools he has attended, and where he spent his summers as a teenager. However, he is incapable of describing a single *event* from his personal life, including what he was doing fifteen minutes ago or what he plans to do tomorrow. When asked about the content of his mental state when he tries to retrieve that kind of information, he says that it's "blank, I guess" (Tulving, 1985, p. 5).

Tulving argued that N.N.'s case supports his class-inclusion hierarchy of memory processes and consciousness. Specifically, N.N. demonstrated a selective deficit in autonoetic consciousness, "the kind of consciousness that mediates an individual's awareness of his or her existence and identity in subjective time extending from the personal past through the present to the personal future," which is necessary for episodic memory (Tulving, 1985, p. 2). That N.N.'s ability to act flexibly on symbolic knowledge of the world, and the objects and events which constitute the world, indicates that his capacity for noetic consciousness, necessary for semantic

memory, was still intact. Tulving does not detail the extent of N.N.'s damage, but corroborating it with reports from patient H.M. strongly suggests hippocampal damage. One could even argue that observing such impairments without knowing the locus of damage controls for expectancy effects and strengthens Tulving's position, and that these impairments are interesting to note regardless of corresponding neural evidence. In any case, N.N. constitutes another important contribution to the neuropsychological literature grounding our understanding of memory systems, and Tulving's phenomenologically-grounded taxonomy of memory remains hugely influential to this day.

1.3.2 Recollection and familiarity

Tulving's taxonomy perhaps would not have persisted so robustly if he had not also developed a method for measuring noetic and autonoetic consciousness in healthy individuals. This consists in asking participants if they made whatever memory judgment the experimenter asked of them on the basis of *recollection* or *familiarity.* In the 1985 study, participants were asked if their response was made because they "remembered" studying the item or if they simply "knew" it some other way. "Remembering" is understood to consist in autonoetic consciousness, whereas "knowing" consists in noetic consciousness. Participants are instructed to respond with "remember" if their experience at retrieval contains information about the initial encoding event. This can be some association the experimenter asked them to form between the object and its context or any independently generated elaborations about the object; e.g., thinking that this picture of a house looks like your neighbor's house, and subsequently remembering that train of thought when recognizing the image on a memory test warrants a "remember" response. The experience of "knowing" is typically likened to the feeling one has when recognizing a face but being unable to determine why the face seems familiar.

1.3.3 Retrieval processes

The qualitative differences between recollection and familiarity correlate with quantifiably different retrieval processes. These are classified as *recall* and *recognition*, and agents regularly undertake both processes several times each day. Recall consists in generating information about a past experience, and as such requires a relatively strong memory trace. In a laboratory setting, this can be done by responding to a cue (some piece of information representing the original experience without encompassing it) or simply generating as much information about specific past events as possible (a paradigm called *free recall*). For example, you might be cued with an image and be asked to generate (recall) the word you were asked to associate with that image at encoding. Recognition consists in viewing a series of old and new items and reporting whether the item seems old or new in the context of the experiment (i.e., whether it was studied or not). Because accuracy can operate on the basis of stimulus familiarity, recognition can generally operate via a weaker memory trace than recall. Generally, successful recall depends on recollective memory. Recognition can be done on the basis of recollection or familiarity, and researchers have been using Tulving's RKN (remember, know, new) procedure as an index of the phenomenology accompanying recognition memory.

There are a number of approaches researchers have taken to quantify retrieval success. A common procedure for calculating recognition accuracy is to calculate the proportion of items in the recognition memory test that are hits (responding "old" to an old item) or correct rejections (responding "new" to a new item). Any number less than 1 indicates that participants either forgot a stimulus item (i.e., a miss) or reported memory for a totally new item (i.e., a false alarm). These false alarms are of primary interest to the current project. Most participants can be expected to false alarm to a few items in a recognition task, usually because of lapses in attention or human error. However, there have been striking demonstrations not only of systematic and

item-specific false alarms but also of recollection (in recall tasks) of information that was never presented during study. Early evidence of these two kinds of false alarms spawned decades of behavioral, neuropsychological, and neuroimaging investigation into false memories and the circumstances under which they arise.

1.4 Empirical evidence of false and distorted memories

1.4.1 Stories, scenes, and semantic associates

Bartlett (1932) is typically credited with the first empirical demonstration of false memories. He asked participants to read a passage from "The War of the Ghosts," a Native American mythology, and then recall the passage from memory. He found that recall accuracy decreased with the number of recall repetitions, but no one has been able to replicate this. In fact, Wheeler and Roediger (1992) found that repeated recall of the same passage *increased* memory accuracy. Despite this curiosity, Bartlett's work remains seminal because he posited the distinction between reproductive and reconstructive memory, and argued that memory is primarily reconstructive: "In a world of constantly changing environment, literal recall is extraordinarily unimportant...if we consider evidence rather than supposition, memory appears to be far more decisively an affair of construction rather than one of mere reproduction" (Bartlett, 1932, pp. 204-205). That is, the reproductive view would predict that memory would preserve all of the details from the first reading, and deficits in accuracy would take the form of forgetting some of those details. What Bartlett found, however, was that people recalled details which had not been present in the first reading, suggesting that memory can operate by *constructing* information, rather than solely reproducing it.

Much stronger evidence for the flexible, constructive nature of memory came around 40 years later. Loftus and her colleagues developed novel techniques for demonstrating systematic

memory distortions for complex, naturalistic events as a function of the linguistic structure of misleading post-event information (MPI). For example, Loftus (1974) showed that, after participants viewed a scene of a car accident, simply asking someone "Did you see *the* broken headlight?" rather than "Did you see *a* broken headlight?" greatly increases erroneous memory reports on a recall test. Interestingly, this effect was time-dependent -- memory errors were much higher when the MPI was communicated directly after viewing the scene than in cases when it was communicated a few days after. This supports the view that consolidation operates to stabilize a memory trace and highlights the malleability of recently formed memories. Loftus and Palmer (1974) showed that when participants viewed identical footage of a car accident, the group asked "How fast were the cars going when they *smashed* into each other?" consistently reported greater speed estimates than the groups where "smashed" was replaced with "collided," "bumped," "contact," or "hit." These demonstrations are now known as the *misinformation effect*, which Loftus (2005) has defined as "the impairment of memory for the past that arises after exposure to misleading information." Though the recent resurgence of naturalistic stimuli like those used by Loftus and colleagues suggests that her work was ahead of its time, there are a relatively limited amount of experimental manipulations one can employ with the MPI approach. And though this was pre-fMRI, the complex nature of the stimuli involved would make identifying neural correlates of these memory errors rather messy.

For these two reasons, Roediger and McDermott's (1995) adaptation of Deese's (1959) semantic list paradigm has since dominated the empirical literature on false memories. In a DRM (Deese-Roediger-McDermott) task, participants study a series of words ("bed," "rest," "pillow," etc.) and then take a free recall or recognition test. Participants will falsely remember studying the *critical lure*, "sleep," as frequently as they correctly remember items presented in

the middle of the list (Roediger and McDermott, 1995). They readily generate critical lures in free recall tests, use R responses on cued recognition tests, and report high levels of confidence in both of these types of responses, indicating that these false memories are phenomenologically rich and not merely artifacts of something like an unconscious decision bias. In fact, participants are often shocked to learn that they had *not* studied a critical lure because their memory for that item is so vivid. The robustness of this effect across a number of environmental conditions has resulted in an extensive research program which has rigorously detailed the conditions under which the effect occurs and how it is modulated by a number of different manipulations.

1.4.2 False or distorted?

In the psychological literature, the distortion and falsity of a memory are often conflated. If one reports seeing something other than what was present in the encoding event, she is said to be having a false memory. While this seems intuitive, De Brigard (2014) points out that there is an important difference between a memory that is false and a memory that is distorted. Unless one posits that the only "true" or genuine instances of remembering are those in which the content of a memory is identical to the content of the initial experience, then not all distorted memories are false. Indeed, on the constructive view, some degree of distortion is nearly inevitable during the reconstruction process.[1] As this has long been the default view in psychology, hardly any researcher would reject this latter point or maintain that genuine remembering only occurs when the content of an experience and the memory thereof is identical (which tacitly assumes the preservative view). This naturally raises the question of just how

[1] Further, a number of internal processes can unfold during encoding of a perceptual event. One could argue that the unverifiability of these processes severely constrains the sense in which any memory can be evaluated as "true."

distorted a memory can be before it is better classified as false. As a first step toward identifying this distinction, I suggest that false memories operate at the event-level whereas distortions occur at the detail-level within a particular event.

This, of course, raises the question of how researchers should define an "event." Traditionally, each item in an episodic memory test is operationalized to constitute an event of its own. This is presumably why Roediger and McDermott (1995) felt justified in concluded their abstract by writing "The results reveal a powerful illusion of memory: people remember events that never happened" (p. 1). However, the validity of claiming that a single word constitutes an encoding event which is unique from encoding the whole list, or is unique from the whole experiment event, is not free from challenges. The intuitive appeal of framing each DRM list as an independent event with each item as a detail is strong, and the appeal for framing the whole memory experiment as an event is perhaps stronger. Indeed, if someone claimed to have participated in a DRM experiment but had never done so, such a report would unambiguously constitute a false memory. Further, if someone recalled the critical lure "sleep" after studying a list where the critical lure was "chair," this report would very clearly represent a memory error. But confidently claiming to have heard the word "sleep" when only the words "bed," "rest," "awake," and "tired" were read also seems to qualify unambiguously as a false memory.

This latter approach has likely been dominant in empirical literature because it increases the total amount of events which in turn increases statistical power, effectively maximizing the likelihood that researchers detect an effect. Although it is less-than-naturalistic or intuitive, the item-as-event design further allows for a great deal of control and systematic manipulation to conduct structured investigations into the specific conditions under which these illusions arise.

On the grounds that recollection or confident recognition of a critical lure constitutes memory for an event which never happened, we can say that false memories consist in reporting memory for information which was not explicitly included in encoding events. Memory distortions, then, consist in misremembering details about information which was explicitly presented during encoding events. The *picture misattribution effect* (Foley et al., 2015) demonstrates this formula nicely. After showing participants a series of items presented either as pictures or words and testing their recognition memory for the item and how it was presented (i.e., the source), the effect occurs when participants correctly identify an item as "old" but report having seen it as a picture when it was actually presented as a word. That is, they correctly remember information included in the initial event (i.e., the item itself) but misremember details about it.

Some researchers have taken the position that picture misattributions represent false memories for source information. On the traditional formula of false memory, this is justified: if a participant studies the word "bread" but then reports seeing an image of bread, this is a memory for something which never happened. But it is not the case that they never encountered the concept "bread" or that images were never used in the experiment. Rather, upon feeling a sense of familiarity after hearing the word "bread" and then being forced to decide between whether that familiarity stemmed from seeing the concept represented as a picture or as a word, they chose to respond with picture. This seems more consistent with misremembering details (albeit rather important details) about an event rather than falsely remembering an event *per se.* There are of course intermediate cases which may challenge this formula, and much more work needs to be done on this topic, but for present purposes I will treat source misattributions as the

"upper bound" on the distorted/false continuum. "Misremembering" will be the blanket term used to refer to both false and distorted memories.

1.5 Misremembering: bug or feature?

1.5.1 Misremembering as a bug in the storehouse model

The storehouse model implies two metaphysical properties of memory that operate in the service of its functional role, which is understood as preservation of the past. These are (1) a unique record (or trace) for each specific memory and (2) a location where these records are kept. The location implication is less restrictive than the record implication – arguing that records are distributed across the brain is logically compatible with the storehouse model, though this does considerably weaken the metaphor. But a storehouse theorist has to maintain the position of a unique record for each experience, as this record is precisely the medium whereby memory preserves past experience. False memories, as defined above, raise significant problems for the storehouse view -- how can there be a record for an event which never happened?

A common response has been to argue that false memories are analogous to perceptual hallucinations, which are natural byproducts of a properly functioning cognitive system. This line of thinking has motivated philosophers to prefer indirect representationalism, the position which maintains that we have a first-order relationship with our memory representations, over direct representationalism, the position which maintains that we have a first-order relationship with the objects of our memory representations. The primary issue with direct representationalism (in this context), is that it leads to disjunctivism, the view that veridical remembering and false memories are states of fundamentally different kinds (Michaelian and Sutton, 2017); i.e., they are products of distinct cognitive systems. This is a large price to pay for those interested in developing a naturalistic theory of memory for two reasons: (1) it is less

parsimonious than accepting that veridical and false memories are products of the same system[2] and (2) empirical evidence demonstrates that these instances of remembering do indeed rely on the same system.

Even on an indirect representational position, memory distortions raise further issues for the storehouse model. Namely, they raise issues for the functional commitment of the storehouse model: the belief that memory functions to preserve the past. If this is an accurate description memory's function, then why is it the case that memory representations regularly contain information for details which were not present during the encoding event? How is a storehouse theorist to account for the incorporation of misleading post-event information into a memory record? In the next section, I show how investigations into the neural mechanisms of constructive memory invalidate the metaphysical implications of the storehouse model, and in the following section, show how this evidence undermines the functional commitment of the storehouse model as well.

1.5.2 Mechanisms of (mis)remembering

Schacter et al.'s (1998) constructive memory framework (CMF) synthesized insights from empirical and theoretical work on the cognitive psychology of misremembering with models of memory processes to create a theory of the neurocognitive processes involved in constructive memory. The success this framework has had in explaining a host of memory phenomena at both neural and cognitive levels has rendered it the received view in the cognitive neuroscience of memory. It takes as its starting point the idea that experiences can be broken

2 Why would evolution have selected for a system which misleads us? If false beliefs are maladaptive (Dennett and McKay, 2009), it seems highly implausible that we would have evolved a system devoted to accurately preserving the past and another devoted to inaccurately doing so.

down into patterns of activation in different modules of the brain which are specialized for processing different features of the experience (e.g., Johnson and Chalfonte, 1994, Damasio, 1989). These distributed patterns of activation are bound together to create a unified representation (Moscovitch, 1994), the reconstruction of which is the function of memory retrieval. Reconstruction operates via a pattern completion process (McClelland et al., 1995), in which a subset of the features comprising a past experience are reactivated, and this triggers reinstatement of the patterns of activation comprising the other features of the experience.

Three important implications follow from this. The first is strictly empirical – the quality of a particular memory representation can be quantified as the similarity between the representations at encoding and retrieval (Kriegeskorte et al., 2008). Second, as highlighted by Damasio (1989), there is no single record or trace of a memory on the CMF. Individuation of episodes is achieved via pattern separation (McClelland et al., 1995), a process which functions to reduce the representational overlap among different experiences. Although the distributed nature of memory representations can be compatible with the second metaphysical commitment of the storehouse model, it is clearly incompatible with first. This is a consequence of the fact that pattern separation functions only to *reduce* the representational overlap between memory representations, rather than altogether abolishing any similarity by creating a totally unique memory representation for each experience. Third, the CMF highlights the fact that regions which are specialized for one function (e.g., perception) are redeployed to serve a different function during memory.

1.5.3 Misremembering as a feature in the episodic hypothetical thought model

Given the magnitude of evidence in support of the CMF, Schacter and Addis (2007) pursue the question of *why* memory should function in this active and constructive manner,

rather than as a passive replay of stored experiences. Drawing on insights from a number of theorists (e.g., Tulving, 2002; Suddendorf and Corballis, 1997; Atance and O'Neill, 2005; Schacter and Addis, 2007a) who emphasize the additional role which memory plays in allowing individuals to imagine future events, they propose the *constructive episodic simulation* hypothesis. This suggests that constructive memory has its origins in the evolutionary advantage gained by the capacity to simulate future events. Because the future is not an exact replication of the past, successful simulation requires a system which can flexibly extract and recombine information from past experience. In support of their hypothesis, they review a host of cognitive, neuropsychological, and neuroimaging evidence which shows that the neurocognitive processes involved in recollection are largely overlapping with those involved in episodic future thinking and imagination.

Although Schacter and Addis's (2007) constructive episodic simulation hypothesis significantly expands the function of memory beyond that implied by the storehouse view, their functional definition is not necessarily incompatible with preservation of the past. Indeed, they write that "The constructive episodic simulation hypothesis does not imply that the only function of episodic memory is to allow us to simulate future events, nor do we believe that its role in simulation of the future constitutes the sole reason why episodic memory is primarily constructive rather than reproductive" (p. 6). De Brigard (2014), however, proposes a more radical theory in which recollection is a subprocess of a larger cognitive system which functions to generate episodic hypothetical thoughts, or "self-centered mental simulations about possible events that we think may happen or may have happened to ourselves" (p.19). He reaches this conclusion by using a mechanistic role function approach (Cummins, 1983; Craver, 2001) to answer the question of what memory is for. This approach consists in considering the role which

a mechanism plays in contributing to the goals of a cognitive organism. Importantly, properly understanding the mechanistic role function of a system requires understanding (1) how the mechanisms that compose the system work and (2) understanding how the system contributes to the function of the larger system within which it is contained (De Brigard, 2014).

With respect to (1), he builds on the CMF by suggesting that retrieval consist in optimal reconstruction of a past experience, where the construction process is constrained by schematic knowledge and prior encounters with the target memory. Schematic knowledge, or schemas, consists of category-specific information which has been abstracted from a number of different experiences belonging to the same category. For example, the schema for an intersection would include the context of being inside a car (whether in the driver or passenger seat depends on how frequently one sits in each), 3 streets that are perpendicular to the current position, and a stop sign in the middle of the right side of the visual field. These further serve to optimize the encoding process by minimizing the amount of information which requires attention and processing by working memory. If perceptual input is consistent with its schema, then memory can simply "tag" that schema as active during this experience and use the extra resources to process and represent information which is unique to this event belonging to a particular schema. If there is a mismatch between schematic knowledge and perceptual input (i.e., a prediction error), then attention is directed toward the locus of the discrepancy and that information is incorporated into the memory representation and is added to the bank of priors comprising schematic knowledge.

De Brigard (2014) makes the point that because encoding operates via schema-tagging, memory needs to be the kind of process that not only binds together disaggregated features of an experience, but also fills in details which, by virtue of adhering to schematic prior knowledge,

were not encoded into the memory representation. It follows, then, that the features which are filled in will be precisely those features which compose the schema for the event – these are the most probable given the nature of the target episode. By virtue of its probabilistic nature, this filling-in mechanism accurately generates information about the encoding experience most of the time. However, in the occasional circumstances where unencoded information deviates from what is contained in a schema, the filling-in mechanism will result in misremembering. But, as emphasized by both Schacter and Addis (2007) and De Brigard (2014a), these instances of misremembering are characteristic of a properly functioning episodic construction system. On the episodic hypothetical thought theory, misremembering is a form of *episodic counterfactual thought*, a simulation of what could have happened.

We can see that with respect to the second requirement for developing a mechanistic role function account for memory, an individual who has the capacity to simulate (1) what was the case, (2) what could have been the case, and (3) what might be the case will have an advantage over an agent who only has the capacity to simulate (1) and (3). In addition to the obvious quantitative advantage, the capacity to simulate what could have been the case can create important constraints on simulations of future actions, which are arguably the most important evolutionarily.[3] And if we situate memory as a system that contributes to the function of a

[3] One might wonder why simulating what could have been the case is important for constraining the content of future-oriented hypothetical simulations. Wouldn't that interfere with setting priors on the basis of past experience, which should presumably be the strongest constraint on any simulation? Although my knowledge of Bayesian updating is too scant to give a thorough how-possibly response, the fact that contemporary systems are capable of separating priors and utilizing them accordingly is evidence for the emergence of this capacity. My hunch is that counterfactual simulations were critical (1) for learning about causal relationships both among actors and items in the external world and between the agent and its environment and (2) for reinforcing instinctive responses to happenings in the world. Imagine an agent

cognitive organism whose goal is survival, then it seems that the episodic hypothetical thought theory gives a better account of *why* memory is constructive: it is a subroutine of a larger system that functions to construct episodic hypothetical simulations.

But with remembering now situated as a subspecies of imagining, it is less clear how agents are able to distinguish between hypothetical and counterfactual simulations. This introduces what Michaelian (2016) has termed the process problem: how do agents determine whether the content of an episodic simulation represents a past experience or is simply drawing on that experience to create a hypothetical simulation? The next chapter presents a version of Michaelian's (2016) process monitoring framework that has been adjusted to cohere with the mechanistic principles of the episodic hypothetical thought theory. Process monitoring is closely related to reality monitoring, a metamemory process that evaluates whether the content of an optimal reconstruction stems from perceptual derived or internally generated information. Reality monitoring itself is a species of source monitoring, the broader metamemory process that

---

narrowly avoids being run over by a speeding car because he instinctively snapped his neck in response to the sound of barreling down the street. Thinking about what would have happened had he not done so happens almost involuntarily. This (1) unconsciously reinforces the instinctive response to look in the direction of loud sounds, (2) reminds the agent to be more attentive to happenings in his environment, which (3) unconsciously (or perhaps consciously, if the agent is particularly reflective and interested in these questions) reinforces his sense of autonomy as an agent in this world, further (4) supplementing unconscious and conscious causal models of the agent-environment relationship. I imagine that the primary advantage to be gained from counterfactual simulations is such generation of causal models, which eventually became but another reinforcer of instinctive responses that were inherited from primordial ancestors who were not endowed with these advanced simulational capacities. One final thought is that organisms capable of simulating possible events before acting on the world occupy a different level on Dennett's Tower of Generate and Test (1969) than agents who can only learn about the world by acting on it. The former are Popperian creatures and the latter are Skinnerian, and each level in the tower represents a qualitative (or nonlinear) advance in fitness and adaptability.

additionally allows agents to discriminate between external (this or that agent?) sources and internal (did I do this or just think about doing it?) sources of memory.

Chapter Two: Monitoring and Reliability

2.1 Clearing up conceptual ambiguities

2.1.1 Levels of explanation

Because of its interdisciplinary nature, cognitive science operates at a number of different levels of explanation. While this is necessary for achieving a comprehensive understanding of cognition, it can result in a good deal of researchers talking past each other when discussing issues which span multiple levels (which is most of them). To help resolve some of this ambiguity, Dennett (1969) proposed a distinction between personal and sub-personal levels of investigation and explanation. The personal level consists in the processes an agent can report using to complete a cognitive task; in effect, personal level processes are consciously and deliberately deployed by an agent to achieve a goal. The sub-personal level consists in the physical and biological processes which support cognition that are inaccessible to an agent by means of introspection, no matter how trained.

Although this is a helpful and reasonably straightforward distinction, much of the language used in cognitive psychology and neuroscience has been regrettably ambiguous with respect to these levels. For example, Schacter and Addis (2007) write, "Retrieval of a past experience involves a process of pattern completion … in which the *rememberer* pieces together some subset of distributed features that comprise a particular past experience, including perceptual and conceptual/interpretive elements" (774, emphasis added). While it is true that agents can engage in this kind of piecing-together process, the process of pattern completion which Schacter and Addis (2007) describe almost certainly occurs at the sub-personal level. If it

were the case that agents experienced or engaged in pattern completion processes when recalling episodic memories, the storehouse model would be totally unfounded -- a hallmark of episodic memories is precisely the unity with which they "unfold before the mind's eye."[4]

2.1.2 The trap of introspection

But, as we have seen, memory functions more like trying to put together a dinosaur from fossilized remains (Neisser, 1967) than a video camera. This mismatch between retrieval processes and the content they transfer is one of the strongest demonstrations of what Place (1956) called the *phenomenological fallacy:* "the mistaken idea that descriptions of the appearances of things are descriptions of the actual state of affairs in a mysterious internal environment" (44). Pylyshyn (2002) slightly revised this formula and dubbed his version *the intentional fallacy*: "when we 'examine our mental image' we are not in fact examining an inner state, but rather are contemplating what the inner state is about – that is, some possible state of the visible world – and therefore this experience tells us nothing about the nature and form of the representation" (158). In both cases, the mistake consists in thinking that there is a necessary relationship between personal level phenomena and the sub-personal level processes which

[4] Dennett (1993) and Pylyshyn (1973, 2002) have argued convincingly against the use of this language when describing the contents of conscious experience. The issue with the mind's eye metaphor, for Dennett, is that it implies a second transduction of the content of consciousness by an internal observer, when clearly no such observer exists. For Pylyshyn the issue is similar, but his emphasis is on the fact that mental images are not depictive in any meaningful sense, such that there is nothing to be interpreted by the mind's eye. Dennett, following Place, suggests making the metaphor explicit by prefacing the comparison with "something like," as in "something like watching an episode unfold before your mind's eye." I will try to follow this suggestion as much as possible. When doing so comes across as syntactically awkward, I will be sure to enclose the metaphor in scare quotes.

facilitate them, or that agents have 'privileged access' to the workings of their mind.[5] There may be some cases where the structure of our experience bears an important relationship with the neural representation which instantiates it, but these are the exception, not the norm.[6] Further, it is indeed the case that agents have privileged access to the *contents* of the cognitive system, but they do not have any sort of first-person access to the *vehicles* (or processes) which facilitate transfers of content (De Brigard, 2014).

In hopes of maintaining conceptual clarity, I will make substantial use of the personal and sub-personal levels in what follows. I begin by identifying the form Michaelian's (2016) process problem takes within the episodic hypothetical thought theory and propose some revisions to his solutions. When the answer to the process problem (memory or imagination?) is memory, agents are then faced with solving the source problem. For Michaelian, the source problem consists in determining whether the content of a memory has its origins in a reliable source, but I will discuss a number of variations on the source problem and argue that Johnson et al.'s (1993) source monitoring framework explains how these problems are solved. I conclude with a specific

[5] It almost certainly the case that most researchers do not actually believe that subjective experience mimics the brain processes which create it. Rather, the point is that no one has explicitly said "these neural processes are not accessible through introspection," or, more subtly, using personal-level language to describe sub-personal processes. Perhaps this has seemed so obvious to cognitive neuroscientists that it did not warrant mention, but this silence has caused a great deal of erroneous conclusions reached through means of armchair introspection.

[6] Kosslyn and colleagues, one of the primary motivators for Pylyshyn in writing his 2002 paper, insist that the retinotopic organization of V1 demonstrates a case where the content of an experience meaningfully resembles the neural representation which creates it, and I generally agree with them. Even though the information conveyed via the retinotopic representation is preserved throughout the visual stream, there is nothing about *those* representations which meaningfully resembles the experiences which correlate with them.

question about encoding conditions contributing to picture misattributions (a source error) and, in Chapter 3, detail the results of an experiment conducted to answer this question.

2.1.3 A note on language

Perhaps no other word in cognitive science has been as burdened by differences in levels as "representation." This should not be surprising — the ambiguity of the word itself naturally leads to application of it to concepts that are still a bit fuzzy. In this project, my use of "representation" is intended to refer only to happenings at the sub-personal level. So when I use the phrase "memory representation," I mean the patterns of neural activity that correlate with behavioral reports of remembering. I use the words "experience" or "simulation" to refer to the personal-level phenomenology that accompanies these neural signatures. Indeed, when I talk about monitoring memory representations later in this chapter, I intend to refer only to sub-personal neural processes. Personal-level monitoring applies only to assessing (1) the content of a simulation and (2) the attitude we have toward it (i.e., the feeling of pastness or familiarity that accompanies it, or feelings of metacognitive confidence).

2.2 Defining the process problem

Like De Brigard (2014), Michaelian (2016) defends a simulation theory of episodic memory. However, in order to allow for incorporation of testimonial information into his account, Michaelian ends up defending a radical simulation theory. He rejects the necessity of any memory trace linking a memory representation to a personally experienced past event, arguing that all episodic memory consists in imagining the personal past. He contends that although this eliminates any metaphysical marker distinguishing remembering from imagining, there are still agent-level markers which allow *them* to distinguish these processes. These are

what solve Michaelian's version of the process problem: how do agents know whether they are engaging in remembering rather than some other kind of episodic simulation?

A key difference between the radical simulation and episodic hypothetical thought theories is that the latter admits of a causal relationship between encoding conditions and memory representations. Namely, that the strength of the disposition to reactivate in the same pattern as encoding is causally linked to the strength of coactivation during encoding.[7] There can, of course, be processes which interfere with this relationship, but they do not necessarily preclude the possibility of causal link. This link establishes the metaphysical marker distinguishing remembering from other episodic simulations: representational similarity. Because remembering consists in optimal reconstruction of a previous experience, one can use multi-voxel pattern analyses (MVPA) to quantify the similarity between a mental representation constructed by memory and a mental representation driven by encoding (e.g., Ritchey et al., 2013). Further, if one had access to a dataset of neural activity involved only in counterfactual simulations and another of neural activity involved only in autobiographical remembering, it should be possible to train a classifier to distinguish representations produced by the same process but with importantly different contents. However, since remembering and hypothetical thinking operate via the same process, the possibility of identifying a metaphysical marker has no bearing on the process problem. Therefore, the process problem on the episodic hypothetical thought theory is identical to the problem on the radical simulation theory.

[7] This is De Brigard's (2014) definition of a memory trace. Any mention of a memory trace in what follows refers to this dispositional property (which, importantly, does not have the ontological status of an object or event).

2.3 Refining Michaelian's solutions

Drawing on Johnson et al.'s (1993) source monitoring framework, Michaelian proposes three classes of markers which agents can use to solve the process problem. Neither of these classes alone is sufficient for solving the process problem, but together they give a plausible account of how agents can reliably discriminate among episodic simulations. The three classes he identifies are formal, content-based, and phenomenal. Formal markers consist in information about how a representation is generated, and these Michaelian identifies as flexibility, spontaneity, and intentionality. Flexibility and spontaneity are highly related -- spontaneously generated simulations are usually less flexible than deliberately generated simulations. Intentionality plays a relatively limited role, as it only applies to instances of deliberate remembering.

Content-based markers point toward the constructive process by appealing to the vivacity, coherence, and valence of an episodic simulation. Simulations of actual events are usually more detailed than hypothetical simulations, and their content is more likely to cohere with an agent's semantic autobiographical knowledge. Further, future events are generally more positively valenced than past experiences (e.g., D'Argembeau et al. 2011). And phenomenal markers aid the agent in distinguishing constructive processes by imbuing her with a sense of pastness or feeling futurity which accompanies the content of an episodic simulation. These content-based and phenomenal markers translate smoothly from the radical simulation theory to the episodic hypothetical thought theory. However, the subtle mechanistic differences between the two motivates some revisions to the formal markers as developed by Michaelian (2016). Further, the inclusion of traces on the hypothetical thought theory allows for a more thorough understanding of the feeling of pastness as a phenomenal marker. I will treat these points in turn.

2.3.1 Intentionality

As on the radical simulation theory, intentionality is the gold-standard when it comes to solving the process problem. According to this marker, the agent's intention in constructing a simulation is what allows her to identify the process which generates it. The power this marker has comes at the cost of ubiquity– intentionality can only be used as a marker in cases of deliberate simulation, which by no means exhaust the circumstances under which simulations can be generated. And it is precisely cases of spontaneous simulation where the process problem is most dire. Therefore, intentionality is just as limited of a marker on the episodic hypothetical thought theory as it is on the radical simulation theory.

However, the intentionality marker does help explain why agents mistake counterfactual simulations for optimal reconstructions of the past. During a memory test, agents are consciously using their memory to answer questions about what was seen during encoding; i.e., their intention in constructing episodic simulations is to optimally reconstruct the past. This leads them to believe that whatever appears "before their minds' eye" will be a representation of the past. But, as we have seen, if a question asks about information which was not encoded into the memory representation for an event (on the item-as-event design), then the reconstruction will be of what could have happened during that event. If the experimenter is interested in misremembering, then the likelihood that the content of this simulation accurately represents the original event decreases significantly. But the intentionality marker is just as strong for the agent (presumably unaware of the experimenter's deliberate manipulation of her memory system), so she confidently endorses the simulation as representing a past event.

2.3.2 Flexibility and spontaneity

The use of flexibility as a process marker finds its roots in Hume, who argued that memory is constrained to preserve experience whereas imagination is free "to transpose and change its ideas" (Hume, 1739, p. 25). Michaelian suggests that rather than positing it as a qualitative (or metaphysical) marker, which is clearly out of line with both simulation theories, flexibility can be used to refer to the quantitative difference between remembering and hypothetical thinking. Because more transformations of information are involved when the system constructs a hypothetical rather than veridical simulation, if an agent has access to the degree of information transformation, he can use that to help solve the process problem. The obvious question then becomes how an agent would gain access to such information. Michaelian does not offer an answer directly, but after devoting an entire section to developing spontaneity as a process marker, he writes "it is not clear whether these [flexibility and spontaneity] should be treated as two separate criteria: access to information about the level of flexibility involved in a given episodic constructive process might manifest itself precisely in a sense that the process unfolds more or less spontaneously" (p. 185).

I think that this insight is correct, especially on the hypothetical thought theory. Flexibility, as Michaelian has defined it, does not seem tenable as a process marker. It is certainly a *characteristic* of different constructive processes, but it is not clear how this would be communicated to an agent in a manner other than through the spontaneity (or automaticity) with which a simulation is constructed. On the hypothetical thought theory, simulation can be modeled as a two-step process: (1) activating traces that reconstruct representations (pattern completion) and (2) binding together the disaggregated representations into a unified mental representation. The spontaneity of experience is constrained by each of these steps. Weaker traces will have a lower disposition to reactivate in the same pattern that encoded an experience,

lengthening the amount of time (and energy) needed to reconstruct that representation. Flexible recombination of representations takes more time than optimally reconstructing a previous global configuration, as there are preexisting connections among dispersed cortical representations which facilitate the binding process.

If a simulation is of an episode with strong traces, then the experiential "lag" which ensues from flexible recombination vs. optimal reconstruction can signal the constructive process to the agent.[8] It is difficult to conceive of an alternative mechanism whereby information about the flexibility of a process can be communicated to an agent without the process problem evaporating completely.[9] It seems, then, that flexibility should be abandoned as a process marker (which, by virtue of the problem, has to do its marking at the agent level) and reformulated as a sub-personal process marker. Spontaneity is the process marker which can signal to an agent the flexibility of the process which generated a simulation. Alternatively, spontaneity could signal the automaticity with which a simulation is generated. This is closely related to the flexibility of the constructive process, but it plays an additional role in the generation of phenomenal process markers.

---

[8] Hume (1739) makes the point that it is impossible "to recall the past impressions [sensory experiences], in order to compare them with our present ideas [imagination], and see whether their arrangement be exactly similar" (152-153). Even though this is glaringly obvious, or perhaps precisely for that reason, I have yet to come across a cognitive scientist (psychologist, neuroscientist, etc.) who has made this point explicit. This omission is innocuous enough, but it clearly implies that the implementation of these markers necessarily depends on memory. I will return to this point shortly.

[9] Although I am not ruling out the possibility of alternative explanations. Dennett regularly warns of the logical fallacy notoriously committed by armchair philosophers of mind: mistaking failures of imagination for insights into necessity. A classic example is the line of thought which claimed that it would be impossible to see an item without also seeing its color – a conclusion that was quickly disproved by Treisman's work on feature integration.

2.3.3 The feeling of pastness

Because the radical simulation theory rejects the necessity of a trace linking a memory representation to its encoding event, it does not have the resources to explain why remembering is accompanied with a feeling of pastness other than statistical regularity. However, I think that the hypothetical thought theory can give a non-arbitrary account of this relationship. If we assume optimal consolidation, then the strength of a memory trace directly reflects the degree of coactivation among the populations of neurons which comprise the disaggregated memory representation; the greater the extent of coactivation, the greater the disposition these neurons will have to reinstate that coactivation. In this way (or, at this level), the strength of a memory trace preserves information about encoding in a format that is content-free. Experiences of different events could result in result in memory traces of equivalent strengths that can be said to be preserving the same thing about those encoding events. I am inclined to argue that traces preserve information about the depth with which information at an encoding event was processed, which I think reflects the intensity of an encoding experience. Even if this is an inaccurate interpretation, the fact that traces preserve some physical information about an encoding event is enough for present purposes.

My suggestion is that the automaticity with which both steps of the episodic construction process are carried out signals to the system that the event being simulated belongs to the agent's personally experienced past. This is communicated to the agent through a feeling of familiarity or pastness, which I think can be understood as gradations of the same experience rather than qualitatively different states. Whereas the speed of the second step is more important in determining the spontaneity with which a simulation is experienced, the speed of the first step is more important for generating feelings of pastness. Because the automaticity with which distributed representations are reinstated stems directly from the encoding event (assuming

optimal consolidation), the system has learned to interpret automaticity of reconstruction as a signal of the "pastness" of a neural configuration – i.e., that it has been instated at a previous point in time. Further, because an event has been represented by the brain at a previous point in time, consolidation facilitates connections among disparate representations that guide the binding process. This increases the automaticity with which the second step of the reconstructive process is carried out, making the signal of pastness even stronger.

This is largely similar to the theory in cognitive science wherein fluency of processing produces feelings of familiarity. Though Michaelian (2016) posits that familiarity is neither necessary nor sufficient for pastness, the class-inclusionary nature of Tulving's taxonomy of memory and consciousness argues against this interpretation. On Tulving's theory, autonoetic consciousness (feeling of pastness) *is* noetic consciousness (feeling of familiarity) but with additional dimensions that distinguish it from its constituent classes of consciousness. Specifically, this is the dimension of mental time travel, or the feeling of the self moving through subjective time. Though this is undoubtedly the strongest process marker distinguishing optimal reconstructions from hypothetical simulations, there are also gradations to the intensity of autonoetic consciousness which accompanies a particular simulation.

It is not clear that the episodic hypothetical thought theory on its own has the resources to account for full-blown autonoetic consciousness, as in awareness of the self traveling through time. But it does have sufficient power to explain why certain simulations are accompanied with a sense of pastness rather than a sense of familiarity. The distinction between these feelings primarily serves to furnish the agent with a sense of confidence that the simulation represents an event from the personal past, which can assist them in discriminating between actual and counterfactual simulations. Further, it helps them distinguish future-oriented hypothetical

simulations from both counterfactual simulations and optimal reconstructions by virtue of the former being accompanied by a feeling of familiarity.

2.4 Two-level belief producing systems

The obvious question that follows is how these sub-personal processes generate process markers for an agent. This is where Michaelian invokes the idea of a two-level belief producing system, which is a standard model of metacognition with some linguistic adjustments that allow for smooth application to issues of epistemic reliability. In these systems, a first process (the "information producer") outputs information to serve as the content of a belief. This is processed by a second level (the "endorsement mechanism") that evaluates the output and determines whether to endorse or reject it. Critically, the endorsement mechanism has what Koriat (2000) has termed a "crossover" mode of operation: it acts unconsciously (or sub-personally) to produce beliefs or feelings that are accessible at the personal level.[10]

Drawing on Kahneman's (2011) distinction between Type 1 and Type 2 processes, Michaelian suggests that the information producer always operates in a Type 1 (heuristic,

[10] I think that this helps explain the property of ineffability that a number of philosophers (most notably Chalmers, 1996 and 2003) have touted as reasons to believe that consciousness defies explanation in terms of contemporary physical theory. Ineffability refers to the insufficiency of a system of symbolic language for capturing the essence of some entity; i.e., the only way we can describe the color red is through metaphor. Perhaps most famously, Chalmers (1996) used ineffability to develop the "hard problem" of consciousness: explaining why particular neural states create this experience (the color red, for example) rather than another. He insists that there must be some necessary relation between information in the world and its representation in the brain which causally links an experience (or "quale") with its neural instantiation, and that this relationship is what causes ineffability. However, if the account given in 2.3.3 is roughly correct, then "ineffability" is simply the consequence of the personal level bottoming-out its access to sub-personal processes by virtue of the crossover mode of operation of endorsement mechanisms.

unconscious, fast, automatic) manner and that the endorsement mechanism does so most of the time. This is supported by the inference that Type 2 processing incurs greater energetic costs, so the system will default to Type 1 processing for both levels. Instances when Type 2 evaluation needs to be undertaken are less frequent on the process problem than on the source problem. These consist primarily of situations where an agent is actively attempting to discriminate whether a simulation is recollective or imaginary. Otherwise, the system operates automatically to generate feelings of familiarity or pastness that allow agents to quickly identify recollective simulations. When an agent is deliberately trying to solve the source problem, this feeling becomes a marker that they can use in conjunction with other markers to determine the process generating a simulation.

2.5 The source problem and its solutions

The source problem can be broadly formulated as the question: where did the contents of this simulation come from? For Michaelian (and Johnson, implicitly), this is the question that agents need to answer after the process generating a simulation is identified as recollection. Because a core tenet of his radical simulation theory is that incorporation of testimonial information is a natural function of memory, the most critical component of Michaelian's source problem is determining whether the information contained in a simulation originates from a reliable agent. On Johnson et al.'s (1993) formula, the question is formulated as "specifying the conditions under which a memory is acquired" (p. 3). However, it seems that a variety of the source problem also exists for hypothetical and counterfactual simulations as construed on the episodic hypothetical thought theory. It consists in recognizing the elements of a simulation that stem from the personal past (e.g., your car) as distinct from the elements that have been fabricated (e.g., a unicorn) in service of constructing the hypothetical simulation. Johnson et al.'s

(1993) framework has the resources to account for all of these questions, so I will place my focus there.

2.5.1 The source monitoring framework

Johnson et al.'s (1993) source monitoring framework grew out of the reality monitoring framework developed by Johnson and Raye (1981). Whereas the reality monitoring framework addresses the question of how agents distinguish between memories of internally generated information and externally derived information, the source monitoring framework is additionally interested in how agents discriminate among external sources of memories (i.e., which agent relayed this information to me?) and memories for internally generated sources (i.e., did I say this or did I just think it?). Both frameworks emphasize that information about the origin of a memory representation is not explicitly encoded into it. Rather, source is inferred via decision processes that capitalize on average differences among memory representations originating in different sources (Johnson et al., 1993).

As such, source monitoring is a metacognitive process that operates in a similar manner as the two-level belief producing systems described in the previous section. There are a number of cues the endorsement mechanism can use to infer source: sensory/perceptual information, contextual (spatial and temporal) information, semantic detail, affect, and cognitive operations. The (mis)match between characteristics of a memory and your schema for a source can also factor into the decision-making process. The speed and accuracy with which source is inferred on the basis of these cues depends on (i) which and how many cues are present in a memory representation, (ii) how unique they are for a given source, and (iii) the efficacy of the judgment process itself and the criteria whereby it operates (Johnson et al., 1993).

2.5.2 Source cues

The reality monitoring framework details the average differences among memory representations that are used to solve the internal-external source problem. Johnson and Raye (1981) propose that

i) Externally generated memories have more sensory and contextual attributes (i.e., are more vivid) than internally generated memories,

ii) Externally generated memories contain more semantic information (i.e., detail) than internally generated memories,

iii) Internally generated memories have more information about cognitive operations than external memories.

Johnson et al. (1993) are less explicit about what average differences are used to solve internal and external source problems. Since external memories should have roughly equal amounts of perceptual and contextual information and semantic detail in addition to relatively little information about cognitive operations, it seems that schema-matching will be critical for distinguishing between external sources. For example, if you have a memory representation of a high-pitched voice conveying information about Disney World, matching the quality of the voice with the schema for your 8-year-old niece (which newly contains information that she visited Disney World) should help identify her as the source of that representation. On the other hand, it seems that the amount of cognitive operations involved in an internal generation process is key for distinguishing between internal sources. Constructing a hypothetical simulation of a

conversation should require more cognitive operations[11] than simply generating a response in a conversation.

2.5.3 Experience-dependency

As highlighted in footnote [8], how representative a collection of cues are for a particular source is largely a matter of statistical regularity. Memory systems need experience (i.e., function over time) in order to be capable of monitoring accuracy. Thus, each correct and incorrect source judgment increases the power memory systems have to discriminate among sources. Additionally, that external source monitoring seems to rely heavily on schema-matching directly implicates experience as a critical factor in source monitoring success. While it is likely that some monitoring capacities emerge simultaneously with the development of episodic memory, Foley et al (1983) and Foley and Johnson (1985), among others, have demonstrated that 6-year-olds are impaired in internal source monitoring, but not reality or external monitoring, compared to adults. This is likely a consequence of the fact that this is the same age around which episodic memory capacities begin to emerge, which means that the system has had limited experience generating episodic simulations. This means that cognitive operations are a new form of information that the system needs to learn how to utilize.

Additionally, Lindsay et al. (1991) showed that age-related differences in source monitoring abilities were greatest when source information was highly similar. For example, 8-year-olds, who should have perfectly intact reality monitoring abilities, had difficulty discriminating between memories for what another person had done and what the child had

[11] I am going to operate on a definition of cognitive operations as conscious or deliberate transformations of information that function to generate episodic simulations. These simulations can be of a stagnant, single image as well as of complex, dynamic events.

imagined that person to have done. This suggests that as sources become more similar on one dimension, the use of additional source cues becomes necessary to correctly discriminate between sources. That 8-year-olds struggled to distinguish between memories for their perceptions and imaginations of another individual's actions lends further support to the idea that the use of cognitive operations as a source cue emerges later on in the lifespan and is dependent on the number of episodic simulations that are carried out by an agent.

2.5.4 Judgment processes

The main point of the previous section was to highlight the fact that there is no necessary relationship between source characteristics and the sources they represent. This is because what they represent is ultimately a matter of how they are interpreted by the endorsement mechanism (cf. Millikan, 1989), which itself operates heuristically (i.e. Type 1) on the basis of average differences accrued through experience. As in process monitoring, however, the endorsement mechanism can initiate Type 2 processing to supplement the interpretation reached on the basis of Type 1 processing. I would like to argue that the crossover mode of operation of the endorsement mechanism, together with the ubiquity of its Type 1 processing, is responsible for creating feelings of confidence in a recollective simulation. The reasoning is similar to that employed in section 2.3.3. If a memory representation is sufficiently detailed in the ways it needs to be in order for the endorsement mechanism to make an accurate Type 1 source inference, then (1) the representation has the details needed to accurately convey the original experience and (2) the automaticity with which this content is simulated signals to the agent that the memory representation is robust.

However, if a memory representation is less-than-determinate, then the Type 1 endorsement mechanism may trigger Type 2 monitoring processes to help solve the problem.

Because Type 2 processing definitionally occurs at the personal level, this recruitment cannot go unnoticed by the agent. There are at least two outcomes after this occurs: (1) the Type 2 search ends relatively quickly and generates a sufficiently detailed simulation that the agent endorses with a fair amount of confidence or (2) the Type 2 search returns something like a 'NaN' or an incomplete source interpretation ("it was probably Maggie" vs. "Maggie by telephone"), leaving agents faced with a decision to endorse to reject the content of a simulation. In either of these cases, the agent's decision depends on task demands and her metamemory beliefs (Johnson et al., 1993). There are some cases, like eyewitness testimony, when agents might require a high degree of confidence in source before endorsing the simulation and reporting memory for the event. In other cases, like a casual conversation, the specificity of source might be less important.[12]

Johnson and Raye (1981) write that "an important point incorporated in the working model of reality monitoring that appears to be born [sic] out by our experiments is the idea that subjects' assumptions about how their memories work will play a critical role in decision strategies and biases operating during reality monitoring" (p. 80). In support of this, they cite evidence of a study they ran where participants reported using perceptual details to decide in favor of an external source and memories of cognitive processing to decide in favor of an internal source (Johnson, Raye, Foley, and Foley (1981). In the same publication, a different experiment found that many participants self-report beliefs that memory should be better for self-generated information. This was corroborated by their finding of a "it-had-to-be-you" effect: when assessing the source of familiarity with a totally new item (i.e., false alarm), participants

[12] It's worth noting that any time agents are asked to answer a question about their memory for source, they are *de facto* engaging in Type 2 monitoring.

consistently demonstrated a bias to attribute it to an external source rather than internal generation.

Although these insights are important to understanding how agents solve the source problem, it is surprising that Johnson et al (1981, 1993) did not propose an account of how overall metamemory beliefs should factor into source monitoring decision processes. I can imagine two alternative predictions: (1) participants with positive metamemory beliefs demonstrate lower criteria for endorsement ("my memory is really good so it's probably right this time too") or (2) participants with negative metamemory beliefs demonstrate lower criteria ("my memory is terrible. This vague sense of familiarity is the best it can do"). In the next chapter I report some data that speak to precisely this question.

In these ways, we can see that source monitoring is a complex metacognitive process that affords simulations their contextual detail and gives agents a sense of how robustly this is preserved in the memory trace. This makes it a key factor in securing the reliability of beliefs formed on the basis of episodic simulations by allowing agents to form attitudes toward the contents of their thoughts (i.e., metacognitive feelings). These can be directed toward the "resolution" (i.e., amount of detail) of the simulation or the trustworthiness of the source that generated that content. The insights that contributed to the development of this framework, however, emerged from structured investigation into encoding and retrieval conditions that lead to errors in source monitoring. Conducting a comprehensive review of the evidence demonstrating conditions under which source monitoring goes awry is beyond the scope of this project, so I will finish this chapter with a discussion of the source error that the experiment in Chapter 3 was designed to test.

2.6 The picture misattribution effect

Originally documented by Durso and Johnson (1980), the picture misattribution effect was named by Foley et al. (2015). The name refers to the disproportionate frequency with which participants report remembering seeing an item represented as a picture when it was actually represented as a word. Per the episodic hypothetical thought theory, these incorrect responses do not reflect a malfunctioning memory system – participants study a series of items, half of which are presented as pictures and half of which are presented as words. Critical to induction of the effect is an encoding question that requires semantic processing on the part of the participant. Durso and Johnson (1980) reported evidence for the effect under a number of different semantic encoding questions, but subsequent research has used only the question that produced the strongest effects: what do you do with this item? This was question not only produced the greatest amount of picture misattributions on a recognition memory test in Durso and Johnson (1980), but was also the only question that resulted in picture misattributions in free recall.

This was a surprising finding for Durso and Johnson (1980) because it was not accounted for by a number of then-prevalent models of memory. Nelson, Reed, & McEvoy's (1977) sensory-semantic model, for example, would have predicted that encoding questions encouraging generation of visual imagery should induce this effect. Durso and Johnson (1980), however, found that source memory was actually *better* for items encoded under these conditions. Because the prediction from the sensory-semantic model was that mental images generated during perception of words should override memory representations for the words on the basis of the picture superiority effect (i.e., images are better remembered than words), Durso and Johnson (1980) adapted this line of thinking to account for their observations.

They suggested that semantic processing of words evoked spontaneous imagery that overrode the memory representation for the veridical percept. Critically, because these mental

images were generated spontaneously, there was no information about cognitive operations to be incorporated into the memory representations. Deliberately generating images, however, did result in this information being incorporated into the memory representation, and Durso and Johnson (1980) argued that this was what allowed participants to accurately remember source information for items studied in that encoding condition. A number of studies have replicated this source memory dissociation introduced by encoding conditions (e.g., Gonsalves and Paller, 2000; Kensinger and Schacter, 2005; Foley et al., 2015; Cooper et al., 2016), lending considerable support to Durso and Johnson's (1980) explanation. The following experiment used divided attention to better understand the conditions at encoding the drive the picture misattribution effect.

Chapter Three: Effects of Divided Attention on Memory Distortions

3.1 Rationale

3.1.1 Manipulating attention during encoding

The present experiment was designed to test the hypothesis that divided attention at encoding could improve source memory by reducing picture misattributions. Theoretical motivation was borrowed from studies of visual attention that demonstrate that a cognitive system with limited resources selectively prioritizes processing of information on the basis of task demands (Carrasco, 2014). Thus, divided attention during encoding should reduce the resources available to generate mental images by virtue of devoting the limited amount of resources to answering the encoding question. Because more resources can be devoted to completing the encoding task, this position also predicts improved recognition memory for information studied with divided attention.

A competing prediction comes from literature demonstrating that divided attention interferes with Type 2 processes while keeping Type 1 processes intact (e.g., Jacoby, 1991; Fernandes and Moscovitch, 2000; Knott and Dewhurst, 2007). If image generation is a spontaneous (i.e., Type 1) process, then this line of evidence would predict that divided attention has no effect on reducing picture misattributions. Additionally, Johnson et al. (1993) raise the point that encoding manipulations that compromise the extent to which an event can be encoded into a unified memory representation should impair source memory accuracy. This follows from the fact that source is inferred on the basis of memory characteristics. Encoding information under conditions that result in weak traces, like stress or divided attention, will limit the specificity with which that information is represented by memory. Poorer representations limit the amount of information judgment processes have to infer source. Altogether, these

considerations lead to the prediction that both recognition and source memory accuracy should be impaired for information studied with divided attention.

Because I was interested in investigating spontaneous imagery, I chose to use the encoding question that has persisted in the post-Durso and Johnson (1980) literature: what do you do with this item? Participants incidentally encoded a series of items presented either as pictures or words. I used a within-subjects design to increase statistical power, such that participants completed half of the encoding phase with full attention and the other half with divided attention. To prevent confounds of double presentation, I used an auditory cue for recognition memory and asked participants to make a source memory response for items identified as old.

3.1.2 Self-report questionnaires

After the experiment, I asked participants to fill out a metamemory questionnaire. The purpose of this was to assess the relationship between extant beliefs about memory, task performance, and response bias (frequency of high confidence responding). Because Johnson et al. (1993) include metamemory beliefs among one of the factors influencing the criteria whereby judgment processes operate, this further addresses elements of the source monitoring framework that (to the best of my knowledge) have not been tested previously. Critically, these data should help adjudicate between the competing predictions I made about how overall metamemory beliefs contribute to decision making processes in the last chapter.[13]

[13] That it is equally plausible that participants with positive and negative metamemory beliefs will demonstrate lower criteria for endorsement. In this context, endorsement is operationalized as proportion of high confidence responding.

A post-study questionnaire asked participants to "explain the strategy you used to determine if an item was presented as a picture or word." This, together with some metamemory considerations, comprise the experimental philosophy portion of the project. A central question of this thesis is understanding the processes that secure the reliability of recollective simulations and, by extension, the beliefs formed on the basis of these simulations. Critical to answering this question is identifying the personal-level resources agents have for discriminating between reliable and unreliable episodic simulations. Although this is largely orthogonal to how they actually accomplish cognitive tasks (and forgetting this would leave me guilty of the phenomenological fallacy), it does seem that agents' beliefs in their cognitive capacities have some important bearing on their willingness to form a belief on the basis of that capacity. In my opinion, there has been a regrettable dearth of investigation into how agents think they accomplish cognitive tasks. I conducted a qualitative, exploratory analysis of their responses to begin addressing this empirical gap and to gain insight into the resources agents have for solving this particular source problem.

3.2 Experimental Methods

3.2.1 Participants

40 individuals participated in the study (19 females, 21 males). All participants were 18-23 years of age (mean = 19.23 years, SD = 1.35) and had no history of neurological or psychological disorders. Three additional individuals participated in the experiment but had to be excluded from analysis because of a computer error during the test phase. All participants are undergraduate psychology students at Boston College and were compensated with course credit. All participants gave informed consent before beginning the experiment, and all procedures were approved by the Institutional Review Board at Boston College.

3.2.2 Task

The task was split into 2 parts, incidental encoding and a memory test. At the beginning of the encoding phase, participants were told that the purpose of the experiment was to functionally categorize items. They were instructed to type out an answer to the question "What do you do with this item?" for 120 stimulus items. Half of the items were presented as grayscale clip-art images of objects and half were presented as words. Items were presented in the center of a white screen for 6 seconds with a 750 ms ISI, and the encoding question remained onscreen for the duration of stimulus presentation. Encoding was split into full (FA) and divided attention (DA) blocks (each containing 60 stimulus items, 50% words), and the order of these conditions was counterbalanced across participants. I used a tone discrimination task to divide attention. For this task, participants were instructed to press the '6' key whenever they heard the background tone switch. The tone was comprised of two 1500 ms sine waves (261 Hz and 329 Hz) alternating randomly (Mickley Steinmetz et al., 2014). The picture/word status of each stimulus item at encoding was randomly determined for each participant.

Participants were instructed on and completed a cued recognition test immediately following both encoding blocks. To avoid confounds of double presentation, retrieval cues were presented auditorily. The duration of each cue was <1s and was presented only once. This fact and the importance of paying attention to the cue were emphasized in the recognition task instructions. In response to each cue, participants indicated whether the named item was old or new (i.e., whether it had just been categorized or not) and their confidence in that response (1-4; definitely old, maybe old, maybe new, definitely new). For items they identified as old, they were then asked to indicate whether they saw it as a picture or a word and their confidence in that response (1-4; definitely picture, maybe picture, maybe word, definitely word). Participants had 5 seconds to make each response, with a 750 ms ISI between questions. The old/new status

of each tested item was also randomly determined. After the test phase, participants completed a version of the Multifactorial Memory Questionnaire (Troyer and Rich, 2018) adapted to the age demographic and a post-study questionnaire specific to this experiment.

3.2.3 Materials

Grayscale clip-art images were obtained from the authors of the Picture Perfect dataset (Saryazdi et al., 2018). I opted to use grayscale rather than colored images to keep the design as similar as possible to Durso and Johnson (1980). The colored versions of the images used have a picture-name agreement of 4.72/5.00 and a name agreement of 88%. Summary statistics for the stimulus items can be found in the Supplementary Materials. The individual sine tones were generated using https://www.audiocheck.net/audiofrequencysignalgenerator_sinetone.php and the background tones were created using Audacity. Retrieval cues were created using an online text-to-speech generator, http://tts.softgateon.net/. For over 60% of the cues generated from this site, I used Audacity to reduce the speaking speed in the interest of optimizing comprehension.

3.2.4 Data analysis

All statistical tests were conducted using the standard alpha = 0.05. For recognition and source memory, I calculated measures of accuracy, confidence and task-dependent metamemory. Recognition accuracy was computed as the proportion of correct old/new responses out of all test items, and source accuracy was computed as the proportion of correct source responses out of items correctly identified as old. I opted for this measure of memory accuracy over corrected recognition or d' because there was no perceptual similarity between study and test items, so recognition could only occur on the basis of semantic familiarity. Confidence was computed as the frequency with which participants used high confidence responses on recognition and source memory tests. Using Fisher's r to z transformation, task-dependent metamemory was computed

as the correlation between accuracy and confidence on a trial-by-trial basis. This measure allows me to quantify the correspondence between a participant's memory accuracy and their assessments thereof; i.e., their metacognitive sensitivity. I used an ANOVA to test the effects of encoding manipulations (Attention: FA, DA; Stimulus Type: Picture, Word) on memory accuracy and confidence. Because unequal amounts of participants failed to use all levels of confidence when making recognition and source responses, I used two-tailed t-tests to compare the effects of attention and stimulus type on subsequent recognition and source metamemory.

I used the "Feelings about Memory" dimension of Troyer and Rich's (2018) Multifactorial Metamemory Questionnaire (MMQ) to obtain a task-independent measure of participants' metamemory. The self-report nature of this measure permits operationalization of it as participants' general beliefs about their memory. I computed Pearson's correlation coefficients to quantify the relationship between MMQ scores and (i) memory accuracy, (ii) memory confidence, and (iii) task-dependent metamemory. Finally, I report the results of my exploratory investigation of the strategies participants reported using to solve the source problem.

3.3 Results

3.3.1 Memory accuracy

A 2 encoding condition (full, divided attention) x 2 stimulus type (picture, word) ANOVA on recognition accuracy scores revealed that divided attention impaired recognition accuracy, $F(1,39) = 19.90$, $p < .001$, and that this impairment was especially strong for items presented as words ($F(1,39) = 4.32$, $p = 0.044$). I found no effect of stimulus type on recognition accuracy, $F(1,39) = 0.17$, $p = 0.68$, which was generally good in all conditions (DA, picture: mean = 0.83, SEM = 0.02; FA, picture: mean = 0.87, SEM = 0.01; DA, word: mean = 0.80, SEM = 0.02; FA, word: mean = 0.89, SEM = 0.02).

The same 2x2 ANOVA was run on source accuracy scores. I found a strong picture superiority effect, ($F(1,39) = 51.28$, $p < 0.001$), and also a source memory benefit for items studied with full attention ($F(1,39) = 5.40$, $p = 0.025$). Contrary to my initial hypothesis, there was no interaction of encoding condition with stimulus type, ($F(1,39) = 0.43$, $p = 0.53$), indicating that dividing attention uniformly impairs source memory for items studied as pictures and words rather than selectively modulating picture misattributions. However, this is in line with predictions made by Johnson et al. (1993). Whereas recognition memory accuracy was between 80% and 89%, source memory accuracy exhibited considerably more variability between conditions (See Figure 3.1 for visual comparisons. DA, picture: mean = 0.95, SEM = 0.01; FA, picture: mean = 0.97, SEM = 0.01; DA, word: mean = 0.70, SEM = 0.04; FA, word: mean = 0.74, SEM = 0.03).

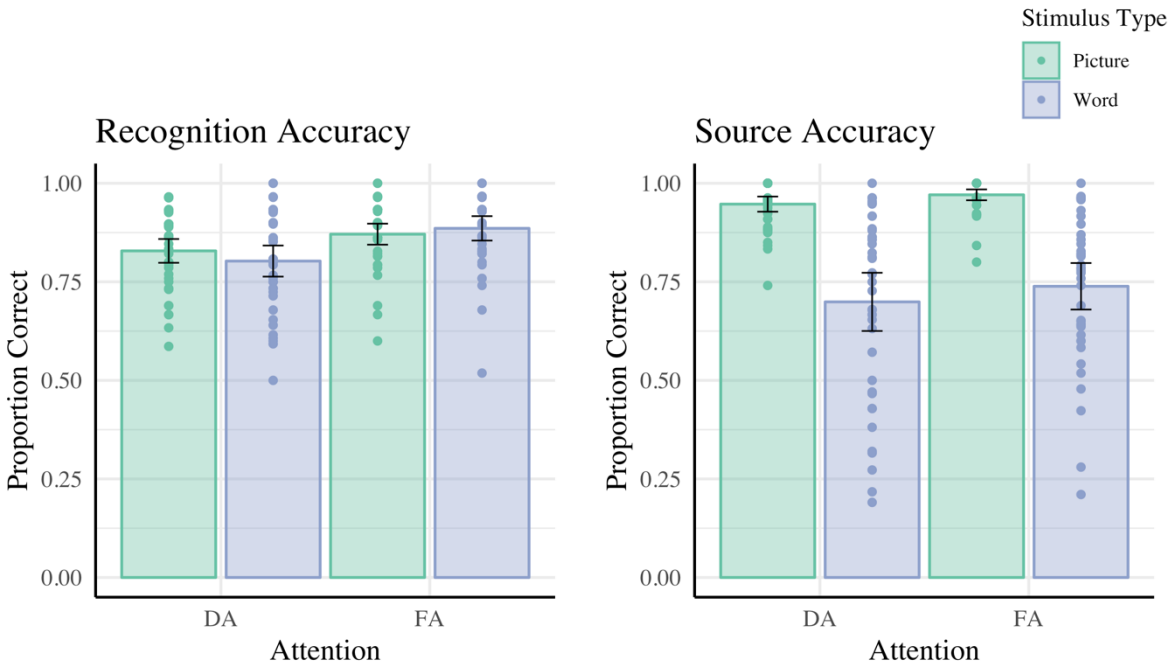


Figure 3.1: Memory accuracy. Error bars represent Mean ∓ 95% confidence interval. Points represent individual participant estimates. DA = divided attention, FA = full attention.

3.3.2 Memory confidence

Memory confidence was calculated as the proportion of high confidence responses each participant used when making memory judgments; i.e., participants responding only with 'definitely old' and 'definitely new' had a score of recognition confidence score of 1. This serves as a measure of the overall level of memory confidence which was modulated by encoding manipulations.

A 2 encoding condition x 2 stimulus type ANOVA on recognition confidence scores showed that divided attention reduced the frequency with which participants used high confidence memory ratings $F(1,39) = 11.298$, $p = 0.002$. There was no effect of stimulus type on confidence, $F(1,39) = 1.04$, $p = 0.31$, and no interaction, $F(1,39) = 0.85$, $p = 0.36$, indicating that attention condition at encoding was the only factor affecting subsequent recognition confidence. Overall, participants very frequently responded with high confidence (DA, picture: mean = 0.91, SEM = 0.02; FA, picture: mean = 0.95, SEM = 0.01; DA, word: mean = 0.90, SEM = 0.02; FA, word: mean = 0.95, SE = 0.01).

An identical ANOVA on source confidence scores revealed that participants used high confidence responses far more frequently for items studied as pictures ($F(1,39) = 53.36$, $p < 0.001$), and that they responded with high confidence source memory less frequently for items studied with divided attention ($F(1,39) = 13.42$, $p < 0.001$). Moreover, encoding condition interacted with stimulus type ($F(1,39) = 4.60$, $p = 0.038$) to disproportionately decrease the frequency of high confidence responses for pictures studied with divided attention. Confidence frequencies for source memory also demonstrated considerably more variability than confidence frequencies for recognition memory (See Figure 3.2 for visual comparisons. DA, picture: mean = 0.85, SEM = 0.02; FA, picture: mean = 0.92, SEM = 0.02; DA, word: mean = 0.70, SEM = 0.04; FA, word: mean = 0.72, SE = 0.03).

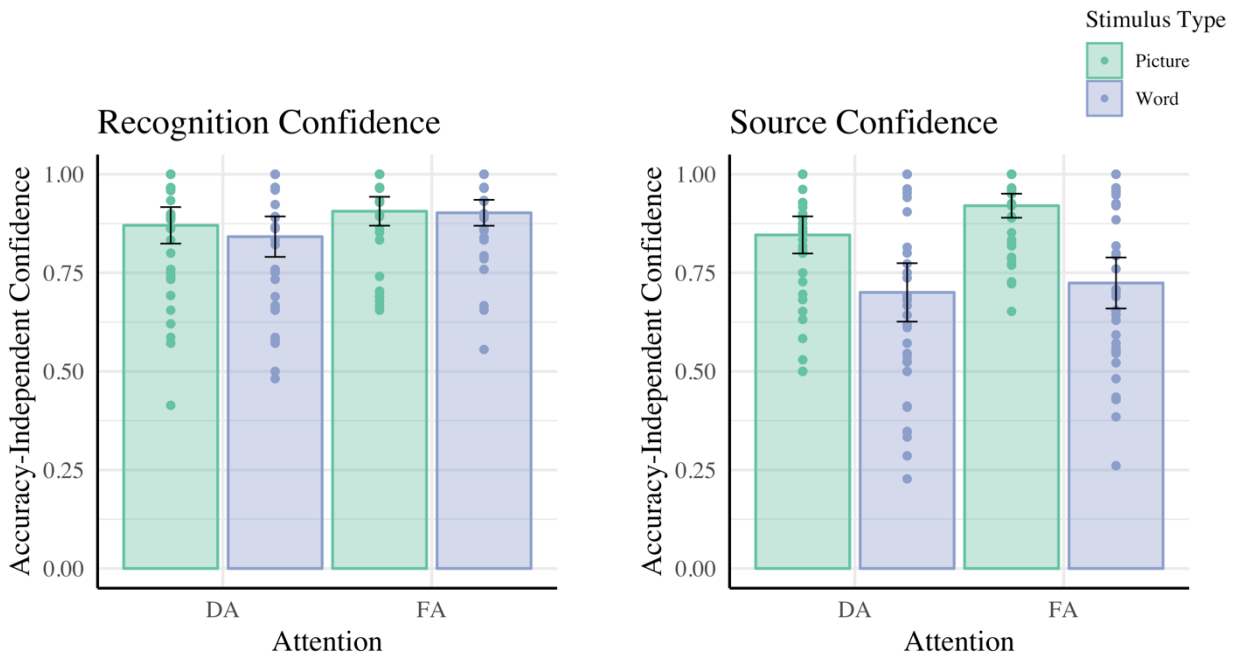


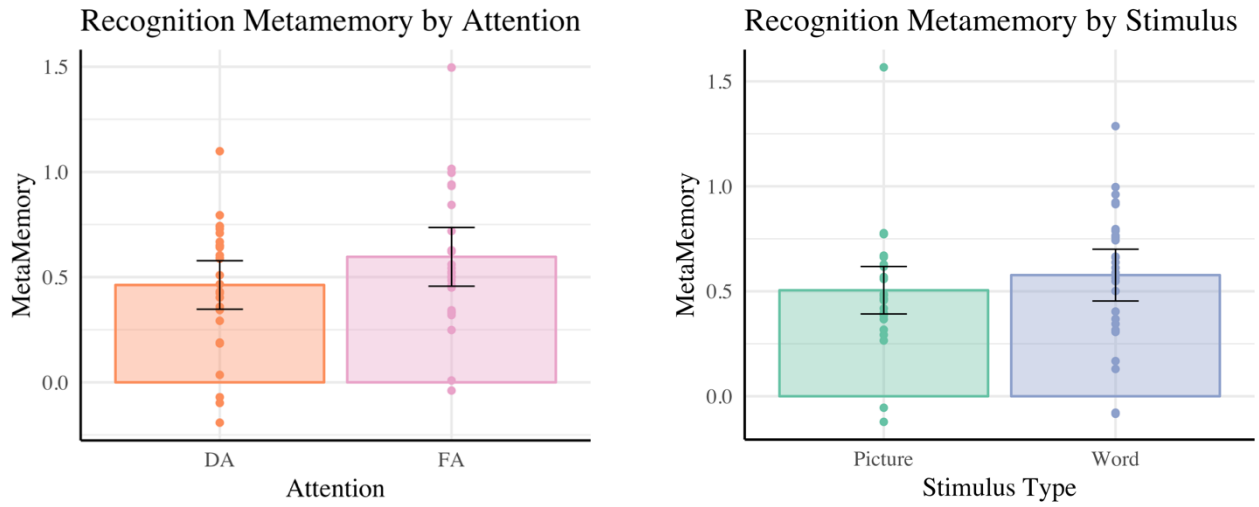
Figure 3.2: Accuracy-independent memory confidence, calculated as the proportion of memory responses made with high confidence. Error bars represent Mean ∓ 95% confidence interval. Points represent individual participant estimates. DA = divided attention, FA = full attention.

3.3.3 Metamemory

To measure metamemory, I used Fisher's r to z transformation to compute the correlation coefficient between recognition memory confidence and recognition memory accuracy on a trial-by-trial basis, within each participant. 13 participants who failed to use all 4 steps on the confidence scale had to be excluded from this analysis, leaving 27 participants. A t-test revealed that recognition metamemory was better for items studied with full attention than items studied with divided attention ($t(26) = -2.33$, $p = 0.028$). However, recognition metamemory was the same for items presented as pictures and words ($t(26) = -1.34$, $p = 0.19$; see Figure 3.3). This trend was reversed for source metamemory. After excluding 8 participants who failed to use all 4 steps on the confidence scale, I found no difference in source metamemory for items studied with

divided attention (t(31) = -0.38, p = 0.71), but significantly better source metamemory for pictures compared to words (t(23) = 3.31, p = 0.003).

(a)



(b)

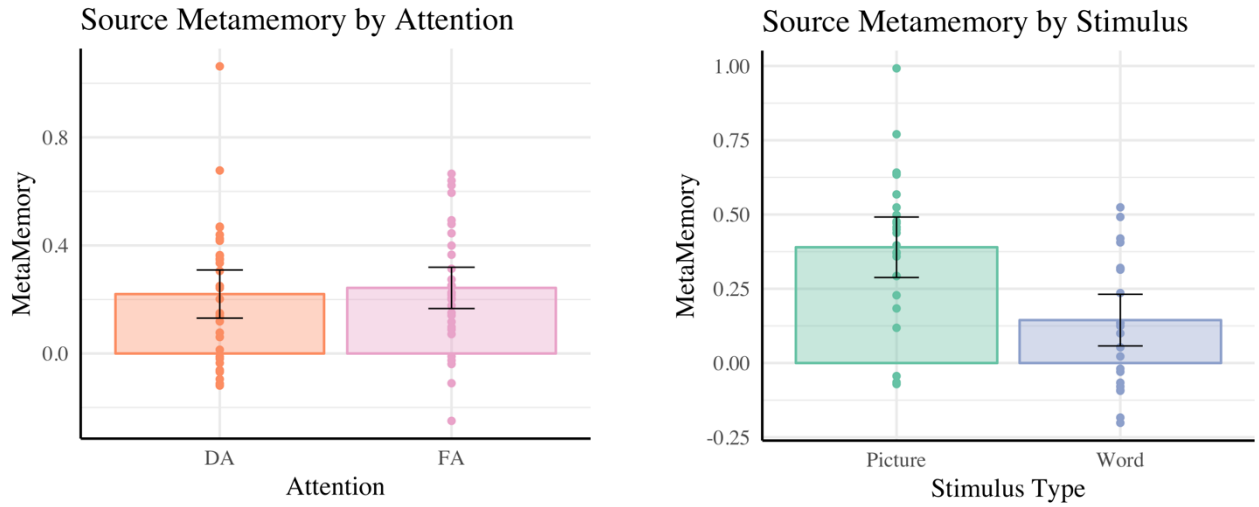


Figure 3.3: Metamemory, computed as Fisher's r-to-Z correlation between source accuracy and accuracy-independent confidence. Error bars represent Mean ∓ 95% confidence interval. Points represent individual participant estimates. (a) Recognition metamemory by attention and stimulus type. (b) Source metamemory by attention and stimulus type. DA = divided attention, FA = full attention.

3.4 Self-report results

3.4.1 Beliefs about memory fail to predict memory accuracy

Using Troyer and Rich's (2018) scoring keys, I computed a task-independent metamemory score for each participant. The range of possible scores on the modified version of the questionnaire is 0-60, with a score of 60 reflecting highly positive feelings about memory. Obtained scores had a mean = 38.25, SEM = 1.41, and ranged from 21-60. To understand the relationship between agents' beliefs about their memory and the reality of their memory abilities, I computed correlation coefficients between each participants' MMQ score and their overall recognition and source memory accuracy on the task. I found that beliefs about memory are a poor predictor of both recognition ($R = 0.29$, $p = 0.071$) and source ($R = 0.071$, $p = 0.66$) memory accuracy (See Figure 3.4).

Interestingly, there was no significant change in the relationship when separate coefficients were computed for DA and FA accuracy scores. Coefficients computed for average accuracy scores are reported for this reason. The direction of the difference between task-independent recognition metamemory and task-independent source metamemory is consistent with the direction of the difference between the corresponding task-dependent measures, though the task-dependent measures indicate a stronger consistency between beliefs about performance and objective performance accuracy.

3.4.2 Beliefs about memory predict likelihood of high confidence responding

Unsurprisingly, beliefs about memory were a good predictor of the frequency with which participants used high confidence memory responses. This relationship was equally strong for both recognition ($R = 0.44$, $p = 0.0047$) and source memory ($R = 0.50$, $p = 0.001$; See Figure 3.4). There was also a significant relationship between MMQ scores and a post-study report

about overall confidence in performance on the task (R = 0.48, p = 0.0018), further supporting the idea that beliefs about memory are a stronger determinant of assessments of memory performance than objective memory accuracy. This also lends support to the prediction that positive metamemory beliefs result in lower criteria for source monitoring judgments.

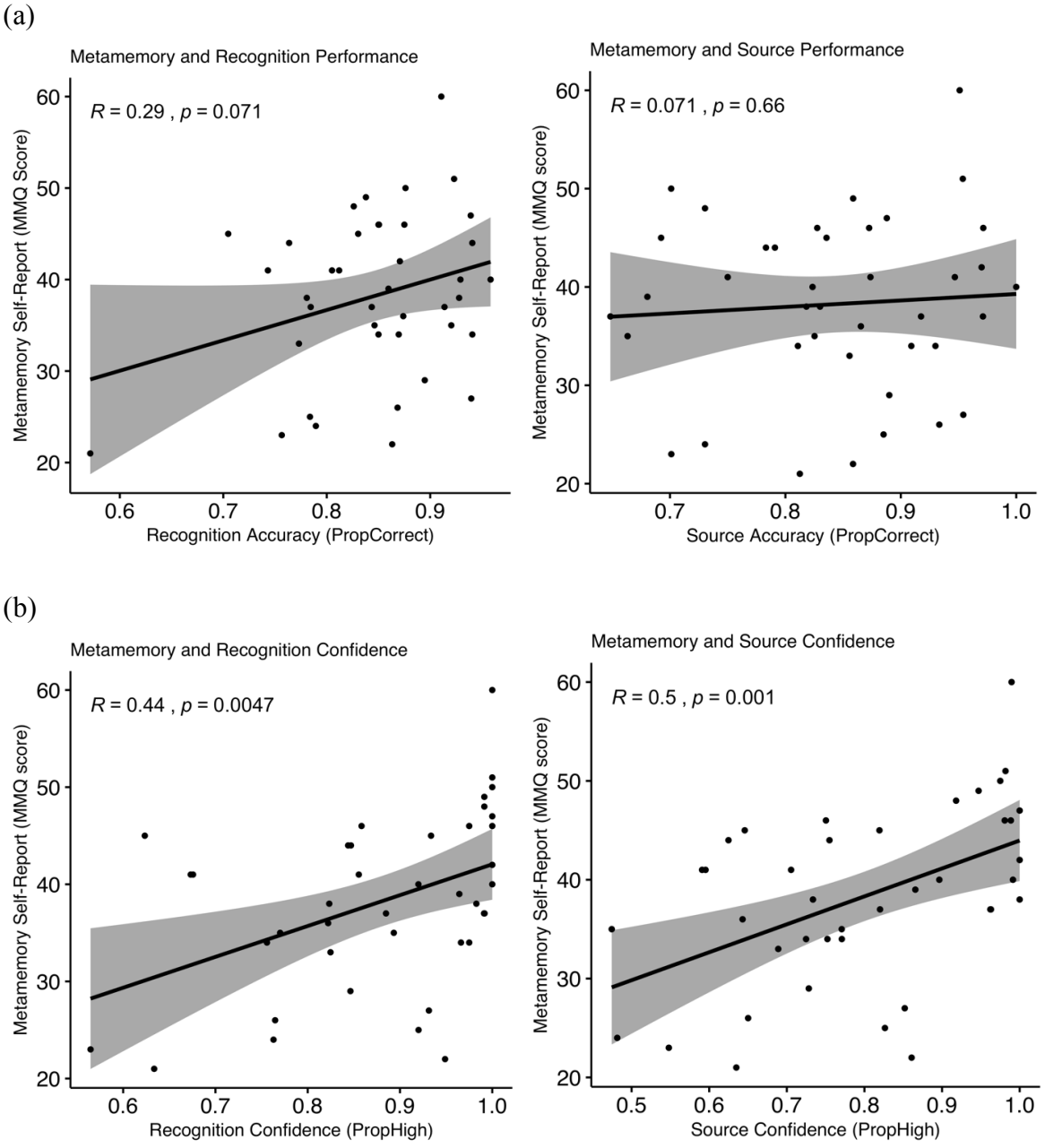


Figure 3.4: Self-report data. Gray regions represent 95% confidence interval. Points represent individual participant estimates. (a) Beliefs about memory fail to predict both recognition and source accuracy. (b) Beliefs about memory reliably predict accuracy-independent recognition and source memory confidence.

3.4.3 Source discrimination strategies

I judged that participants responses to the prompt "describe the strategy you used to determine if an item was presented as a picture or word" could be classified into 5 general categories: simple memory, mental time travel, conditional search, visualization, and encoding responses. A number of participants' responses used language which classified them for more than one of these categories, though I did not find that any response could be classified for more than 2 categories. A table containing the responses and my classifications, as well as the criteria a response had to meet to be included in category, can be found in the Supplementary Materials. Again, I would like emphasize that what follows should be viewed as qualitative, exploratory analyses. These classifications were not corroborated by a second rater and the language participants used was sometimes ambiguous. Readers are encouraged to consult the Supplementary Materials to corroborate the classifications themselves.

Of the 40 respondents, 6 used language that indicated they did not understand the question and 1 claimed to have had "no specific strategy". Evidence of misunderstanding the question consisted in responses aimed to explain how participants determined whether the item presented at encoding was presented as a picture or word, as the only information on screen during retrieval was the retrieval question, fixation cross, and old/new or picture/word confidence scale. The categories that the 33 remaining responses fell into can be found in Figure 3.5. "Simple memory" responses simply cited memory as the means through which they solved the source problem. "Mental time travel" responses claimed to use this capacity to solve the problem. "Conditional search" responses conducted a serial memory search, primarily anchoring decision responses in whether they could remember a picture at study to associate with the verbal cue at retrieval. "Visualization" responses attempted to visualize the item on the screen to solve

the item on the screen to solve the problem. And "encoding responses" tried to use idiosyncrasies of their responding patterns at encoding as cues for stimulus type.

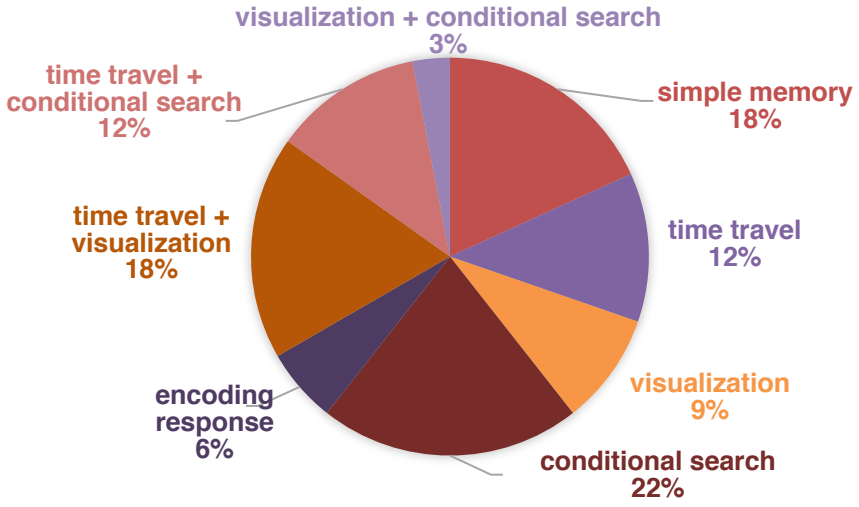


Figure 3.5: Source monitoring strategies. A full list of responses and categorizations can be found in the Supplementary Materials.

3.5 Discussion

This experiment tested how dividing attention during encoding modulates the accuracy of recognition and source memory, as well as the accuracy with which participants were able to monitor their own memory performance. Specifically, I was interested in testing the hypothesis that dividing attention inhibits imagery at encoding, the predominant explanation for how encoding may cause subsequent picture misattributions. I found that dividing attention decreased both recognition and source memory accuracy, consistent with Johnson et al.'s (1993) prediction that divided attention compromises the extent to which an experience can be encoded into a memory representation. Whereas this impairment was greatest in recognition memory for items presented as words, divided attention equally impaired source memory for items presented as

pictures and words. Thus, my results suggest that dividing attention at encoding weakens the strength of a memory representation rather than selectively inhibiting spontaneous imagery. Because dividing attention is known to interfere with controlled processes while leaving automatic ones intact, these results lend further support the position that the encoding process(es) contributing to picture misattributions are Type 1; i.e., participants are not deliberately generating images to answer the function question.

That participants less frequently used high confidence responses when making recognition and source memory judgments for items studied with divided attention supports the idea that this manipulation compromised the integrity of the memory representation. In the case of source memory judgments, this impairment was selective to items presented as pictures. This is consistent with the finding that the most commonly used strategy for solving the source problem in this experiment was to conduct a conditional memory search for a recently viewed image that corresponds to the item name coming out of the speaker. If divided attention compromises the strength of memory traces, then both picture and word representations for items studied with divided attention are less determinate than representations for items studied with full attention. This should increase the time needed for the system to construct a simulation of the experience that corresponds to the auditory cue, which should decrease the agent's confidence that the simulation is being generated by recollection. However, because picture representations contain more detail than word representations, their determinacy has more room to vary before becoming unreliable sources of information. I think this is the best explanation for the selective source confidence impairments for pictures.

Altogether, I found that incidental encoding of information with divided attention leads to decreased accuracy and confidence in recognition and source memory. Further, it impairs

subsequent monitoring accuracy for recognition memory but has no effect on source monitoring accuracy, which itself was considerably less accurate than recognition metamemory. This is consistent with an interpretation that divided attention impairs the extent to which an experience can be encoded into a unified memory representation. That all dimensions of memory were better for pictures than for words supports an interpretation that picture representations are more detailed than word representations. This gives the system more information with which it can construct a simulation, leading to more accurate and confident responses. However, the high amount of detail that can be included in a picture representation also increases the degree to which identifying characteristics of a picture can be encoded into a memory representation. That participants demonstrated selective impairments in confidence for pictures studied with divided attention supports this claim. The next chapter takes these findings, together with the philosophical and theoretical considerations of previous chapters, to lay the groundwork for an epistemology of episodic memory.

Chapter Four: Toward an Epistemology of Episodic Memory

4.1 What does episodic memory contribute to knowledge?

In order to conduct structured investigation into the necessary and sufficient conditions for knowledge, epistemologists have undertaken many analyses of the concept of knowledge in the attempt at creating formulas that indicate its presence. The most famous is known as the tripartite analysis of knowledge, which claims *S* knows *p* iff:

i. *p* is true;

ii. *S* believes that *p*;

iii. *S* is justified in believing that *p*.

This is often abbreviated as the "JTB" analysis, which holds that justified, true belief is necessary and sufficient for knowledge. Investigation into what determines the truth of *p,* what it means for *S* to believe that *p*, and what it takes for that belief to be justified can then ensue. Edmund Gettier (1963) famously formulated two counterexamples that demonstrated the insufficiency of the JTB analysis to give an account of knowledge. One of them goes like this:

> Smith and Jones are interviewing for a job. While they wait, Jones takes the coins out of his pocket and he and Smith count there to be 10 coins, which Jones puts back into his pocket. After the interview, the interviewer tells Smith that Jones is going to get the job. Thus, Smith has strong evidence to believe the conjunctive proposition (a) that Jones is the man who will get the job, and that Jones has ten coins in his pocket. Smith sees that (a) implies the proposition (b) that the man who will get the job has ten coins in his pocket, and forms this justified, true belief. However, suppose that the manager overrides the interviewer and chooses Smith for the job, and that, unbeknownst to him, Smith also had 10 coins in his pocket. Does Smith still know (b)?

Clearly, it seems that Smith does not know (b) because (b) is only true in virtue of the amount of coins that are in Smith's pocket, and Smith's knowledge of (b) rests on his belief in the number of coins in Jones' pocket; i.e., Smith has a justified, true belief that (b) but he does not *know* that (b). Since this demonstration, most epistemologists accept the left-right biconditional (knowledge requires justified true belief) but reject the right conditional (justified true belief is knowledge).

This chapter aims to incorporate all that I have established regarding the nature and function of episodic memory to begin laying some foundations for an epistemology of episodic memory; i.e., what is the appropriate way to formulate the role that episodic memory plays in our having knowledge? I want to argue that episodic memory is a basic and generative epistemic source, similar to perception, and that agents form beliefs about their personal past and possible future on the basis of simulations constructed by the hypothetical thought system. This is a significant departure from traditional theories of memorial justification, which, assuming the storehouse model, posit that memory (1) preserves beliefs and (2) provides justification for believing at a later date. However, this may not be as radical of a position as it may seem if we consider the fact that most epistemology of memory has focused exclusively on information that is processed and transferred by the semantic memory system, which serves a markedly different function that is accompanied by an equally different memorial experience.

The focus of this chapter is on justification. Developing an account of how episodic memory justifies belief raises at least three problems that I will address in turn. The first is a central question in theories of justification: internalism or externalism? Ultimately, I will argue that the multi-step nature with which episodic memory functions to justify belief warrants an externalist-internalist hybrid theory of justification. Because the externalist dimension invokes

process reliabilism, I will first handle two problems that this position faces: the problem of generality and the threshold problem. After that, I present an argument for episodic memory as a basic epistemic source both for beliefs and justification. Fully defending that position requires elucidating the internalist dimension of my theory of justification.

4.2 Externalism about Justification

The central question dividing internalists and externalists about justification is whether an agent needs to have some perspective on the J-factors (or justifiers) of a belief. Internalists maintain that this is a necessary condition for having a justified belief – i.e., if you cannot articulate the justification for your belief, it is not justified. Externalists reject the necessity of this condition, arguing that beliefs can be justified via J-factors that are not accessible to an agent. A popular variant of this position is process reliabilism, which argues that a belief is justified if it is produced by a reliable belief-formation process (Goldman, 1993). Because this position does not require that an agent have perspective on the belief-formation process, it is in line with externalism about justification.

4.2.1 The problem of generality

Process reliabilism is challenged by the problem of generality: at what level does a process need to be reliable? This is a problem at both the personal and sub-personal levels. At the sub-personal level, the problem consists in questions like do the enzymatic processes facilitating neurotransmitter synthesis need to be reliable? Or is it more important that circuit-level interactions are reliable? At the personal level, the problem arises from the fact that "any belief-forming process token falls under indefinitely many process types" (Michaelian, 2016; p. 48). That is, when I form a belief about what I did last week, I do so by relying on memory, episodic memory, episodic memory in a cognitively normal adult, episodic memory in a cognitively

normal adult on a Tuesday, etc.[14] Michaelian (2016) suggests that a naturalist epistemology, which is best understood as a chapter of cognitive science that is concerned with "natural kinds of belief-forming processes," (p. 48) can inherit the individuation of process levels from cognitive science. By this he means that epistemologists can say that they are addressing episodic memory in cognitively normal adults in the same way that cognitive scientists individuate populations to avoid such unnatural process types as "episodic memory in a cognitively normal adult on a Tuesday." On the basis of the commitments outlined in Chapter 1, it should be clear that this is an approach I wholeheartedly support.

Even after adopting process individuations from cognitive science, we still face the problem of generality. Do the sub-personal or personal level processes need to be reliable? I want to argue that a necessary condition for the justification of beliefs formed on the basis of episodic simulations is that the simulations are generated by a reliable process. There are two ways that epistemologists have understood reliability: in the sense of frequency (how regularly the process produces true beliefs in the actual world) and in the sense of propensity (its ability to generate true beliefs in nearby possible worlds). Because of the commitments outlined in Chapter 1, I am going to operate on the frequency formula of reliability to argue that the probabilistic nature with which the episodic hypothetical thought system constructs simulations gives agents *prima facie*[15] justification for forming beliefs on the basis of episodic simulation. This follows

[14] Example adapted from Michaelian (2016). This may seem like a nonsense problem to a scientist. But it is one of many problems that arise from the nature of thought experiments in philosophy, as described in Chapter 1, and I would be remiss to write about process reliabilism without mentioning this central issue that it faces.

[15] *Prima facie* = "on the face of it." This is a weak kind of justification that can be defeated on the basis of new information.

directly from the fact that optimal reconstructions are *statistically* optimal by virtue of their reliance on schemas, which are literally constructed on the basis of statistical regularity. In this way, the first step of simulation construction is a reliable process. The second step, binding, unifies the disparate reconstructions and increases the disposition these representations have to trigger each other's reconstruction. This further contributes to the statistical reliability of reconstruction, warranting a claim that sub-personal reconstruction processes are reliable. Finally, the fact that humans can effortlessly use their memory to communicate with other humans suggests that it is a process that operates reliably across tokens of this process type.

4.2.2 The threshold problem

This naturally leads me to the threshold problem: how statistically reliable does a process need to be to classify as a reliable epistemic source? Goldman and Beddor (2016) hint at this problem when discussing statistical reliability, but answer it simply by saying that statistical reliability should be significantly greater than 0.5 but less than 1.0. Because this can be seen as a variant of the generality problem for frequency reliabilists, I think that taking the same approach suggested by Michaelian (2016) in response to the generality problem will give us a satisfactory answer. In cognitive science, accuracy levels between 0.80 and 0.90 are generally interpreted as evidence of a properly functioning memory system that is being probed by a sufficiently challenging task. If participants demonstrate accuracy considerably higher than 0.90, then the task is probably too easy. If their accuracy is considerably below 0.80, then the task is probably too hard (or your manipulation is working).

My data on recognition and source accuracy presented in the previous chapter lend support to this claim. Recognition accuracy varied between 0.80 and 0.89. Source accuracy varied between 0.70 and 0.97. If we average across conditions, however, source accuracy was

0.84. I think that these numbers nicely highlight the critical caveat of frequency formulas of reliability: they speak to the ratio of true beliefs to total beliefs formed by the system. This necessarily means averaging across conditions where accuracy is very high and conditions where accuracy is very low. But because I am not doing armchair philosophy, my theory of reliability does not need to be compromised by this theoretical smear (i.e., that the overall reliability of memory is a composite of different circumstances where it functions more or less accurately). Rather, by building on a host of previous investigation, I have identified specific encoding conditions that compromise the accuracy with which beliefs can be formed on the basis of episodic memory. On the assumption that decision processes in an experimental setting are relatively fixed, this highlights another caveat to the process reliabilist formula: the accuracy of reconstruction is contingent on encoding conditions.

4.2.3 The accuracy of episodic simulations is graded

Importantly, the accuracy of an optimal reconstruction is not all-or-none. The simulation it generates could contain wholly accurate details (e.g., morning breath, feelings of disgust) but fail to convey any important information associated with those details (e.g., what that person was saying). The converse is also true, as noted in Chapter 2. A simulation can convey important information that was communicated during the encoding event but fail to specify how that information was acquired. Further, it could specify only a portion of how that information was acquired (e.g. "Maggie" vs. "Maggie by telephone"). If we assume optimal consolidation, then the specificity of a memory representation, which is directly linked to the accuracy of optimal reconstruction, is a function of encoding conditions. Specifically, encoding conditions that impair (i) the extent to which attention can be devoted to relevant information, (ii) the extent to which an event can be consolidated as a unified representation, or (iii) the extent to which

information can be processed will manifest as impoverished memory representations. Rather than cutting its losses, the episodic construction system supplements missing information with schematic knowledge. This increases the system's chances of the agent forming beliefs and acting on its basis, which allowed it to persist over time.

Is the contingency of reconstruction on encoding conditions a sufficient reason to doubt the reliability of the reconstructive process? I am inclined to argue that it is not, on the basis that the schemas it uses to compensate for this contingency are statistically probable in nature. Thus, agents retain *prima facie* justification for forming beliefs on the basis of episodic simulations. However, it does seem that agents who could notice differences between accurate and less-than-accurate simulations should have an advantage over agents who lack this capacity. As pointed out in footnote [8], agents do not have the ability to directly compare different simulations to each other. This means (i) the discrimination process must be learned over time and (ii) it has to occur simultaneously or immediately after the experience of a simulation.

In fact, it seems that assessing the accuracy of a reconstruction should have its origins in the process problem. Simulation processes are differentiated by the constraints placed on the content to be included in a simulation. Whereas there are few constraints on the content to be included in hypotheticals, the content contained in an optimal reconstructions is importantly constrained by accuracy. That this cognitive architecture is already in place suggests that agents do have the capacity to assess the accuracy of a simulation. Thus, it seems that an adequate account of how beliefs are justified by episodic memory requires an appeal to internalist factors of justification. I will return to this shortly.

4.3 Generativity and belief formation

4.3.1 How does episodic memory generate beliefs?

As an example of how engrained the functional commitment of the storehouse model is in epistemology of memory, here is a quotation from Robert Audi's (2011) *Epistemology,* which is a textbook on the topic:

> Memory is a source of beliefs in the way a storehouse is a source of what has been put there, but it is not a source of beliefs in the generative way perception is. Clearly our memory, as a mental capacity, is a source of beliefs in the sense that it *preserves* them and enables us to *call them up*. We do this when we solve mathematical problems using memorized theorems. We may also be guided by other kinds of presupposed premises without having to call them to mind. Remembered propositions (and patterns) can be like routes we know well: they can guide our journey while we concentrate on the road just ahead. (p. 75, his emphasis)

I like this quotation because it highlights not only the pervasiveness of the storehouse model, but also the fact that epistemologists of memory are concerned almost exclusively with the semantic memory system.[16] Michaelian (e.g., 2011; 2012; 2013; 2016) has been making good progress on epistemological problems facing our empirically-grounded understanding of episodic memory, but he is relatively alone in this undertaking – most philosophers only go as far as treating the metaphysical problems episodic memory faces. In any case, this excerpt from Audi represents the default view in epistemology of memory: memory is a basic, generative source for justification (i.e., a belief you hold can be justified on the basis of memory), but not for beliefs.

---

16 I think my favorite example comes from Senor's (2014) Stanford Encyclopedia Entry on "Epistemological Problems of Memory." After highlighting the fact that a separate kind of event-based memory exists, he writes "Although it is clear that this distinction is real and significant, I don't propose to have too much more to say about it here. Since epistemology is primarily the study of knowledge and rational belief, and since knowledge and belief are propositional in nature, we'll here limit our focus to propositional memory." I'm sure you can imagine my shock upon encountering this while putting together my thesis proposal!

I think that the episodic hypothetical thought theory challenges this formula. I suggest that episodic simulations are basic and generative epistemic sources, and that optimal reconstructions are also basic and generative sources of justification. When I say that a simulation is a basic epistemic source, I mean that it generates beliefs without the dependence or contribution of another source (Audi, 2011); that is, memory generates content without drawing on other cognitive systems. This is different from saying that memory is a basic source of justification, which means that a belief can be justified by virtue of being retrieved from memory. Lackey (2005) was the first to suggest that memory can be generative. She argued that because information can be stored in memory without an agent initially forming a belief with that content, retrieving that memory creates a belief with that content which, assuming her memory is reliable, is justified. This is a marked departure from the default position that memory is only a belief-dependent process; i.e., it takes a belief as input and then outputs the same belief at a later time. If no belief is inputted into the 'storehouse,' but it later outputs a belief with the content X, and there is no evidence for memory malfunction, then memory can be said to have generated the belief *that X;* i.e., memory can be a belief-independent process as well.

This shows that memory can generate not only beliefs but also justification on the basis of previously stored content, if we assume a reliabilist account of justification. Additionally, it paints a plausible picture for how memory may function as a belief-independent process. As I have shown, however, simulation theories (e.g., De Brigard, 2014; Michaelian, 2016) push the belief-independency of memory even further by arguing that properly functioning memory systems regularly generate content that was not even present during an encoding event. And, by

virtue of the schema-consistency with which content is generated, the belief in that content being present at encoding is largely accurate.[17]

4.3.2 How does episodic memory justify beliefs?

However, it is important to keep in mind that the content of an optimal reconstructions can be more or less accurate and specific with respect to the initial encoding event. On this premise, Michaelian (2016) argues that "the standard, binary notion of belief as an all-or-nothing state – either an agent endorses a proposition or he does not – is not straightforwardly applicable to episodic memory… because the agent may have differing levels of confidence with respect to different aspects of a retrieved episodic representation" (p. 54). He continues to point out that this greatly complicates the project of assessing the reliability of episodic memory, and that we can simplify it by concerning ourselves primarily with episodes that agents wholly reject or wholly accept, while keeping in mind that this is not the whole story. Because confidence is created by metacognitive monitoring processes, it is a definitionally internal factor. I want to argue that the fact that agents, at the personal level, always have varying degrees of confidence in both the accuracy of a simulation and the process that generates it necessitates including an internal dimension to the account of how beliefs formed on the basis of episodic simulations are justified.

If the account of monitoring given in Chapter 2 is roughly correct, then Type 1 process and source monitoring occurs every time a simulation is constructed. These automatic monitoring processes signal to agents both the process generating a simulation and the conditions

[17] Again, this is a consequence of the fact that schemas represent statistical regularities in the agent's environment. Laboratory experiments, in this sense, are highly unusual settings that generate conditions specifically designed to induce inaccuracies in generation of episodic content.

under which the information contained therein was acquired. That is, it is never the case that agents are wholly unaware whether they are remembering or imagining. And even if it is the case that daydreaming generates spontaneous simulations, there are a number of different markers agents can use to identify the process generating a simulation and, in the case of hypotheticals, which elements of the simulation belong to the personal past and which have been fabricated by flexible recombination.

Whereas it is difficult to conceive of a situation where a neurocognitively normal agent was utterly incapable of determining whether a simulation was generated by remembering or imagining, there are a number of cases when a memory representation is largely ambiguous with respect to source. Further, there are cases when source is completely indeterminate. When Type 1 monitoring processes are incapable of inferring source, and this is relatively important for completing the task at the agent's hand, then they will recruit the agent to undertake her own Type 2 investigation into plausible sources. This includes reflecting on the content and quality of a simulation, considering which elements of it cohere with semantic knowledge of the personal past, and assessing how these contents fit your schemas for the information represented. There is hardly a more internally justified belief than one which is formed on the basis of Type 2 source monitoring.

As I have pointed out, though, it is not the case that all endorse/reject decisions are made on the basis of such deliberate metacognition. Most of the time, these decisions are made on the basis of feelings that are produced by sub-personal metacognitive processes. However, as theorized in Chapter 2 and subsequently shown in Chapter 3, metamemory beliefs have an important bearing on how willing an agent is to endorse or reject a particular simulation. Even though I found no relationship between metamemory beliefs and memory accuracy,

metamemory beliefs were a strong predictor of high confidence responding for both recognition and source memory. This was true even for source memory, despite the fact that participants failed to demonstrate reliable task-dependent monitoring for this dimension of their memory (see Figure 3.3; the strongest correlation between confidence and accuracy was around 0.30 for source, notwithstanding whether the calculations were made on the basis of encoding condition or stimulus type).

I think this nicely highlights Michaelian's point that agents can be confident in certain dimensions of a simulation while being less confident in others. That participants were reliably able to endorse or reject the belief that an item was old on the basis of their confidence in that belief, they were less reliably able to do so when assessing specific details about that event representation. Because they were forced to make a response for each test item, it is difficult to use these data to make broad claims about the reliability of agent-level endorsement processes, as participants may have been compelled to report a belief when they would have naturally withheld belief-formation. Further, making an old/new judgment can recruit resources beyond those afforded by the episodic construction system (e.g., cortex can facilitate familiarity judgments), and these may have given agents the metacognitive advantage in task-dependent recognition monitoring.

4.3.3 How can justification be defeated?

Because I am arguing that agents are *prima facie* justified in forming beliefs on the basis of episodic simulations, identifying factors that contribute to how confidently agents endorse or reject the content of a simulation can be seen as identifying defeaters for that *prima facie* justification. Epistemologists distinguish between undermining and rebutting defeaters. Undermining defeaters challenge your reasons to believe *p* whereas rebutting defeaters refute the

truth of *p*. Low source confidence and ambiguous process signals are undermining defeaters for endorsing the contents of a simulation. Noticing an incoherence between the content of a simulation and existing belief (that you are confident is accurate) via Type 2 monitoring would be an undermining defeater for endorsing the content of a simulation. In addition to these internal defeaters, there are also external defeaters to *prima facie* justification. Undermining external defeaters largely consist in neurocognitive abnormalities, such as traumatic brain injury or severe mental illnesses. For example, if an agent suffered head trauma that damaged part of his hypothetical thought system, this would be an external defeater of his *prima facie* justification for believing the content of a simulation. Rebutting external defeaters can be testimony from another agent (assuming no misleading intent) or digital media. For example, if I experience a simulation of myself dancing with my friend at a party that I think is being generated by memory but then see the text messages from her apologizing for not coming, then this defeats my justification for forming the belief that she was at the party.

In summary, the episodic hypothetical thought theory of human remembering motivates a revision to traditional formulas of memorial justification. Rather than simply functioning to preserve previous justification, episodic memory functions as a basic and generative source of *both* justification and belief. The beliefs that are formed on the basis of episodic simulations are ultimately justified by virtue of the reliable processes that generate them. However, because sub-personal metacognitive processes have a "crossover" mode of operation, they automatically generate feelings that are critical to the beliefs that agents form on the basis of the simulation. Finally, because the accuracy of episodic simulations is not all-or-none, the beliefs formed on this basis do not necessarily need to be all-or-none. Agents can endorse particular dimensions of a simulation while rejecting or withholding belief about others.

Conclusion

> *So that imagination and memory are but one thing, which for diverse considerations hath diverse names*.
>
> — Thomas Hobbes, Leviathan 1.2

I have argued that empirical investigation of human memory speaks against the commonsense conception of memory as some kind of "storehouse" that preserves records of previous experiences and acquired beliefs. Early evidence against this intuition were demonstrations of false and extremely distorted memories, prompting psychologists to defend a position where memory preserves the past by reconstructing previous experience. More recent evidence demonstrating substantial overlap of the neurocognitive systems involved in remembering and hypothetical thinking has motivated researchers to suggest that remembering is most properly understood as a subprocess of a larger system that functions to generate episodic hypothetical thoughts. Coupling this revised functional understanding of memory with its reconstructive nature casts substantial doubt on its reliability as a justifier of knowledge.

However, there are at least two metacognitive monitoring processes that give agents some perspective on both (1) how a simulation is being generated and (2) how accurate or plausible that simulation is. These are process and source monitoring, respectively, and they function by evaluating characteristics of the representation generating a simulation to infer how it is being generated and/or where the content comes from. This happens in a Type 1 manner virtually every time a simulation is constructed. Importantly, these monitoring processes have a crossover mode of operation that allows them, at the sub-personal level, to communicate their

interpretation to the agent that occupies the personal level. This communication is experienced by the agent as feelings — feelings of pastness, of familiarity, of confidence, etc. It is on the basis of these feelings that agents form and express beliefs about their personal past.

Because these metacognitive monitoring systems operate on the basis of average differences between kinds of representations, one can manipulate characteristics of a representation such that it is interpreted incorrectly by the monitoring system. A well-documented case of one such misinterpretation are picture misattributions, which occur when agents mistakenly remember seeing information presented as a picture even though they actually saw it presented as a word. The standard explanation for this phenomenon is that semantic processing, which the effect is contingent on, results in spontaneous imagery that interferes with accurate encoding of the perceptual experience. I used divided attention to test this explanation and found no effect on picture misattributions, supporting an explanation that the encoding factors contributing to picture misattributions are Type 1 in nature. I also found that divided attention selectively impaired confidence for source judgments about stimuli presented as pictures, though it had no effect on participants' ability to monitor their source performance. It did, however, decrease recognition metamemory accuracy.

I argued that the aforementioned considerations motivate positing episodic memory as a basic and generative source of *both* knowledge and justification, contrary to the general consensus in epistemology that memory is only a basic source of justification. This follows directly from the fact that perception and recollection operate via schema-tagging, a point that was made in the context of a broader argument that false and distorted memories do not represent memory malfunction in any epistemically important sense. That they are labeled false and distorted is an artifact of the functional commitment of the storehouse model: that memory

functions to preserve the past. Even with episodic memory metaphysically situated as a subprocess of a hypothetical thought system, agents have the ability to discriminate between simulations of actual or hypothetical events as well as the ability to assess the accuracy with which a simulation is constructed. These manifest as feelings that can serve as defeaters for the *prima facie* justification agents have to form beliefs on the basis of reliably (re)constructed episodic simulations. Because the accuracy of an episodic memory can vary with respect to information processed during the encoding event, agents can choose to endorse certain aspects of a simulation while withholding belief in or rejecting others.

***

Ultimately, this was a thesis interested in investigating the nature of our mental states and the knowledge we can have of them. Of primary interest were mismatches between how we think our mental states function and what cognitive science tells us about how they function. I think these cases are especially interesting because they highlight the fact that we can have false beliefs about how our minds work, despite our intuition that we are the ultimate authority about what goes on in our minds (which itself is a false belief). The orthogonality of phenomenology to mental state function raises a number of theoretical and practical implications. Perhaps the most pressing theoretical concern is whether this means that we have no choice but to accept total skepticism about mental states and how they inform us about the external world. I have tried to show that the answer to this question is absolutely not — cognitive systems have evolved a number of hacks that give agents (1) some control over the manipulation of mental states and (2) some perspective on how these states are instantiated. This perspective need not map neatly onto how the states are *actually* instantiated; it merely needs to inform agents about their mental states in whatever way is sufficient for communicating the "big picture" message. Adopting an

evolutionary perspective largely dissolves concerns of external world skepticism. Mental states evolved *in response to* selection pressure from the environment. That they have persisted this long is evidence of their correspondence to conditions in the environment.

Practically, this raises the question of how agents can know when to endorse or reject belief in the contents of a simulation on the basis of phenomenology. I have tried to show that feelings of pastness and familiarity are reliable markers of the pastness of a simulation or familiarity of perceptual input and, as such, can be justifiably used to form a belief that the content of a simulation or perception has been processed by the system at a previous point in time. However, if there is a case where a feeling of familiarity is particularly faint, agents should use their discretion in endorsing belief in the pastness of content. What tools do they have for exercising discretion? Type 2 source monitoring, corroboration with semantic knowledge or other episodic memories that they are more sure of than this one, testimony from external sources, etc. Further, if they can remember something about their mental states surrounding and during the to-be-remembered event, this can aid in the decision process. Possessing scientific knowledge of encoding conditions that compromise the integrity of a memory representation is a huge asset to the decision process, as this allows agents to formulate a more objective explanation for vague feelings of familiarity.

Thus, we can see that even though we have false beliefs about how mental states function, we are still able to form justified beliefs about them and the content they represent, at least in the case of episodic memory. And though the reliability of episodic memory varies from person to person, and the accuracy of different memories can vary tremendously within an individual, its evolutionary design secures its status as a reliable source of knowledge about the personal past.

Supplementary Materials

S.1 Stimuli statistics

I opted to use Saryazdi et al.'s (2018) Picture Perfect stimulus set after learning that the Snodgrass and Vanderwurst (1980) dataset used by Durso and Johnson (1980) was no longer freely available. Saryazdi et al. (2018) took high-resolution photographs of everyday items and converted them to clipart. Then, they collected normed data both in lab and online for both stimulus sets and posted everything in a Google Drive folder accessible to the public. Saryazdi generously emailed me line-drawing and grayscale versions of the clipart photographs as well. I opted to use the grayscale clipart images rather than line drawings after a number of different individuals in lab had difficulty coming up with the proper name for the item rendered in the line drawing, as picture-word agreement was a critical factor in the recognition test.

The following data were obtained from the normed data included in the Picture Perfect set. Importantly, ratings were made on colored versions of the clipart images I presented in grayscale. However, I chose the images keeping this in mind and thus did not include any items whose corresponding image required color for discrimination (e.g., I did not include "tomato"). 160 items were randomly assigned old/new, picture/word, and FA/DA presentation status. They had an average familiarity rating of 3.57/5.00 (SD = 0.63) and a picture-word agreement rating of 4.72/5.00 (SD = 0.17), and an average image agreement (i.e., how closely the clipart matches the mental image participants generate when thinking of the item) of 4.16/5.00 (SD = 0.53). 88.36% of participants generated the same name for a target item, and the variability in their responses (measured as an *H* value) was 0.7. 56.27% of participants generated the same verb for a target item, and those responses had an *H* value of 1.72.

S.2 Self-report classification scheme

S.2.1 Simple memory

These responses made no explicit reference to any other cognitive process other than remembering; e.g. “memory” or “I remembered some of them vividly as either being a picture or a word.” A number of the remaining responses make use of memory-related verbs (i.e., “recall,” “remember,” etc.) but did so in conjunction with other mental operations which distinguish these responses from those which only used memory-related verbs to describe their strategy.

S.2.2 Mental time travel

In order to classify for the mental time travel category, responses either had to include “think(ing) back” or some reference to visual information presented during encoding. The response “I tried to think back to when I was describing the picture/word and see if an image or a word came back to me” exemplifies the first criterion and the response “I just thought really hard about what I saw on the screen during part 1” exemplifies the second. 4 responses classified for the mental time travel category only. 6 responses used a combination of mental time travel and visualization strategies, and 4 used a combination of mental time travel and conditional search strategies.

S.2.3 Visualization

To classify as using a visualization strategy, a response had to include language referencing generation of a visual image. 3 responses classified for the visualization category only, with the most involved of these being “based on mental image memory in my head and associating the item either with the written word or drawn image.” 2 responses explicitly cited imagination processes as responsible for constructing their visualization, and both of these qualified for both the visualization and time travel categories. A total of 6 responses classified

for both of these categories, and the response "I visualized what I saw on the screen and chose based off of what I remembered" is a good example of this overlap. Only 1 response classified for both visualization and conditional search strategies.

S.2.4 Conditional Search

To classify as using a conditional search strategy, responses had to indicate some criterion which served as a source marker. Some responses articulated the steps taken after their memory experience did not meet the criterion used to determine a source, whereas others listed only the criterion; e.g., "if I could think of the clip art associated with it." Virtually all of the responses which used only a conditional search strategy reported anchoring the decision process in their visual memory for pictures; e.g. "I remembered the way it looked so if I didn't have an image in mind but remembered the item I knew it was a word." The only respondent who anchored the decision process in visual memory for text also used a mental time travel strategy: "I tried to remember whether I saw the word in text, then if I couldn't, I chose picture." 7 responses qualified for this category alone, and the 5 responses which overlapped with time travel or visualization make conditional search the most frequently reported strategy.

S.2.5 Encoding response

Encoding response was the least commonly used strategy, with only 2 of the 33 responses qualifying for this category. The responses in this category did not overlap with any category. To qualify for this category, responses had to cite generating encoding responses as a marker of stimulus type; i.e., "I tried to remember the word I put down in response to the item" and "sometimes remembering what I typed/mistyped."

S.3 Self-reports and their classification

Classification	Response
n/a	I had no specific strategy
misunderstood	For items presented in a word, I often copy wrote what it does, while I wrote what the item is for those presented as a picture
	i looked at it and determined
	I tried to just analyze what was in front me
	Not really aware of strategy, just visually processed what was on the screen
	It just appeared in my mind when I read the computer. If it was text a picture, if not then it was a word.
	as a picture, i would picture myself performing an action with the item. When a word, I would picture the item itself
simple memory	memory
	memory
	I remembered some of them vividly as either being a picture or a word
	recalled using my memory
	I remebered in my head, could see it spelt or as a picture
	trying to remember if i saw it or not
time travel	I just thought really hard about what I saw on the screen during part 1
	I tried to think back to when I was describing the picture/word and see if an image or a word came back to me.
	recalling when what I saw during the first part
	By recalling what I saw
visualization	Mainly based on mental image memory in my head and associating the item either with the written word or drawn image.
	Tried to visualize the picture
	Visualization
conditional search (pic*)	I tried to recall whether I had seen a picture of it before or not
	I tried to remember if I had seen a picture, and if I did not but still remembered it, it was a word.
	if i could think of the clipart associated with it

	I thought about whether or not the item could evoke visual imagery. If it could not evoke visual imagery, then I knew it was most-likely a word and not a picture.
	If I could think of the picture I picked picture but if not I assumed it was a word
	If I could recall the picture I put in picture, if I couldn't I then thought if I recalled the word. If I remembered the item but didn't know whether as a picture or a word I took my best guess.
	I remembered the way it looked so if i didnt have an image in mind but remembered the item i knew it was a word.
encoding response	I tried to remember the work I put down in response to the item.
	sometimes remembering what I typed /mistyped
time travel & visualization	I would imagine in my head or try to recall the picture or what the word would look like and see if I could recall
	I visualized what I saw on the screen and chose based off of what I remembered.
	I visualized the picture or the word that I saw on the screen
	I pictured it on the screen to see if I saw that picture or not
	visualize it like I saw it minutes before
time travel & conditional search (pic)	If I could remember what it looked like it was probably a picture, if I didn't or I remember the words helping me come up with a description for it, it was probably a word.
	I tried to remember in my head, if it was a picture, how the picture looked on the screen.
	if i could remember seeing the word
	I tried to remember whether I saw the word in text, then if I couldn't, I chose picture
Visualization & conditional search	Try and visualize the picture in my mind, and if I can't it's probably a word.

Bibliography

Atance, C. M., & O'Neill, D. K. (2005). The emergence of episodic future thinking in humans. *Learning and Motivation*, *36*(2), 126–144. https://doi.org/10.1016/j.lmot.2005.02.003

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/S0079742108604223

Audi, R. (2011). *Epistemology: A contemporary introduction to the theory of knowledge* (Third Ed.). Routledge. Retrieved from https://www.taylorfrancis.com/books/9781136934476

Bartlett, F. C. (1932). Remembering: An Experimental and Social Study. *Cambridge: Cambridge University*. Retrieved from http://books.google.com/books?hl=en&lr=&id=WG5ZcHGTrm4C&oi=fnd&pg=PR9&dq=bartlett+remembering&ots=BAeWcuInfI&sig=rFLXsRDPpEtodXcTFIMScqhAPec

Carrasco, M. (2014). Spatial Covert Attention: Perceptual Modulation. In *Oxford Handbook of Attention* (pp. 183–230). Retrieved from http://www.psych.nyu.edu/carrascolab/publications/Carrasco_OxfordHandbookofAttention_2014.pdf

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. Retrieved from https://market.android.com/details?id=book-oVqsjJvWgkMC

Chalmers, D., & Others. (2003). Consciousness and its place in nature. *Blackwell Guide to the Philosophy of Mind*, 102–142. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470998762#page=114

Bibliography

Cooper, R. A., Plaisted-Grant, K. C., Baron-Cohen, S., & Simons, J. S. (2016). Reality Monitoring and Metamemory in Adults with Autism Spectrum Conditions. *Journal of Autism and Developmental Disorders*, *46*(6), 2186–2198. https://doi.org/10.1007/s10803-016-2749-x

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, *68*(1), 53–74. https://doi.org/10.1086/392866

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62. https://doi.org/10.1016/0010-0277(89)90005-X

D'Argembeau, A., Renaud, O., & Van der Linden, M. (2011). Frequency, characteristics and functions of future-oriented thoughts in daily life. *Applied Cognitive Psychology*, *25*(1), 96–103. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1647

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155–185. https://doi.org/10.1007/s11229-013-0247-7

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/13664879

Dennett, D. (1969). Personal and sub-personal levels of explanation. *Content and Consciousness*, 17-20.

Dennett, D. C. (1993). *Consciousness explained*. UK: Penguin.

Durso, F. T., & Johnson, M. K. (1980). The effects of orienting tasks on recognition, recall, and modality confusion of pictures and words. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 416–429. https://doi.org/10.1016/S0022-5371(80)90294-7

Bibliography

Eichenbaum, H. (1997). Declarative memory: insights from cognitive neurobiology. *Annual Review of Psychology*, *48*, 547–572. https://doi.org/10.1146/annurev.psych.48.1.547

Fernandes, M. A., & Moscovitch, M. (2000). Divided attention and memory: evidence of substantial interference effects at retrieval and encoding. *Journal of Experimental Psychology. General*, *129*(2), 155–176. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10868332

Foley, M. A., Bays, R. B., Foy, J., & Woodfield, M. (2015). Source misattributions and false recognition errors: examining the role of perceptual resemblance and imagery generation processes. *Memory* , *23*(5), 714–735. https://doi.org/10.1080/09658211.2014.925565

Foley, M. A., & Johnson, M. K. (1985). Confusions between memories for performed and imagined actions: a developmental comparison. *Child Development*, *56*(5), 1145–1155. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/4053736

Foley, M. A., Johnson, M. K., & Raye, C. L. (1983). Age-related changes in confusion between memories for thoughts and memories for speech. *Child Development*, *54*(1), 51–60. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6831988

Gettier, E. (1963). Is justified true belief knowledge? *1963*, 273–274. Retrieved from https://books.google.com/books?hl=en&lr=&id=Gp9Umi2VEh8C&oi=fnd&pg=PA175&dq=gettier+1963&ots=OHA2VrZMYY&sig=N9Q5sB7OwuGRNLVKB1G0fa0EyXo

Goldman, A. I. (1993). Epistemic Folkways and Scientific Epistemology. *Philosophical Issues. A Supplement to Nous*, *3*, 271–285. https://doi.org/10.2307/1522948

Goldman, A., & Beddor, B. (2016). Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/reliabilism/

Bibliography

Gonsalves, B., & Paller, K. A. (2000). Neural events that underlie remembering something that never happened. *Nature Neuroscience*, *3*(12), 1316–1321. https://doi.org/10.1038/81851

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. MIT Press. Retrieved from https://market.android.com/details?id=book-1lCMDwAAQBAJ

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Johnson, M. K., & Chalfonte, B. L. (1994). Binding complex memories: The role of reactivation and the hippocampus. *Memory Systems*, *1994*, 311–350. Retrieved from http://memlab.yale.edu/sites/default/files/files/1994_Johnson_Chalfonte_Binding.pdf

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8346328

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67. Retrieved from http://psycnet.apa.org/journals/rev/88/1/67/

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive Operations and Decision Bias in Reality Monitoring. *The American Journal of Psychology*, *94*(1), 37–64. https://doi.org/10.2307/1422342

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York, NY: MacMillan.

Kensinger, E. A., & Schacter, D. L. (2006). Neural processes underlying memory attribution on a reality-monitoring task. *Cerebral Cortex* , *16*(8), 1126–1133. https://doi.org/10.1093/cercor/bhj054

Bibliography

Knott, L. M., & Dewhurst, S. A. (2007). The effects of divided attention at study and test on false recognition: a comparison of DRM and categorized lists. *Memory & Cognition*, *35*(8), 1954–1965. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18265611

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*(2 Pt 1), 149–171. https://doi.org/10.1006/ccog.2000.0433

Lackey, J. (2005). Memory as a Generative Epistemic Source. *Philosophy and Phenomenological Research*, *70*(3), 636–658. https://doi.org/10.1111/j.1933-1592.2005.tb00418.x

Lindsay, D. S., Johnson, M. K., & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology*, *52*(3), 297–318. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1770330

Loftus, E. F. (1974). Reconstructing memory: The incredible eyewitness. *Jurimetrics*, *15*, 188. Retrieved from https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/juraba15&section=38

Loftus, E. F. (2005). Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & Memory* , *12*(4), 361–366. https://doi.org/10.1101/lm.94705

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

Bibliography

McClelland, J. L. (1995). Constructive memory and memory distortions: A parallel-distributed processing approach. *Memory Distortions: How Minds, Brains, and*. Retrieved from https://books.google.com/books?hl=en&lr=&id=P5fTQ2vYkrgC&oi=fnd&pg=PA69&dq=Constructive+memory+and+memory+distortions:+a+parallel-distributed+processing+approach&ots=P_on1_6vfh&sig=A1eLEwW-vnNDUBsMB7DTlQ4dFh8

McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *The Behavioral and Brain Sciences*, *32*(6), 493–510; discussion 510–561. https://doi.org/10.1017/S0140525X09990975

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/09515089.2011.559623

Michaelian, K. (2012). Metacognition and endorsement. *Mind & Language*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0017.2012.01445.x

Michaelian, K. (2013). The information effect: Constructive memory, testimony, and epistemic luck. *Synthese*. Retrieved from https://link.springer.com/article/10.1007/s11229-011-9992-7

Michaelian, K. (2016). Mental time travel: Episodic memory and our knowledge of the personal past. Retrieved from https://books.google.ca/books?hl=en&lr=&id=P9CZCwAAQBAJ&oi=fnd&pg=PR7&ots=gcIE3Migun&sig=RJ-I7JTCZCCTuTnuwvSrgPE1R8I

Michaelian, K., & Sutton, J. (2017). Memory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/memory/

Bibliography

Mickley Steinmetz, K. R., Waring, J. D., & Kensinger, E. A. (2014). The effect of divided attention on emotion-induced memory narrowing. *Cognition & Emotion*, *28*(5), 881–892. https://doi.org/10.1080/02699931.2013.858616

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, *86*(6), 281–297. https://doi.org/10.2307/2027123

Moscovitch, M. (1994). Memory and working with memory: Evaluation of a component process model and comparisons with other models. *Memory Systems*, *1994*(369-394), 224. Retrieved from https://books.google.com/books?hl=en&lr=&id=4gdHL81eaQgC&oi=fnd&pg=PA269&dq=moscovitch+1994&ots=mUYSC2ALBB&sig=9dO9jaoXVgC43bZs80MyXHln2Cc

Neisser, U. (1967). Cognitive Psychology (New York: Appleton-Century-Crofts). *Google Scholar OpenURL Yorkville University*.

Nelson, D. L., Reed, V. S., & McEvoy, C. L. (1977). Learning to order pictures and words: A model of sensory and semantic encoding. *Journal of Experimental Psychology. Human Learning and Memory*, *3*(5), 485. Retrieved from http://psycnet.apa.org/record/1978-11577-001

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, *70*(19), 699–711. https://doi.org/10.2307/2025079

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1. Retrieved from https://psycnet.apa.org/journals/bul/80/1/1/

Pylyshyn, Z. W. (2002). Mental imagery: in search of a theory. *The Behavioral and Brain Sciences*, *25*(2), 157–182; discussion 182–237. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12744144

Bibliography

Ritchey, M., Wing, E. A., LaBar, K. S., & Cabeza, R. (2013). Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. *Cerebral Cortex* , *23*(12), 2818–2828. https://doi.org/10.1093/cercor/bhs258

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *21*(4), 803. Retrieved from http://psycnet.apa.org/fulltext/1995-42833-001.html

Saryazdi, R., Bannon, J., Rodrigues, A., Klammer, C., & Chambers, C. G. (2018). Picture perfect: A stimulus set of 225 pairs of matched clipart and photographic images normed by Mechanical Turk and laboratory participants. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1028-5

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*(3), 501–518. https://doi.org/10.1037/0278-7393.13.3.501

Schacter, D. L., & Addis, D. R. (2007a). Constructive memory: the ghosts of past and future. *Nature*, *445*(7123), 27. https://doi.org/10.1038/445027a

Schacter, D. L., & Addis, D. R. (2007b). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289–318. https://doi.org/10.1146/annurev.psych.49.1.289

Bibliography

Senor, Thomas D., "Epistemological Problems of Memory", *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/fall2014/entries/memory-episprob/>

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology. Human Learning and Memory*, *6*(2), 174–215. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7373248

Sperling, G. (1967). Successive approximations to a model for short term memory. *Acta Psychologica*, *27*, 285–292. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6062221

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13515–13522. https://doi.org/10.1073/pnas.93.24.13515

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs*, *123*(2), 133–167. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9204544

Troyer, A.K., Rich, J.B. (2018). *Multifactorial Memory Questionnaire: Professional Manual.* [online version]. Retrieved from https://www.baycrest.org/Baycrest_Centre/media/content/form_files/MMQ-Manual-2018_ebook.pdf

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1. Retrieved from http://psycnet.apa.org/journals/cap/26/1/1/

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology*, *53*, 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Bibliography

Wheeler, M. A., & Roediger, H. L. (1992). Disparate Effects of Repeated Testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results. *Psychological Science*, *3*(4), 240–246. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Yoshihara, M., & Yoshihara, M. (2018). "Necessary and sufficient" in biology is not necessarily necessary - confusions and erroneous conclusions resulting from misapplied logic in the field of biology, especially neuroscience. *Journal of Neurogenetics*, *32*(2), 53–64. https://doi.org/10.1080/01677063.2018.1468443