THE FUNCTION OF EPISODIC MEMORY:
IS EPISODIC MEMORY A PRESERVATIVE COGNITIVE SYSTEM?

DANIEL CASEY

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY
YORK UNIVERSITY
TORONTO, ONTARIO

February 2024

ABSTRACT

In contemporary philosophy of memory, the dominant account has been the causal theory of memory. Established by Martin and Deutscher in 1966, the causal theory proposes that episodic memory can be differentiated from other cognitive processes, such as imagination, by referencing a causal connection to past experiences. This causal connection is essentially the idea of a memory trace: that information is stored at the time of an experience and can be called to mind when an individual is engaged in the process of remembering. A possible implication of this account is that episodic memory is a *preservative cognitive capacity*. That is to say that the cognitive faculty of episodic memory is concerned with the storage and retrieval of information.

More recently, however, studies of various types of memory errors (DRM Paradigm, Loftus paradigm) suggest that the content in episodic representations might not be entirely preservative, and that the episodic memory system might be responsible for generating some of this content. These studies present a particular challenge for strictly preservative models of the function of episodic memory, and have prompted two lines of response. The traditionalist response is to maintain that episodic memory is still primarily a preservative cognitive capacity that requires a memory trace, but to leave some room for content generating features. On the other side, the revisionist response is to propose a full-scale rethinking of the function of episodic memory, possibly even dispensing with the need for a memory trace.

The focus of this dissertation is to assess the challenge presented by these kinds of memory errors, and to make the case that they are not inconsistent with a traditionalist model of episodic memory. By comparing the challenging cases of memory errors with more common types of memory error that are not understood to be problematic, I aim to show precisely what it is about these memory errors that presents a challenge for preservative models of memory. With this established, I will assess both the traditionalist and revisionist arguments, and argue that the cognitive function of episodic memory should still be understood as being broadly preservative.

ACKNOWLEDGEMENTS

I'd like to thank my supervisor, Muhammad Ali Khalidi, for his gentle perseverance and guiding insights—without him, this dissertation wouldn't have come to be. I'd like to thank my supervisory committee, in particular Jacob Beck, for his encouragement and support over the course of this project, and Kevin Lande for sharing his helpful advice. I'd also like to thank Sarah Robins, whose ideas have strongly influenced my work, and who was kind enough to review my dissertation.

Most importantly, I'd like to thank my family. My mother for always listening to my thoughts, and my father for always pushing me to make them real. I'd like to thank my sister, Kate, for being my confidante, my brother, Patrick, for being my best-friend, and my youngest brother, Richard, for being the voice of reason I find myself listening to more and more. I'd also like to thank my girlfriend, Aliesha, for putting up with my nonsense and loving me none the less for it. I love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Episodic Memory and the Causal Theory

The focus of this thesis is to assess the cognitive function of episodic memory. The role of episodic memory is an issue that has received considerable attention in recent philosophical and cognitive science discussion. The more contentious positions in this discussion maintain that episodic memory is not, in fact, a distinct cognitive system. Proponents of these positions argue that episodic memory is actually part of a larger cognitive system, one that concerns itself with first-person mental simulation—episodic hypothetical thinking (De Brigard 2013), or mental time travel (Michaelian 2016b). On these views, episodic memory does not have a distinct cognitive function, and can better be understood alongside future-oriented mental time travel, episodic counterfactual thinking, and mind reading, as a species of episodic imagination (Michaelian 2016b). On the other side of this debate, more moderate positions maintain that episodic memory does constitutes a distinct cognitive system, with a cognitive function that amounts, roughly, to representing past experiences.

The contentious side of the debate is bolstered, primarily, by a growing body of evidence from psychology and neuroscience that suggests episodic memory is not as reliable as it has typically been understood to be (Deese 1959, Roediger and McDermott 1995; Loftus 2003). This, coupled with evidence from neuroscience that shows considerable overlap in the brain regions responsible for episodic memory, counterfactual thinking, and hypothetical simulation (Szpunar 2010), gives reason to question traditional accounts of episodic memory. In response to this evidence, those holding the more contentious positions have proposed a wholescale revision of episodic memory (De Brigard 2013; Michaelian 2016b). On the other side of this debate, those holding more moderate positions maintain that such a revision is premature (Bernecker 2004; Robins 2016). They look to account for these evidentiary discrepancies within the traditional framework of episodic memory.

The traditional framework of episodic memory corresponds with the way that we typically conceive of memory as working: that our memory systems are responsible for the storage and retrieval of representations of past experiences. This is a conception of memory that is deeply imbedded in our theoretical and practical understandings of human cognition and behaviour.

This framework is appealing, in part, because it fits so well with our intuitive, or folk, understanding of how memory works. Much of what we do, how we understand ourselves, how we treat each other, and the rules, laws, and social practises that we follow, are predicated on this understanding of episodic memory. The stakes, then, in this debate are by no means insignificant, with implications that extend through theoretical positions in philosophy, cognitive science, and psychology, but also into the more practical applications of personal knowledge, personal identity, agency, testimony, and many other areas that make use of this model of episodic memory.

The goal of this project is to reconcile the evidentiary discrepancies, or inconsistencies, put forward by revisionary accounts of episodic memory with an account of episodic memory that is consistent with more traditional models. By assessing the cognitive function of episodic memory, I hope to show that many of these issues can be accommodated without proposing a full-scale revision. If the function of the episodic memory system can be refined, so as to show that many of the problems faced by a traditional conception of episodic memory can be accounted for as functional errors, or features of the function of episodic memory, then this will go a long way towards vindicating the traditional understanding of memory.

1.1 Introduction

According to traditional models, an important function of episodic memory is that it represents past experiences. Just because the function of episodic memory, on these traditional models, is to represent past experience, this doesn't mean that it always does so successfully—episodic memory, like many other cognitive faculties, is capable of malfunctioning. An implicit assumption of these traditional models, however, is that when functioning properly, the representations available in episodic memory are faithful representations of what actually happened, or at least of the experience of what happened. Generally, we take it that the memories we have are of events that happened to us, and that the details in those memories are consistent with the details of those events. In other words, episodic memories are accurate and authentic representations of past experiences. Consistent with these assumptions is the idea that, when these representations are not accurate and/or authentic representations of past experiences, they

should be understood as failures of the memory system (Robins 2016b, p. 436). This is the point on which revisionary accounts of episodic memory depart from traditional models: they don't see cases of misremembering necessarily as failures in episodic memory, since they don't take the function of episodic memory to be representing past experiences. Instead, they see episodic memory as part of a more inclusive cognitive system that functions to represent hypothetical past, present, and future-directed episodic content (De Brigard 2013, p. 174-5). This is an important point of disagreement between the two sides of the debate, and relates directly to how the function of episodic memory should be understood.

The idea that episodic memory doesn't have a function of representing past experiences is unsettling, in part because of how ingrained in human behaviour and attitudes this concept is. The idea that episodic memory functions to represent past experiences has a long history and provides a foundation for how we understand the world and our place within it. In this chapter, I want to explain this concept of episodic memory, and why it's so important for making sense of many of our beliefs, behaviours, and attitudes. Within this understanding of episodic memory, the assumption that episodic memories are accurate and authentic representations of past experiences is key to making sense of how, and why, these models of episodic memory are so influential. This, I will argue, is an important assumption in traditional models of episodic memory, and something to consider when we discuss the function of episodic memory in greater detail.

In the first section, I'll begin by setting the parameters of this discussion with a straightforward question: What is episodic Memory? This section will outline the general concept of memory and explain how episodic memory is differentiated from other classifications of memory, in particular semantic memory. In the second section, I'll look at how the traditional model of episodic memory, our folk concept of memory, is widely featured in our theoretical and practical understanding of human attitudes and behaviours. Within this traditional model of episodic memory, there is an implicit assumption that the representations produced in episodic memory are accurate and authentic representations of past experiences. In the third section, I will look at theories of remembering: how philosophers have addressed the concept of episodic memory. Here, I'll look at several important issues that come up in philosophy of memory. I'll also outline

the dominant account of memory in philosophy and cognitive science—the causal theory of memory—and explain how it fits with traditional ideas about memory. In the fourth, and final, section, I'll explain the idea of a memory trace, a central idea in the causal theory and in working models of memory. Overall, this should provide a useful overview and develop some key insight regarding traditional accounts of episodic memory and how they understand the function of episodic memory.

1.2 What is episodic memory?

Episodic memory, sometimes referred to as autobiographical memory[1], is the memory we have of specific events, or episodes, that we have experienced in our lives. When you recall a particular experience, bringing to mind the vivid details with which you initially experienced it, it is thought to be the episodic memory system that is responsible for the phenomenon of mentally reliving this event. In this way, episodic memory is a sub-category of the broader concept of memory, one that deals specifically with the recollection of first-person experienced past events.

Whether the broader concept of memory, as it has come to be used, actually refers to a distinct category or kind, is debatable, as is the proposed taxonomy, and how to go about delineating the different kinds of memory that fall within this classification. Nonetheless, a generally accepted breakdown of memory in the cognitive sciences follows (see fig.1). Types of memory are usually divided into two main classifications: declarative, and non-declarative memory. These classifications roughly correspond with the common distinction between memory *of* (declarative) and memory *how* (non-declarative).

[1] Episodic memory is sometimes distinguished from autobiographical memory, because episodic memory relates to particular episodes from our personal past, whereas autobiographical memory relates to a more complete understanding of our past, that may entail knowledge form other memory systems.

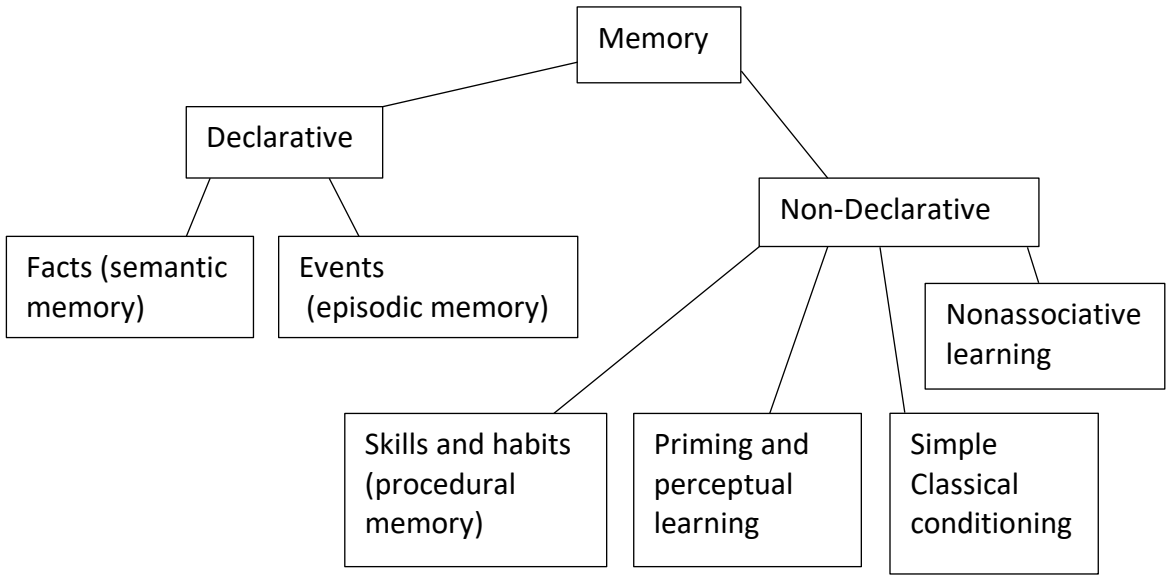

Figure 1 Squire's taxonomy of mammalian long-term memory systems (2004).

The declarative memory system consists of cognitive memory systems that store and retrieve information that can be expressed in terms of its propositional content. These include episodic memory, memory of specific events from a person's past, and semantic memory, memory of facts or propositions. Non-declarative memory, by contrast, consists of memory systems that do not involve the cognitive storage and retrieval of information with propositional content. This classification has a broader range, and is less clearly defined, consisting of various phenomena that share similar functional properties but lack the cognitive availability, and propositional content, characteristic of declarative memory (Michaelian 2016b, p. 20). These include procedural memory (memory of skills and habits), priming and perceptual learning, simple classical conditioning, and non-associative learning.

Because of the wide range of systems that are grouped under the classification of non-declarative memory, it is unlikely that they display sufficient similarity for the category to constitute a natural kind—a category that represents a grouping as it occurs in the world, absent human interests. Similarly, the broader classification of memory takes on this same difficulty and, more likely, constitutes a nominal kind—a weaker classification that is based on human convention rather than a distinction in reality. Michaelian argues that, if we analyse the general classification

of memory according to Marr's three-level description of cognitive systems, then there are significant differences between declarative and non-declarative memory at each of these three levels (2016b, p. 22-3). If we accept his arguments, this would rule-out the possibility of memory being a unified cognitive system and, subsequently, a natural kind. Declarative memory, however, and its sub-categories of episodic and semantic memory, have more clearly defined parameters, and are better candidates for natural-kind status. Khalidi suggests two possible routes for making the case that episodic memory constitutes a natural kind: one being with respect to the distinct phenomenology of episodic memory, and the other being with respect to the etiological, or causal, properties of episodic memory (2023, p. 123-157). Accordingly, there is reason to suspect that the categories of semantic and episodic memory refer to specific cognitive systems and have the potential to be classified as natural kinds.

1.2.1 Episodic and semantic memory

While both species of declarative memory, episodic memory and sematic memory, are functionally similar with regards to the storage and retrieval of information, they differ with regards to the way in which the retrieved information is consciously experienced. Using the term "episodic" as a means of differentiating memories of past experiences from semantic memories was first proposed by Tulving (1972, p. 384), although the distinction had been implicit in psychology for some time, and this has since become the generally accepted distinction. According to this distinction, sematic memory is thought to be associated with noetic consciousness—in remembering a fact, one is consciously aware that they are doing so. Episodic memory, on the other hand, is thought to be associated with autonoetic consciousness—in remembering a past experience, one is not only consciously aware that they are remembering the associated event, but also that they have experienced it, first-hand (Tulving 1983). For example, when I remember that the capital of France is the city of Paris, I am aware that I am remembering this information (noetic consciousness); however, when I remember having had a specific meal in Paris, I am not only aware of where and when that meal was served, but I have the accompanying awareness of the experience of eating that specific meal at that specific time and place (autonoetic consciousness).

There is undoubtedly an important connection between episodic and semantic memory. We see this in how our memories of certain facts are often connected with our memories of past experiences (as in I am only aware that the specific meal I ate in Paris was served at the time and place I ate it because I, in fact, did experience eating it then, and there). It can also happen that we retain memories of past experiences precisely because they are associated with novel experiences and information: I may only remember that specific meal, of so many others, because it was a meal I was as yet unfamiliar with (boudin noir)[2]. Intuitively, this connection between the experiences and facts that we remember is straightforward. Such a connection between the contents of episodic memory and semantic memory would likely suggest an overlap, or interconnectedness, between the two memory systems. Despite this interconnectedness, however, there is good reason to suspect that the two classifications refer to distinct systems.

In particular, there is evidence of patients observed where defects in episodic memory are not apparent in semantic memory. Frequently cited in support of this distinction is the extensive study of patient KC, a man who suffers amnesia as a result of a head injury that damaged his left frontal-parietal and right parieto-occipital lobes (Tulving, Schacter, McLachlan, and Moscovitch 1988). Patient KC suffered extensive memory impairment from the period before the injury that caused the onset of his amnesia (retrograde amnesia), as well as in the period following the onset of his amnesia (anterograde amnesia). In these studies, he was shown to retain general semantic information from before his injury, but was unable to recall any personal experiences, either from before, or after. Over the course of the study, through prompting, he was eventually able to recall semantic details about his personal life before the accident—about his work, his family, and his friends—, but could not recollect any events in the context of his personal experiences.

This study is taken as evidence of a single dissociation between episodic and semantic memory. The reasoning follows: if episodic and semantic memory are, in fact, distinct systems, then it would be possible for an injury to damage one system (in this case the episodic memory system) while leaving the other (semantic memory), at least partially, intact. Because of the

[2] This connection might, in part, result from an increase in attention: we have a tendency to pay more attention to new, or novel, experiences, and attention has been shown to influence the encoding of episodic memory (Aly and Turk-Browne 2016).

interconnectedness between the neural regions associated with both systems, however, such injuries are rare. KC's is the most extensive, and telling study, but other studies have confirmed similar results (Kapur 1999; Wheeler and McMillan 2001). What these studies are taken to show is that it is, in fact, possible for such an injury to occur, which provides empirical support for the disassociation between episodic and semantic memory systems.

The implications of a single disassociation are, however, limited. It's possible to damage the memory system such that one capacity (episodic memory) could be impaired while another (semantic memory) remains intact. This would show that the two capacities are responsible for different processes, but not that they are independent systems. If the memory system were damaged such that an impairment in semantic memory also resulted in an impairment in episodic memory, this would still be consistent with a single disassociation, but not indicative that the two systems were independent. To make the latter case, we would want to show evidence of a double disassociation.

More controversial are cases of semantic dementia. In such cases, patients show a degeneration in basic semantic memory, but retain relatively intact episodic memory from recent past events (Irish, Addis, Hodges & Piguet 2012). These types of cases give the flip-side of the amnesia cases discussed above: they show that it is possible for an individual to have an, at least partially, functional episodic memory system, while their semantic memory systems is impaired. Together, then, these cases of amnesia and semantic dementia are evidence of a double dissociation between the semantic and episodic memory systems: the semantic memory system can function without the episodic memory system, and vice versa. This evidence supports the notion that these two classifications of declarative memory refer to distinct cognitive systems: if the two systems were not independent, we should assume that a break-down in the function of one system would necessarily carry-over into the other, which doesn't appear to be the case.

While there is more to be said about the relation between semantic and episodic memory, and we will return to this throughout this project, it suffices for now to point out that the important distinction between them is the type of consciousness with which we recollect a given information: whether we recollect it as a first-person experience, or as a fact. And that there is

good evidence, in the cases of Alzheimer's and semantic dementia discussed above, of memory impairments in one memory system not persisting in the other, to suspect that these two classifications refer to distinct cognitive systems. Since this dissertation is concerned primarily with the function of the episodic memory system, unless otherwise specified, the term "memory" will be used in reference to episodic memory—apart, however, from discussions on the relationship between episodic and semantic memory, the paper will not extend into other classifications of memory.

In the next section, I will discuss folk conceptions of episodic memory, and how this understanding of episodic memory plays a significant role in many of our beliefs, attitudes, and behaviours. Implicit in these folk conceptions is the assumption that the representations made available in episodic memory are accurate and authentic representations of past experience—an important feature to consider in relation to the function of episodic memory.

1.3 Folk conceptions of episodic memory

Although the term "episodic memory" might be a relatively recent coinage, the concept is not. The idea that we can deploy first-person recollections of events from our past has probably been around since as early as humans began to recognize and understand their cognitive faculties—and the many ways that we have come to rely on this cognitive faculty probably far outdate that. As early as Aristotle, memory was understood as the distinct way through which we relate to our past experiences (Michaelian 2016b, p. 4). This idea that humans can cognitively store and retrieve information about past experiences is something that we've naturally come to assume about our cognition. In this way, memory is foundational for human knowledge—in fact, it is often classified as a basic epistemic source, along with perception, introspection, inference, and communication (Michaelian 2016b, p. 8). It would be hard to imagine our cognitive processes working as they do without this ability to store and recollect facts and experiences.

Storing information, however, is not exclusive to episodic memory. While the importance of declarative memory as a whole is undeniable, this doesn't tell us about the particular importance of episodic memory, as a system of memory distinct from semantic memory. At many levels, the

relation between semantic and episodic memory is not fully understood, which creates a problem for identifying the exact functional significance of either system: do the facts and propositions we recall derive from stored past experiences, and so does episodic memory constitute the general foundation for all declarative memory; do we first store the propositional content of our experience, and then use it to recreate first person episodic content; or, do the two systems occur somewhat independently, and simply arise from the same aggregate of past experience, one taking the form of a recollection of facts, and the other a recollection of experience? Without being able to answer these kinds of questions, it's impossible to determine the precise scope of either of these memory systems, which makes it difficult to say exactly what the cognitive function of episodic memory is. This, in turn, creates a difficulty for determining how imbedded in human cognition and behaviour episodic memory truly is.

We'll return to this dilemma in greater detail when we look specifically at the function of episodic memory in chapter four. For now, we simply want to get an idea of how our common-sense understanding of episodic memory features in everyday life. One way to get a handle on this is to look at human beliefs, behaviours, and attitudes that rely on a traditional understanding of episodic memory. This should give a rough idea of how deeply rooted in human behaviour common notions of episodic memory are.

1.3.1 Why are these models of episodic memory important?

I doubt I can give an exhaustive account of the ways that folk concepts of episodic memory feature in our everyday beliefs, attitudes, and behaviours, but a sensible starting point, I think, is to consider the distinction that is unique to episodic memory—namely, that it entails autonoetic, or first-person experienced consciousness. From this point, it is fair to assume that many of the human behaviours and attitudes that rely on episodic memory will share, or in some way make use of, this first-person concern.

If we take declarative memory to be a basic epistemic source, then a good starting point is to look at the kinds of knowledge that are related directly to our personal experiences. One form of knowledge that stands out in this regard is the knowledge we have of our personal past.

Typically, we expect that much of what we know about ourselves comes from our memory of past experiences. Of course, we can learn things about our past from the testimony of others as well—the stories of our parents, siblings, and peers— and it might be the case that, even if we knew nothing about ourselves from our own memories, we could still have a substantial basis of knowledge about our past obtained purely from these kinds of indirect sources (stories, diaries, photo albums, etc.). Still, it seems that even the most thorough account of one's personal past, derived entirely from second-hand sources, would be significantly lacking in comparison with the kind of personal knowledge that the average adult human enjoys. This is because much of what constitutes the average adult human's knowledge of their personal past is made up of their recollections of past experiences—their episodic memories. In this respect, then, it seems that an important part of our knowledge about our past experiences comes from episodic memory.

Along similar lines, or possibly even as a result of this knowledge, comes another important example—personal identity. In philosophy, questions about personal identity are concerned with determining the criteria for identifying a person over a period of time. One prominent suggestion is that this criterion can be found in human consciousness. Locke argued that a person extends as far as their consciousness of their past thoughts and actions. He thought that it was precisely this feature of our cognition, our ability to recollect our past experiences, that determines the extension of a person over a period of time: "[for a person] can consider itself as itself, the same thinking thing in different times and places" (Locke 1689, 2.27.9). In particular, for Locke, it is what we now refer to as autonoetic consciousness that is essential to personal identity. It is not just that we can recollect things from our past, but that we can recollect things as we experienced them—from a first-person perspective. This theory of personal identity, psychological continuity, has been refined by more recent proponents (Shoemaker 1970, Parfit 1984), but still maintains its reliance on episodic memory[3].

[3] While memory accounts of personal identity have been dominant in the philosophy of personal identity, these are not without criticism. Early criticisms of Locke's view include Butler's circularity objection (1736) and Reid's transitivity of identity objection (1785). In response to these criticisms, Shoemaker suggests the concept of quasi-memory (memory that does not presuppose selfhood) (1970), and psychological continuity (overlapping memory connections) as opposed to direct memory connection (1984); Parfit ultimately dispenses with the identity component in questions of personal identity, suggesting what really matter (from a practical standpoint) is psychological continuity (1984).

While questions of personal identity might seem to present abstruse philosophical problems, for Locke the issue took on practical significance in that he argued establishing personal identity was necessary for attributing moral culpability— "in this personal identity, is founded all the right and justice of reward and punishment" (Locke 1689, 2.27.18). Others, too, have used accounts of personal identity (or psychological continuity) as foundational to their ethical theories (Parfit 1984, McMahan 2002)[4]. The central idea being that in order to track moral responsibility, we have to be able to track the same moral agent over a period of time, where personal identity looks to establish this through identifying the continuity of a person.

This concept of a moral agent, or even that of agency in general, features prominently in assessments of moral and rational deliberation. We generally consider the agent, herself, as being *the* central component in any kind of deliberation, and agency as a prerequisite for attributing rationality and responsibility. Bratman proposes that human agency consists in three core features: reflectiveness, planfulness, and the conception of our agency as being temporally extended (2000, p. 35-46). He argues that an interdependence between these three core features is required to create the kind of purposeful agency found in human beings. Agency requires the ability to make plans, to think of the past, commit to the future, and see those commitments fulfilled. To do this, the agent has to be able to conceive of their self as something capable of actualizing these diachronic plans. For Bratman, making sense of this temporally extended sense of agency relies on a variation of psychological continuity, which is, essentially, constituted by episodic memory.

Gerrans and Kennett argue that mental time travel, episodic memory and imagination, and, in particular, the quality of autonoetic consciousness, are necessary components of moral agency (2010, 601-2). Although it's not implausible to imagine conceiving of and fulfilling plans on the basis of purely semantic memory, they argue that the autonoetic quality of episodic memory, and the ability to project oneself into the future, creates a sense of ownership, or "mineness", that is essential for agency. On their view, without this sense of ownership, the commitments we make lose their impetus: an agent requires this personal, or intimate, connection with the episodes that plot their moral path in order for them to feel compelled to follow it. The same framework can be

[4] For an overview see *Personal Identity and Ethics: A Brief Introduction* David Shoemaker (2009)

applied to practical reasoning: we can successfully create, and execute, the plans that we've created because we can conceive of ourselves as temporally extended agents capable of doing so. Without this, they argue, the sorts of normative demands made by moral and practical reasoning lose their force.

There are, however, alternative accounts of agency that do not call for this degree of rational deliberation. Some even place no requirements on rational deliberation (Doris 2016). The same can be said of moral theories: there are moral theories that place little emphasis on moral deliberation—some version of sentimentalism, for example (Haidt 2001). These views, however, are generally revisionary, and I think it's fair to say that they don't accord with our traditional intuitions about agency and moral responsibility. At a profound level, then, it seems moral responsibility, or at least our common-place notions about moral responsibility, and many derivative human practises and behaviours, rely on a traditional model of episodic memory.

The practical implications of this reliance on a traditional understanding of episodic memory are extensive. Even if we learn about ourselves, or the world we live in, through the stories of others, we expect that at some point, much of this information is derived from the stored content of past experiences. The notion of testimony gains its significance precisely because we take past experience as a reliable epistemic indicator[5]. There'd be no sense asking for someone's recollection of an event, or calling a witness to trial, if we didn't believe that these recollections of first-person experiences were accurate sources of information. So many moral practises—promises, favours, friendships, and all sorts of relationships—are predicated on sharing memories of particular past experiences that are assumed to be accurate. At a point, the list of the practical significance of our conception of episodic memory becomes exhaustive, with the implication being how extensively we deploy these common-sense ideas about episodic memory.

[5] The connection between episodic memory and testimony has long been observed, but for a more contemporary perspective on this relationship see Mahr and Csibra (2017). Their view is discussed in more detail in chapter four.

1.3.2 Accuracy and authenticity

What these common-sense ideas about episodic memory tell us, or what all of these practises assume about episodic memory, is that it is a cognitive feature whose function it is to store and retrieve representations of our past experiences. In these cases, there is an implicit assumption that the representations made available in episodic memory are, at least in a significant way, faithful representations of actual experiences. This is to say that they are accurate, in that the content made available in these representation is consistent with the initial experience, and authentic, in that they represent events that were actually experienced.

We assume, for example, that the events we remember from our life, are events that actually happened to us, that the plans we started, and the promises we made, are worth seeing through because we actually committed to them, and that the testimony of others can verify occurrences to which we were not witnesses, because they were actually there, as witnesses. We rely on episodic memory in theses cases, precisely because we believe that the representations made available in episodic memory are authentic representations of past experiences. That is to say, we believe that a given episode is available in one's episodic memory because it represents an event that the individual experienced. What's more, we believe that the details made available in such a representation are consistent with those of the actual event, or at least consistent with that individual's experience of the event.

In certain cases, it might be that greater significance is placed on either the authenticity or the accuracy of a memory. For example, in cases of personal identity, the contents of the memory are not as important as is the relationship between the present individual and their cognitive awareness of a past event. In such cases, one might maintain the authenticity of the memory is the more important feature. In other cases, for example when I ask you to recall the memory of a certain event (say, for example, where you parked your car), it is the content of the memory, and not the authenticity of its origin that is most relevant. But, often, the two dimensions, accuracy and authenticity, are interconnected: we tend to assume that details (accuracy) are correct because the event was experienced (authenticity); or, conversely, that because the details are correct (accuracy), the event was actually experienced (authenticity). In cases of testimony, for

example, we assume that the two dimensions are related: because you were there, you can give an accurate rendition of the event. It is, however, important to recognize both of these dimensions when making sense of the significance of episodic memory.

All of these common beliefs and practises maintain that episodic memories are representations of past experiences, and that these representations are generally faithful representations of experienced past events. These assumptions underly many of the theoretical and practical ways that traditional models of episodic memory feature in human behaviours. If these assumptions were false, it would be a monumental task of revising our everyday beliefs and practises.

This section is intended to provide a rough outline of our common intuitions about episodic memory—how we typically conceive of it, and how this conception is featured in everyday belief and behaviour. I should make it clear that this is not an independent reason for accepting that the function of episodic memory is that of representing past experiences—if our common-sense assumptions about episodic memory turn out to be wrong, then we should revise the way we understand memory, and just because we rely so heavily on this common-sense conception of episodic memory, doesn't necessarily mean it's correct. It is, however, important to note how prevalent traditional models of episodic memory are.

In the next section, we'll look at some of the philosophical questions and accounts that address episodic memory.

1.4 Theories of remembering

In philosophy, theories of remembering are concerned primarily with identifying the distinct criteria that define episodic memory. What is it, precisely, that distinguishes a memory, and the process of remembering, from other cognitive phenomena and processes? Early empiricist accounts focused on how to differentiate memories from other cognitive phenomena such as beliefs, ideas, and imaginings.

Before Tulving's useful distinction between autonoetic and noetic consciousness, a prominent suggestion was that memories were identifiable by their unique phenomenology (Hume 1739; Russell 1921). If you consider the things that you know about yourself, the knowledge of your personal past, there is an important difference between the cases where you actually remember a particular experience from your past, and those where you merely remember information about yourself. In cases where you remember a particular experience, this is typically accompanied by a distinct phenomenology—a vividness, or familiarity, that is characteristic of memory. Early empiricists took this phenomenal quality as the defining feature of memory—something which could differentiate memories from beliefs and ideas. Presumably, this same phenomenological quality, the first-person experiential feature of memory, is what motivates Tulving's distinction between noetic and autonoetic consciousness, and which now constitutes the conventional distinction between semantic and episodic memory.

Another important distinction that preoccupied early thinkers was the difference between memory and imagination. While the phenomenal quality of a memory might distinguish it from other cognitive phenomena, the same can't necessarily be said of imagination. Like memory, imagination is considered to have a particular phenomenological quality. Although many would argue that there is a distinct phenomenological difference between our memories and our imaginings, pinpointing exactly how to differentiate these two cognitive phenomena puzzled early empiricists, and still puzzles some contemporary philosophers (Michaelian 2016b; De Brigard 2013).

Establishing these different cognitive classifications prompted some of the early questions confronted in the philosophy of memory, questions that are still central in discussions today. On the one hand, there is the question of how to differentiate episodic and semantic memory, and on the other, how to differentiate episodic memory from episodic imagination. With regards to the first question, Tulving's distinction between autonoetic and noetic consciousness provides a useful, and conventionally accepted, distinction between these two types of memory. Still, however, there is more to be said about this distinction and the complex relationship between these two classifications of memory, which we'll discuss in greater detail in chapter four. With regards to the second question, how to differentiate episodic memory from episodic imagination,

or whether such a distinction can even be made, is still an ongoing debate in philosophy of memory, and closely related to the central topic of this thesis. An important step in response to this second question came when Martin and Deutscher proposed that episodic memory was distinguished by a causal criterion (1966). In the next section, we'll look at their proposal in more detail.

1.4.1 The causal theory

In a paper titled "Remembering", Martin and Deutscher argued that memory was identified by a causal criterion. At the time, this wasn't a widely accepted view, but has since come to be the dominant position in theories of remembering.

When they published their article, the idea that belief was a necessary condition for memory was still common (Hume 1739; Russell 1921). Martin and Deutscher, however, made the case that, while belief might be intuitively associated with memory, it can't be a necessary condition. To see this, one simply has to recognize the possibility of having a memory without realizing it to be such, and therefore not having a belief in its content. If you've ever experienced the perplexing sensation that an idea, phrase, or image that you've conjured in your imagination might not be the genuine product of your own creative thinking, but rather a representation of something that you've seen, or heard, elsewhere and stored away in the back of your mind, then their reasoning should be apparent. The experience of confusing memory with imagination, and vice versa, is a common phenomenon—*cryptomnesia* occurs when a memory we've forgotten returns without us recognizing it as a memory, and instead take it to be an original representation or idea. This is often associated with cases of unintentional plagiarism (Schacter, 2002. p. 108-9). Although Martin and Deutscher don't explicitly identify *cryptomnesia* in their paper, the example they describe, wherein a young painter recreates a pastoral scene from his youth without realizing it was something he witnessed in childhood, is consistent with this kind of memory bias (1966. p.167-68). Such cases rule-out belief as a necessary condition for memory. If it's possible to have a memory without believing it represents an actual past experience, then it's possible to have and episodic memory without the corresponding belief, therefore belief can't be a necessary

criterion of episodic memory. Instead, Martin and Deutscher proposed that a causal condition is necessary for memory.

On their view, remembering requires 3 criteria. In order for a person to remember something, they must: 1) represent that thing, 2) have experienced that thing, and 3) their experience of that thing must be operative in producing their subsequent representation of it (Martin and Deutscher 1966). The first criterion is that in order for someone to be remembering something, they must in some way, either outwardly (such as through speech or writing), or inwardly, (such as through introspection, or recollection), be representing that thing. The second criterion is that the person must have experienced the thing that they are now representing—if a person never experienced something, then they can't rightly be said to remember it. These first two criteria are straightforward, since for someone to be remembering something, they must, at the time of remembering, have the corresponding memory, which is effectively a representation, and that representation must be of something they experienced—otherwise it could be simply an imagining.

Their third criterion introduces the causal condition. The idea that the initial experience now being represented in memory must be operative in producing the representation that constitutes the memory, implies that there is a causal connection between the initial experience, and the representation being remembered. This can't, however, be any type of causal connection (or series of successive causal states). Simply having a causal connection between the experience and the representation would allow for all sorts of deviant causal chains that wouldn't intuitively qualify as cases of memory. Consider, for example, the case of someone suffering from Alzheimer's. Such an individual might have an experience, write down that experience in their diary, and then read their diary out loud at a later date. In this case the individual would satisfy the three criteria: they would (1) have had the experience (the experience they wrote about in their diary), (2) be representing it (through writing it down and reading it out loud), and (3) their initial experience would be operative in the production of their current representation (via the diary entry). In this case, most people would agree that the individual in question is not, in fact, remembering.

While the above example isn't quite as creative, or thorough, as some of the examples Martin and Deutscher provide, it makes the case for why any kind of causal condition is not sufficient for memory. For something to be considered a memory, the past experience and the present representation must be connected via the appropriate causal connection. To this criterion, Martin and Deutscher include the two additional stipulations: (1) the causal connection must be operative in the right kind of way. This is intended to rule-out cases of prompting. Specifically, they discuss cases where the individual doing the remembering is prompted to the contents of the memory, and can only represent the content of their memory to the extent that's been prompted (as in the example of the diary discussed above). And, 2) that the causal connection goes via memeory trace. They explain that such a memory trace would be a structural analogue of the state resulting from the initial experience. The idea of a structural analogue is simply a representation of the requisite details of the thing being represented— as example, they suggest Wittgenstein's structural analogy between the music being played on a gramophone and the grooves in the record: for an increase in pitch, there is a corresponding increase in the frequency of the grooves, for an increase in loudness, there is a corresponding increase in the grooves' deflection, etc. (Martin and Deutscher 1966, p. 190). A complete structural analogue would represent all of the features of the content it represents. They settle, however, for a structural analogue that is, at least, consistent with all of the features being represented by memory: in other words, the memory representation can't contain more information than is available in the trace.

Since Martin and Deutscher proposed this account of remembering, the idea of a causal condition has become the dominant view in theories of remembering. This is likely because it captures so neatly our common-sense notions about how episodic memory works—that memories are stored representations of past experiences that are called to mind when we are engaged in remembering. In defining the causal condition, they've satisfied the implicit assumption that memories are authentic and accurate representations of past experience. The causal condition requires that memories are both connected to the experience of the actual event (authenticity) and that the memory trace is a structural analogue of the experience of that event (accuracy). Thus, their account fits neatly with common sense ideas about how episodic memory works.

An important issue worth noticing, here, is that the causal theory fits neatly with preservative models of episodic memory—models that understand episodic memory as being a strictly preservative cognitive capacity—because the causal condition implies the storage, or preservation, of information. This, however, does not mean that a causal condition, or memory trace, is incompatible with other models of episodic memory. It's possible, for example, to maintain something similar to the causal connection suggested by Martin and Deutscher, and posit a model of episodic memory that is not strictly preservative. This point will come up again in chapter four (section 4.4), but it's worth making this distinction clear.

Another important point to consider regarding Martin and Deutscher's account is that it captures common sense ideas about episodic memory so clearly because their analysis relies primarily on intuitive judgements about what qualifies as a memory. Through their analysis, they establish the criteria of memory such that it admits the cases that would intuitively be classified as memories, and rules out the cases that wouldn't. Thus, the causal account of memory matches up with conventional ideas about episodic memory. This kind of conceptual analysis works very well in so far as we expect it to accurately establish what constitutes the concept of memory, but when it comes to applying this concept of memory to the underlying cognitive system that generates memory, intuitions can't have the last word.

In this case, however, where we're talking about memories—a central construct in our mental lives—intuitions are a respectable place to begin. They do, after all, track common sense, but they are certainly not infallible. And, what common sense would tell us about the process of remembering is that there must, in some way, be a kind of causal connection such that a past experience can be represented at a later time, with at least a sufficient degree of accuracy. But even though Martin and Deutscher's causal theory is proposed as a conceptual analysis of the verb "to remember" and our ordinary concept of memory, their causal condition can also be taken to be an empirical hypothesis about the capacity of memory. In what follows, in discussing the causal condition and the closely related notion of a memory trace, I will mainly be concerned with them as theoretical posits in a naturalistic theory of memory, rather than merely as elements in a conceptual analysis of the ordinary concept of remembering.

This concept of a causal connection has become the dominant view in philosophy of memory. Without positing a causal connection, it's difficult to explain how a past experience, that no longer exists, could be represented at a present time. This notion is precisely what makes the causal condition of memory so forceful—that it corresponds with our intuitions about how memory must work.

We'll discuss this idea of a memory trace more in the following section.

1.4.2 Memory trace

In their explanation, Martin and Deutscher are deliberately non-committal on exactly what constitutes such a memory trace. They explain that they don't need to understand the neurophysiological nature of such a trace in order to claim that it must be present: it suffices to rely on a simple structural analogy—like the imprint of a coin in wax, or the gramophone example discussed above (Martin and Deutscher 1966, p. 190). They appeal to the common-sense notion that in-order for something that was in the past, to be represented now in the present, there must be this causal connection, or memory trace.

Although Martin and Deutscher are credited with establishing this view, De Brigard (2014, p. 402-3) notes that the idea of a memory trace has a long history in philosophy, going back to as early as Plato and Aristotle. He points out that the arguments in favour of such a memory trace usually take the form of an inference to the best explanation: we can't seem to make sense of the idea of a past experience being represented in the present without appealing to this kind of trace.

As Robins points out, the notion of a memory trace implicitly associates memory with a sort of storage and retrieval function, which she characterizes as the "archival" views of memory: our memories are stored away, like books in a library, and retrieved when we call a memory to mind (2016b, p. 435-6). According to Robins the cognitive level at which the memory trace occurs varies according to different theories: it can be either personal, such that the memory is like an image held in the mind's eye; or sub-personal, such that it is a cognitive mechanism (2017, p. 77-

8). Furthermore, the memory trace can be understood to be static, like a fixed image, or dynamic, like a reconstructive process.

According to Robins the notion of a memory trace serves to explain four aspects of memory (2017, p.79-85). There is the need to *represent the past*: this idea we've already outlined, and is essentially the common-sense intuition that in order to represent a past experience, which no longer exists, at a present time, requires some recourse to a memory trace (Robins 2017, p. 79-80). There is the notion of *preventing causal gaps*, which was seen in the discussion of Martin and Deutscher—a memory trace is required to avoid deviant causal chains that can make past representations available in the present, without intuitively qualifying as memory. Another reason is to *distinguish memory from relearning*—in this context, *relearning* refers to cases where an individual forgets information they had acquired in the past, and then comes to re-acquire it from a different source (Robins 2017, p. 82). Consider, for example, the case of writing down an experience in a diary, forgetting it, and then reading the diary years later and coming to recollect it. This is importantly different from a case where one can recollect the experience without having to consult their diary. In order to differentiate these two cases, however, requires appealing to a memory trace. And, finally, there is *making sense of memory science*. The idea of a memory trace features in memory science, and discoveries in the empirical study of memory appear to support this idea (Robins 2017, p. 83). Specifically, the Encoding-Storage-Retrieval (ESR) model is the working model for studying the memory process. On this model of memory, *encoding* refers to the process through which experiences are attended to and processed for storage; *storage* refers to the process through which this information is retained, and *retrieval* refers to the process through which these stored memories are cognitively retrieved at a later date.

In certain contexts, the term *engram* is used to describe a memory trace, specifically regarding its neurophysiological constitution. In the path towards identifying the neural location of the engram, De Brigard notes two significant discoveries in the 20th century (2014, p.407-11). The first of these was in 1957, where a study by Brenda Milner on a patient H.M. revealed that the hippocampus was necessary for creating new conscious memories. And the second of these was the first description of long-term potentiation (LTP)—the discovery a neural mechanism that

could preserve the effects of a stimulus (De Brigard 2014, p. 408)—made by Bliss and Lomo in 1973. These two discoveries offered ways of identifying memory traces: 1) in terms of the neural substrate as a system (using fMRIs and ERPs) and 2) in terms of the neural substrate at a cellular level (De Brigard 2014, p. 408). These discoveries substantiated the concept of a memory trace, or engram, on a neural level.

In this section, we can see how Martin and Deutscher's idea of a causal connection, or memory trace has become so enduring. Not only does it correspond with traditional intuitions about how memory works, but it gains strong support for other reasons. Notably, those surveyed by Robins (2017): that it is needed to represent the past, prevent causal gaps, distinguish memory form relearning, and, possibly most telling, that it corresponds with scientific models of memory.

1.5 Summary

This chapter has been a basic overview of episodic memory, outlining the concept of episodic memory, and detailing how it is featured in human beliefs and behaviours.

In the first section, I began by defining the concept of episodic memory, providing a brief taxonomy of memory in general, and looking at specific intricacies related to episodic memory. Of particular importance is the distinction between semantic and episodic memory. These two systems of declarative memory are distinguished by the phenomenology and content of their cognitive state: episodic memory takes the form of recollection of first-person experienced events, and semantic memory of the recollection of facts and propositions. I made the case that there are good reasons to support this distinction. Notably, there are theoretical reasons for treating them as different cognitive, or natural kinds, and there are empirical reasons for thinking that they refer to distinct cognitive systems: there are studies on Alzheimer's and semantic dementia that support a double disassociation between the two systems.

In the second section, I discussed how common sense, or folk, concepts about episodic memory are featured in everyday human beliefs, behaviours, and attitudes. In particular, ideas of personal knowledge, personal identity, agency and testimony all rely on these folk ideas about episodic

memory. An important feature of these beliefs and practises involving episodic memory is that they hold the function of episodic memory to be that of representing past experiences, and that implicit in this understanding is the assumption that, when properly functioning, these representations are accurate and authentic representation of past experience.

In the third section, I addressed how ideas about episodic memory have been treated in philosophy. Theories of remembering have focused on two important distinctions: differentiating episodic memory from sematic memory, and differentiating episodic memory from episodic imagination. The first of these is an important issue, despite the generally accepted distinction between autonoetic and noetic consciousness, which we will return to in chapter four. The second is closely related to the central aims of this thesis, and will be addressed throughout the dissertation. An important theory in terms of addressing this second distinction is Martin and Deutscher's proposal of a causal criterion. Their account has been the dominant theory of remembering in philosophy and cognitive science. A strength of the causal theory of memory is that it fits well with common sense ideas about how memory works: the idea of a causal connection, or memory trace corresponds with the traditional model that episodic memory represents past experiences, and that these representations are accurate and authentic representation of past experience.

While the idea of a memory trace fits well with our intuitions about how memory works, there are other reasons for accepting this idea as well. In section 4, I outline the concept of a memory trace in detail, and suggest other reasons, beyond the intuitive appeal, that support this concept. Probably the most significant of these is that the idea of a memory trace corresponds with scientific models of how memory works, and that these models have been corroborated by empirical evidence.

Altogether, this shows that episodic memory is a well-established concept that refers to a distinct cognitive system. According to traditional models, episodic memory functions to represent past experiences, and there is an implicit assumption that these representations are accurate and authentic representations of past experience. These models feature prominently in human beliefs

and behaviours. The causal theory of memory, and the idea of a memory trace, accurately portray these traditional models of memory, and these are the dominant models in the cognitive sciences.

1.6 Moving forward

At the beginning of this chapter, I outlined a central debate in philosophy of memory: whether the function of episodic memory is broadly preservative or generative. Much of this chapter was a review of traditional concepts of episodic memory and an account of the causal theory.

The question of whether memory is broadly preservative or generative results, primarily, from the study of certain kinds of memory errors. In *chapter two*, I will discuss different kinds of memory errors in more detail. I'll look at different cases of memory errors, some of which are problematic for causal theory accounts and others that aren't. This will give us a better understanding of why the problematic cases of memory errors are so. I'll also explain the dilemma these cases of memory errors creates, and the challenges faced on either side of this issue.

Of the two sides of this debate, I think the more traditional, hybrid, views, offer a better account of episodic memory. In *chapter three*, I will assess how both sides of the dilemma propose dealing with problematic kinds of memory errors. Here, I will argue that revisionist accounts face more difficulties than those that maintain the causal connection.

Effectively, this debate comes down to a question of the function of episodic memory: what does our episodic memory system do? In *chapter four* I will assess the function of episodic memory in more detail. Here, I will outline some contemporary accounts of the function of episodic memory. Ultimately, I will be arguing that episodic memory functions as a broadly preservative cognitive capacity.

Chapter 2: Challenges of Memory Errors

At the outset of this dissertation, I outlined a central debate in contemporary philosophy of memory—that is the question of how to properly understand the function of episodic memory. If you recall, there are two main positions in this debate: the revisionists, who want to move away from traditional models of episodic memory (De Brigard 2013; Michaelian 2016b), and the traditionalists, who want to preserve something resembling the causal theory of memory (Bernecker 2004; Robins 2016b). This debate arises, largely, because of a growing body of evidence in psychology that speaks against the reliability of episodic memory. This evidence consists of studies in cases of misremembering and false memory (Deese 1959; Roediger and McDermott 1995; Loftus 2003). These studies provide the backbone for arguing that causal theory accounts of memory, which emphasize the role of a memory trace, and so correspond to a broadly preservative theory of the function of memory, are unable to account for certain kinds of memory errors, and so provide an inadequate account of memory.

The line of question that motivates these revisionist positions follows: if these cases of misremembering and false memory are as prevalent as these studies suggest, then should this evidence be taken as grounds for rethinking the function of episodic memory—that is to say, does episodic memory really function to preserve personal past experiences, both accurately and authentically? If so, then how can we account for the prevalence, and systematic nature, of the types of memory errors that these studies reveal?

In order to gain a better understanding of this line of reasoning and to understand the significance of this evidence in detail, we will need to look more closely at these studies. While it's clear that these studies identify prevalent and consistent patterns in the function of episodic memory, how we classify these patterns, whether they are understood to be errors, and the extent to which the prevalence of such errors undermines the functional account of the causal theory, remains unclear. Of course, to fully appreciate the extent to which these cases undermine our understanding of episodic memory requires that we first establish a functional theory of episodic memory, something that will be the focus of chapter four. For the purpose of this chapter, however, we can proceed by tentatively assuming that episodic memory is broadly concerned

with preserving fairly accurate representations of past experience. This position is consistent with causal theory accounts of memory and will allow us to evaluate and organize the different kinds of memory errors, which will be the focus of this chapter.

In the first section, I will begin by discussing some of the more common, everyday kinds of memory errors. This should allow for a better understanding of why certain kinds of memory errors are consistent with the causal theory, and others aren't. In the second section, I will introduce the more problematic cases of memory errors. These include errors revealed in the DRM paradigm, cases of false memories, and general memory distortions. In the third section, I will analyze the significance of these problematic cases of memory errors with regards to both traditional, causal theory accounts of memory, and revisionary accounts of memory. Finally, in the fourth section, I will provide a summary for this chapter.

2.1 The causal theory and memory errors

When I say that the prevalence and consistency of certain types of memory errors stands to undermine the functional account of episodic memory provided by the causal theory, it should be made clear that I am referring to specific types of memory errors, and not just any old memory error. Namely, these memory errors are well-documented, predictable errors, the most notable of which are the DRM paradigm and the kinds of false memory cases seen in the Loftus experimental paradigm.

Concerning memory errors in general, there's nothing inherently problematic for a biological system, like episodic memory, to malfunction. We typically expect that over time, and under certain circumstances, our memory systems will suffer from a variety of malfunctions. I don't expect, for example, that I should be able to vividly recall the experience of every meal I've had over the past year, or that I can consistently recall the finer details of past experiences on cue. But, if we take the function of episodic memory to be the comprehensive preservation of completely accurate representations of past events, then these types of cases qualify as malfunctions. Nonetheless, we accept these kinds of "errors" without taking them as reasons for

reevaluating our ideas about how episodic memory works. So, what is it about certain kinds of memory errors that challenges the causal theory's understanding of episodic memory?

Before we look at the more problematic kinds of memory errors, let's consider some of the kinds of memory errors that don't create a problem for the causal theory's account of episodic memory. This will give us a better understanding of how the theory accommodates these errors, and the important differences between common kinds of memory errors and the more problematic cases discussed in the literature.

2.1.1 Common memory errors

While certain kinds of memory errors are a central feature in this debate about how to understand episodic memory, there are many other kinds of memory errors that are well-understood but haven't tended to create a problem for causal accounts of memory.

The fact that I can't remember the experience of eating a specific meal last year, although perhaps upsetting, isn't entirely shocking. Over time, memories fade and, even though some memories fade from our minds more rapidly than others, even the most enduring memories suffer from the passage of time. This is known as *transience*: the decreasing accessibility of memories over time. First studied by Ebbinghaus in the 1880's, he noticed that mere hours after memorizing nonsensible strings of syllables, his retention had significantly decreased (Schacter 2001, p. 13). Transience is a well-documented memory error that has been studied for over a century, and probably one that people have been familiar with for much longer than that, and yet the prevalence of this memory error hasn't led researchers to reject the causal theory of memory.

Another common memory error occurs when a person is certain that they have a specific memory, but can't quite bring the corresponding information, or experience, to mind. This type of memory error is known as *blocking*. Blocking is generally associated with the recall of names and is particularly apparent in the retrieval of proper names (Hanley, Kay 1998). For example, you remember that a person's name begins with the later "J", which brings up the name "Josh", and causes you to block on the actual name, "Justin". While blocking on the names of people and

places constitutes a semantic memory error, the phenomenon of blocking has also been observed in episodic memory. A series of studies had participants perform basic tasks (hammering a nail into a piece of wood, or pointing out a country on a globe). Some subjects were then shown photographs of themselves performing these tasks. These subjects experienced improved recall of the activities reviewed in the photographs, but suffered worse recall of activities that were not reviewed than control groups that didn't review any photographs (Koustaal et al. 1999). The suggestion, here, is that the retrieval of certain memories (the photographed tasks) has caused participants to block on other tasks (those that were not photographed). Blocking has been extensively studied, both as it relates to semantic memory, as well as episodic memory, but hasn't been suggested as a reason for rethinking the function of episodic memory.

Another, frustratingly common, kind of memory error has to do with the weekly, or daily, misplacing of mundane objects and the forgetting of routine tasks. You're hurriedly returning to your house to get the hat that you'd forgotten and realize you've misplaced your keys. You frantically search the house, only to realize that you'd left your keys in the front door. *Absent mindedness* occurs when a memory either isn't properly encoded to begin with or was missed in the retrieval process (Schacter 2001, p. 41-60). A common cause of absent mindedness is a lack of attention. In the above scenario, you're so preoccupied with quickly retrieving your hat that you fail to properly encode a memory of leaving your keys in the door. Studies have shown that divided attention during encoding significantly increases the likelihood of absent mindedness (Baddeley et. Al 1984; Craik et. al 1996). Although you might, technically, be experiencing certain things, if you aren't paying attention to them, it is unlikely that they will be available in memory. Again, absent mindedness is a well-documented and commonly experienced memory error, but it doesn't present a problem for a causal theory account of memory.

There are, of course, many other kinds of memory errors, but the three described above—transience, blocking, and absent mindedness—are exceedingly common, and predictable, memory errors. Likely, most people are familiar with all three of these, and each has been the subject of extensive study. If the issue were just a matter of frequency and predictability, then these memory errors would be just as problematic for the causal theory's account of memory as

the kinds of memory errors that feature in this debate (the DRM Paradigm and false memory). But, as I will go on to show in the following section, they aren't.

2.1.2 Explaining common memory errors

The main reason these common kinds of memory errors aren't understood to be problematic is because they are compatible with a version of the causal theory that combines a causal account of memory with the three staged model of memory used in memory science, the encoding-storage-retrieval model of memory.

The Encoding – Storage – Retrieval (E-S-R) model of memory is the working model for memory used in memory science. According to the E-S-R model, remembering involves three stages, or processes. This first of these, *encoding*, is where information is taken in and processed by the memory system. *Storage* is the process of maintaining the memory encoding over a period of time. And, *retrieval* is the process through which these stored memories are then made consciously available when a person is engaged in remembering. These three stages correspond to the idea that a memory is stored in the mind via a memory trace. The trace represents the content that the memory system encodes, stores, and then retrieves (Schacter 2001, p. 138). During the process of remembering, this trace is called to mind.

If we look back over these three kinds of memory errors, we can see that each corresponds with a failure in one or more of the processes outlined by the E-S-R model. Transience is caused by the degrading of the memory trace over time. This is a failure of the *storage* process of memory. And it is explainable since we can recognize that the human mind has a finite storage capacity. Over time, a working memory system will continue to store new memories, and this inevitably impacts the storage of older memories.

Blocking occurs when the retrieval of one memory, or certain features of a memory, interferes with the retrieval of other memory content. This is a failure in the *retrieval* process—the memory has been encoded, and stored, but upon retrieval, the memory system fails to produce the corresponding representation. Cognitive processes, like those involved in the memory system,

are not infallible and under certain constraints (like remembering a name or experience on cue), do not always perform optimally. What's more, in some cases retrieval can be cue-dependent, and the retrieval of some memories can only be retrieved given certain types of cues—without the requisite cue, the memory system is unable to retrieve the corresponding memory.

Absent mindedness occurs when an experience isn't properly stored in memory. This is often a failure in the *encoding* process (though it can also result from cases of failed *retrieval*). Because the individual experiencing absent-mindedness wasn't attentive to certain features of an experience, they failed to properly process those features of experience as memories. This makes sense if we consider that humans have a limited capacity for attention, and that the memory system has a limited capability of processing experiences into memory—when parts of an experience get left out during encoding, absent mindedness occurs.

So, while all of these memory errors are prevalent and predictable, they don't present a problem for a causal theory of memory because they correspond neatly to the stages laid out in the E-S-R model of memory and are consistent with an account of memory that serves, broadly, to encode, store, maintain, and retrieve representations of experiences in the form of traces, although it might not do so maximally or comprehensively. A theory needs to be able to account, not only for the successes of a system, but also for its failures. The important point, here, is that these kind of memory errors are explainable according to a causal theory account of memory. The memory errors that feature prominently in this debate (the DRM paradigm and false memories), however, create a particular problem for causal theory accounts of memory in this respect. In the next section, we'll look at these memory errors in more detail.

2.2 Problematic memory errors

In the previous section, we looked at common types of memory errors that didn't present a problem for causal theory accounts of memory. We saw that an important reason for why these kinds of memory errors don't present a problem is that they fit within a causal theory account of how memory works. In other words, they are easily explained by a scientific model (the E-S-R model) that is consistent with the causal theory. The more problematic varieties of memory error

for causal theories to accommodate are problematic for precisely the opposite reason: they are not easily explained using a causal theory framework. That is not to say that causal theory accounts of memory cannot accommodate these errors—most contemporary proponents of the causal theory allow for some constructive processes to accommodate these kinds of errors (which we'll touch on in more detail in section 2.3.1)—but simply that they present a challenge for a strictly preservative interpretation of the causal theory.

There are several varieties of memory error that fall within this not-easily-explained-by-the-causal-theory category, but the most frequently discussed cases in the literature are the DRM paradigm and cases of false memory, or confabulation. Other examples include memory distortions like boundary extension (memories in which the periphery is extended), field-to-observer perspective (in which memories are viewed from a third person perspective), and the telescoping effect (in which the proximity between memories is misjudged).

We'll look at the specifics of these memory errors in detail and the experiments used to confirm them.

2.2.1 The DRM Paradigm

The DRM paradigm refers to an experimental technique that has been shown to consistently generate a particular kind of memory error. The technique was originally tested by Deese (1959) and his results were later confirmed and made more widely known by Roediger and McDermott (1995)—hence, the DRM paradigm. In these experiments, subjects mistakenly identify items as being on a previously shown list when those items have a strong association with list items. The lists initially consisted of words, but experiments have since reproduced the effects with pictures, faces and dot arrays (Robins 2016b, p.434). A standard variation of the experiment follows:

Experimenters provide subjects with lists of words, each of which is associated with a non-represented, "critical", word. So, for example, if the critical word was "*spider*" the list would contain words that would be closely associated with that word: *web*, *insect*, *bug*, *fright*, *crawl*, *small*, etc. Importantly, the critical word from which each list is drawn (in this case *spider*) does

not appear on the actual list. The lists are ordered such that words with the strongest association with the critical word (e.g., *web*) appear earlier on the list, while words having weaker associations with the critical word (e.g., *small*) appear later. After hearing the lists, subjects are then given another list, containing a mixture of words from the original lists (e.g., *insect*), and lures (words that did not appear on the original lists). The lures include the critical words and words that neither appeared on the lists nor were associated with the critical words (e.g., *sandwich*). Subjects are then asked to identify whether each word was from one of the original lists, or if it was a new word. If they identify the word as being from an original list, they are asked to further specify whether they actually r*emembered* the word as being from the list (had the phenomenological experience of hearing/seeing the word), or simply *knew* the word was on the list (remembered the word from the list but didn't have the corresponding phenomenological association with having heard/seen the word).

The second step in this memory recognition test, the remember-know distinction, was added to Deese's original experimental technique by Roediger and McDermott (1995). They based this on Tulving's remembering/knowing experimental technique (1985). The purpose of including this second step of having subjects self-report on their state of awareness was to determine whether false recall on the test was actually testing false memory, or whether these instances of false recall could be explained simply by having activated associative networks (Roediger & McDermott 1995, p. 804). It was reasoned that if subjects were simply picking up associations with the critical lure, then they wouldn't have the corresponding phenomenological experience of *remembering* the word on the list.

Results from the studies showed that subjects' recognition of list words from the original list varied in accordance with their positioning. Words that were more closely associated with the non-represented critical word, those that appeared earlier on the list like "*web*", had the highest recognition rate (70-80%). Words that appeared in medial positions on the list, so with a lower association with the target words, like "*crawl*", had a lower recognition rate (40-50%). Regarding the lures, lure words with no association with the target list words, like "*sandwich*", had the lowest recognition rate (<10%). Surprisingly, non-represented critical words like "*spider*" had a recognition rate comparable to the recognition rate of strongly associated list

words (70-80%), which is significantly higher than the recognition rate of the weaker associated list words, like "*crawl*". Basically, subjects falsely recognized the non-represented critical words more frequently than they remembered many words that appeared on the original lists, and at a similar rate to list words with the highest recognition rate. What's more, subjects reported that they actually *remembered*, as opposed to merely *knew*, the non-represented critical words 72% of the time, which is, again, comparable with the rate at which they reported remembering strongly associated list words.

What the DRM paradigm is taken to show is that the memory system consistently and predictably generates false memories. If we accept the remembering-knowing distinction, then it looks like in these cases, subjects are having false memories of words being on the list that were never actually there. To these subjects, the memory of hearing "*spider*" and of hearing "*web*" would be indistinguishable, with the interesting distinction being that they *did* hear the word "*web*" but *didn't* hear the word "*spider*". This is troubling for causal theory accounts of memory because it does not seem to be explained by the theory as previously sketched out. In these cases, the episodic memory system is producing a representation of an experience (hearing the word "*spider*") that was never actually experienced.

We'll look at the significance of these experiments in more detail, but first let's look at another, perhaps even more troubling, series of experiments on false memory.

2.2.2 False memory cases

On October 4th, 1992, an EL Al cargo plane suffered from engine failure shortly after taking off from Schiphol Airport in Amsterdam. The pilots couldn't successfully land the plane, and it ended up crashing into an eleven-story apartment building, killing all four crew and thirty-nine residents. The tragedy was heavily televised and featured in media in the Netherlands. Ten months after the plane crash, a team of Dutch psychologists questioned the memory of people from their university. Fifty-five percent of respondents admitted to seeing footage of the planes crash, and in another interview, this number went up to two thirds. They described details like

the angle of the plane upon hitting the building, the appearance of the plane heading into the crash, and the subsequent wreckage. Here's the kicker: there was never any footage of the crash[6].

False memories gained public attention towards the end of the 20th century, with cases of false eyewitness testimony leading to wrongful legal convictions, and psychiatry sessions intent on uncovering repressed memories turning out instead to produce wildly inaccurate memories (Loftus 1997). In cases like the plane crash, the team of psychologists deliberately used leading questions to illicit false memories in respondents. Persistent police and legal interrogation can, however, unintentionally solidify false memories in eyewitnesses. And the constructive imagination techniques used by some psychiatrists, can end up having patients come to believe that fictional imaginings actually happened.

Loftus describes an experiment the goal of which was to implant a specific memory in subjects (1997). The memory was of being lost in a shopping mall as a child. To implant this memory, experimenters asked subjects to recollect childhood events. Then, enlisting the help of a relative or partner, experimenters created a booklet for each subject containing three stories about events that the subjects had actually experienced, and one that they hadn't (the false memory). Subjects were then asked to read the booklet and write down what they remembered about each event, or if they didn't remember the event, to make note of it. Results showed that 68 % of the subjects remembered information about the events that actually happened; and 29% of the subjects falsely recalled some details of the fictional event. Loftus notes that there were some discrepancies between the descriptions provided for actual events and false memories, such as they used more words describing their actual memories and rated these memories as clearer (Loftus 1997, p.72). Still, even with these discrepancies, having over a quarter of the subjects report false memories is significant.

Hyman et al (1998) conducted a similar experiment with similar results. They found when subjects were initially interviewed after being presented with the details of a fictional memory, none of the subjects falsely remembered the fictional event. Later, in a second interview, this number was up to 18%, and by a third interview, up to 25%. In this case, it would appear that

[6] Example taken from Schacter, *Seven Sins of Memory* (2001, p.112-3).

rehearsing details of the false event in subsequent interviews, over time, increases the likelihood of subjects developing a false memory of the fictional event.

Loftus (2003) notes that several features have been shown to contribute towards the creation of a false memory. Having a confederate (a family member or someone close to the subject working with experimenters) duplicitously confirm that the false event did, in fact, occur, increases the likelihood of subjects developing a false memory. Also, exercises in guided imagination can contribute to creating a false memory (Loftus 1997, p. 74). In another study, doctored photographs, with an image of the subject superimposed onto the scene of a hot-air balloon ride, succeeded by the third interview in having 50% of the subjects remembering something about this fictious hot-air balloon ride (Wade, Garry, Read, Lindsay 2002). In these cases, having false, corroborative, evidence, contributes to creating false memories.

While the experimental results in these studies aren't quite as widespread as those for the DRM paradigm, they are nonetheless impressive due to the scope of the false memory. In these cases, the false memory of an entire event is being created. And in many instances, subjects are highly confident in the veracity of this false memory, going on to add to the content of the false memory, and embellish the details in personalized ways (Loftus 2003, p. 232).

Just like the cases produced by the DRM paradigm, these cases of false memory are problematic for causal theory accounts of memory. If memory is understood to function to store and preserve the content of experienced events, as the causal theory suggests, then how can these cases of false memory be explained? The events in question never even occurred, let alone were experienced by the subjects that came to *remember* having experienced them.

This is a problem that causal theories face when confronted with these cases of memory errors. In the next section, we'll look at how different accounts of memory address these sorts of errors.

2.3 How to treat memory errors

In the previous section, we outlined two significant kinds of memory errors—the DRM paradigm cases, and cases of false memory—and saw that they were problematic for causal theory accounts of memory for similar reasons. In both cases, these memory errors involve subjects coming to falsely remember content that they never actually experienced.

Schachter groups common kinds of memory errors into two broad categories that are helpful for making sense of the difficulty faced by causal theory accounts (2001, p.4-5). Errors of *omission* occur when content from the initial experience is left out. Errors of *commission*, on the other hand, occur when content is added to the content available from the initial experience in memory. Effectively, some kinds of memory errors involve leaving out information that should have been included in memory, and other kinds of memory errors involve including information that was never present in the initial experience.

If we look back over the types of memory error that are easily explained by the causal theory, we can see that they fall squarely into this first category—they are errors of omission. Cases of transience, blocking, and absent mindedness all involve features of an initial experience being left out in later memories of that experience. The problematic cases from the DRM paradigm and false memory, however, involve the addition of new content in memory—they are errors of commission. In these cases, new content, other than that which was initially experienced, is being made available in memory. Explaining how this new content comes to feature in memory is where causal theory accounts face a challenge.

The problem for causal theories in this respect is that, according to a classical model of the causal theory, memory is a mainly *preservative* capacity. That is to say, memory functions solely to store and retrieve representations of the experienced content of past events. This goes back to Martin and Deutscher's idea that a memory trace should be characterized as a structural analogue of the experience (1966, p. 190). Accordingly, the memory trace can only contain a representation of information that was available in the initial experience and, importantly, no new information can be represented—this is not to say that in the process of remembering, new

information (i.e. context) cannot be included, but simply that the trace itself is a direct structural analogue of the experience, and doesn't contain any information beyond what was made available from the initial experience. This is the problem, then, for an account of memory that treats memory as a purely preservative capacity: how do we make sense of content being added, or generated, in episodic memory when our theory doesn't allow for content creation. While this is sometimes treated as a problem for accounts of memory that maintain a causal connection, it is more precisely only a problem for those accounts that understand the function of memory as being purely preservative. In other words, it is the preservative models of memory, which posit a memory trace, but not the basic idea of a memory trace, that run into problems.

This dilemma—how can a purely preservative memory system account for new content in memory—is responsible for the two positions in the debate. On the one hand, traditionalists want to maintain that memory is still essentially a preservative capacity, with the difficulty now being how to account for these errors of commission. And, on the other hand, revisionists argue that memory should no longer be understood as preservative at all, indeed some would deny a distinct faculty of episodic memory. This plays out differently for both sides of the debate.

2.3.1 Traditionalist dilemma

For the traditionalists, who want to maintain that memory is still largely a preservative capacity, the difficulty is how to make sense of these kinds of memory errors—the DRM paradigm and cases of false memory—within the causal theory framework.

According to a classical interpretation of the causal theory, as laid out by Martin and Deutscher (1966), this presents a genuine problem, at least if the causal theory is understood as an empirical hypothesis about the capacity of episodic memory. Particularly, if we want to maintain that the memory trace is a structural analogue in the strict sense. In this respect, the prevalence and consistency of the kinds of memory errors resulting from the DRM paradigm and false memory studies show that memory cannot be strictly preservative. Memories are not static representations of experience that are stored away in the mind, available for recall.

This, however, does not necessarily mean we should abandon the idea of a memory trace altogether. Just because the process of remembering often involves more information than what was presented in the initial experience, doesn't mean that the memory trace isn't central to memory, or that there isn't a memory trace at all. There are several ways of accommodating errors of commission within the causal theory: one is to allow for some constructive processes such that the memory trace can be supplemented with additional content. Another, is to allow that the memory trace is not a fixed structural analogue in the strict sense outlined by Martin and Deutscher. A third option would be to deny that the mental states brought about in the DRM paradigm and Loftus cases are memories at all. Instead, they could simply be fake memories, which would mean that we don't have first person authority over our own memories—what, phenomenologically, seem to be memories, are not always such.

Of these three possibilities, most contemporary advocates of the causal theory advance something in-line with the first option. They accept that memory is primarily preservative, but allow for some constructive, or generative, features. Bernecker allows for a form of what he calls *moderate generativism* in the preservation of a memory trace—the content of memories can be selective, and can be updated, but no new content can be added (2008, p. 164). Before taking a more revisionist position, Michaelian argued for a generative account of memory that maintains the idea of a memory trace but includes constructive features of memory—memory might be stored via a *gist* trace, for example, that provides a constructed schema of the initial experience (2011). Robins argues for a middle ground between traditional "archival" views on the one hand, and modern "constructive" views on the other (2016b, p.443-5).

Technically, then, most contemporary proponents of the causal theory argue for what are known as *hybrid* theories of memory. These theories combine the preservative features of the causal theory but include constructive features as well. While the general direction of these ideas is promising, determining the exact relation between the preservative and generative features of memory presents a challenge. Bernecker's *moderate generativism*, for example, is still too weak to accommodate the cases involved in the DRM paradigm and false memory studies. Although the selecting and updating of memory allowed for by moderate generativism might account for some kinds of memory distortions, Bernecker's clause explicitly denying new content effectively

makes his account less accommodating, since the specific memory errors in question involve new memory content (Michaelian 2011, p. 336). The problem with *gist* type schematic accounts of memory is that they have difficulty accounting for the accuracy of memories. On these views, gists are generalized schematic representations created from the content of past experiences. The challenge, here, is that if memories are typically stored as gists, then how can we explain the cases where remembering yields finer grained details than gist structure predicts (Alba Hasher 1983)? When we recall the specific details from a past experience, for example, it seems as though the corresponding representation must be more detailed than a mere schema. Robins (2016b) suggests we need to incorporate constructive features of memory into the causal theory in order to account for these kinds of memory errors. To this effect, she proposes a useful taxonomy of memory errors, which we'll discuss in detail in chapter three. Her account, however, focuses more explicitly on the classification of memory errors, and still leaves open the issue of reconciling the preservative and generative features of episodic memory.

Overall, then, hybrid views struggle with resolving this discrepancy between reconciling the preservative and generative features of memory.

2.3.2 Revisionist dilemma

Revisionists take on the other side of this dilemma. They move to dispense with the preservative understanding of memory, instead focusing on the generative features of memory.

De Brigard argues that memory is better understood as part of a system of *episodic hypothetical thinking*—a cognitive system that deploys first person inferences about past and future possible scenarios (2013, p. 174-5). On this view, episodic memory, along with episodic imagination and episodic counterfactual thinking, are all part of this system of episodic hypothetical thinking, which is wholly generative. Michaelian (2016a) argues that by applying a reliabilist framework to memory, we can effectively dispense with the causal condition put forward by Martin and Deutscher altogether. Michaelian (2016b), later argues for a view of *mental time travel* that is more in sync with De Brigard's idea of episodic hypothetical thinking. According to these views, memory should not be understood as a preservative capacity at all.

While these revisionary accounts propose a complete reconceptualization of how episodic memory works, dispensing entirely with a causal condition, or memory trace, isn't a cheap solution. Similarly to the problems associated with gist-type schematic views, dispensing with a memory trace makes explaining the accuracy of memory much more complicated. It's a bit of a case of throwing the baby out with the bathwater, so to speak. As we saw, there are important kinds of memory errors—transience, blocking, and absentmindedness—that are successfully explained by reference to a memory trace. What's more, explaining why certain outcomes succeed without recourse to a causal connection becomes mysterious. Without positing, at least theoretically, a memory trace, something as simple as returning to your car in a crowded parking lot becomes difficult to explain. A causal connection of some kind is needed to make sense of such successful behaviour, otherwise we would be left with a causal gap (De Brigard 2020; Khalidi 2023). So, effectively, the proposal to forgo the explanatory power afforded by a memory trace with regards to both the successes and certain failures of memory, in favour of better explaining other kinds of failures, isn't a clean fix.

This, mind you, is provided that these revisionary views can actually provide better explanations of the DRM cases and false memories. Robins points out that explaining the DRM paradigm is a particularly complicated matter because the results support both preservative and generative features of memory (2016b, p. 434-5). The DRM cases are interesting because subjects are undoubtedly retaining some memory from the original lists. Otherwise, we wouldn't see the high degree of correlation between critical lures and strongly associated list words. Accordingly, these are not cases of wholly fabricated memories, but memories that clearly retain most of the important information from the initial experience. Explaining this without a memory trace becomes difficult.

Similarly in the cases of false memory described by Loftus. While it might be striking that over 25% of the subjects ended up coming to possess false memories, if we dispense with the idea of a memory trace and suggest instead an inference-based system of hypothetical thinking, explaining why almost three quarters of the subjects *didn't* come to have false memories looks even more baffling. If we expect our memory system to process the likelihood of past events and

develop conscious episodic simulations accordingly, then shouldn't strong evidence, along with the testimony of trusted relations, suffice to illicit such a memory? In other words, if memory does function to generate plausible information about the past and future based on background information, testimony, and perception, etc., then we would expect that most, if not all, of the subjects in the Loftus paradigm cases would endorse the false memories. The fact that after being exposed to such thoroughly contrived false evidence and testimony, only 25% of the participants came to recall false memories, suggests that if this is, indeed, how memory works, then it doesn't do so very well. Incorporating a memory trace in the explanation, however, makes this discrepancy much less striking.

Michaelian (2016a) argues against Robins (2016b) on this point. He claims that just because a system of episodic memory doesn't need to posit a memory trace, this isn't to say that memory can't, and doesn't often, employ one. Proponents of episodic hypothetical thinking are free to claim that memory often involves a memory trace, and that some kinds of memory errors are always explained by reference to a memory trace (Michaelian, 2016a). This characterization of the issues, however, might be overly simplified. If we accept that certain kinds of memory errors always involve a memory trace, then presumably the same can be said of certain kinds of episodic memories—that they always involve a memory trace. If this is the claim, then a proponent of the causal theory needs simply to say that it is precisely those cases of mental representations, that always involve a memory trace, that constitute episodic memories. Other mental representations that do not involve a memory trace might be other things, they might be memory errors, they might be imaginings, but they are not episodic memories.

So, while revisionists in this dilemma maintain that episodic memory does not necessarily involve a memory trace, the explanatory value of a memory trace is something they still need to contend with.

2.4 Summary

In this chapter, we've reviewed several important kinds of memory errors to determine how accounts of memory, both traditionalist and revisionist, can deal with them.

Common types of memory errors include *transience*, the decreased accessibility of memories over time, *blocking*, where retrieval of a given memory is inhibited, and *absent mindedness*, where memories aren't properly encoded. These kinds of memory errors are particularly prevalent and have been the subject of extensive study. Nonetheless, they don't create a problem for traditional causal theory accounts of memory. This is because they can be explained by the ESR (encoding storage retrieval) model of memory—the leading model for memory used in memory sciences. Causal theory accounts successfully accommodate these three processes with a causal component (the memory trace) to explain these instances of memory error.

The more problematic kinds of memory errors for causal theories to explain are cases of misremembering and false memory. The main examples of these in the literature are the DRM paradigm and the Loftus false memory paradigm. These experiments show that episodic memory consistently and predictably generates misrememberings and false memories. These kinds of memory errors are problematic for causal theory accounts because they involve the generation of new content—content that wasn't present in the initial experience—in memory. Schachter classifies memory errors into two broad categories: errors of *omission* and errors of *commission*. It is this latter category, errors of commission, that present a problem for causal theory accounts of memory. This is because, according to the causal theory, memory is mainly a preservative capacity. These errors—the DRM and Loftus paradigms—represented cases of generative memory, and so present a challenge for traditional causal theory views.

The challenge of reconciling these kinds of generative memory errors with the causal theory is what prompts the two main sides in this debate. The traditionalists argue for a variation on the causal theory that allows for some constructive features of memory (Bernecker 2008; Michaelian 2011; Robins 2016b)—these might more accurately be referred to as *hybrid* views since they incorporate constructive features of memory along with the causal condition. Revisionists, on the other hand, argue that we can dispense with the causal condition entirely (Michaelian 2016b.); or, that episodic memory is not, in fact, a distinct cognitive system at all (De Brigard 2013). These accounts take the consistency and prevalence of generative memory errors as grounds for rethinking the idea that episodic memory is a mainly preservative cognitive capacity.

Both side of this debate face challenges. The main challenge for the hybrid views is how to go about reconciling the preservative and generative features of memory. The difficulty, here, is the extent to which a causal account of memory can allow for generative content whilst still retaining the importance of a memory trace and a primarily preservative function. For the revisionists, the challenge is how to make sense of episodic memory without positing a trace. As we've seen, the notion of a memory trace, even if simply theoretical, is important for making sense of both the successes and failures of episodic memory. To dispense with the notion of a memory trace as a solution for accounting for generative memory errors seems unwarranted. I think, then, that hybrid views present the more promising solution. In the following chapter, I will argue for this: that revisionist positions face an insurmountable difficulty in attempting to dispense with the notion of a memory trace, and that an account of the function of episodic memory is consistent with that of a broadly preservative capacity involving a memory trace.

Let's look now at how we might classify these cases of false memory.

Chapter 3: Errors and Revisions

In the previous chapter, we looked at different kinds of memory errors. We saw that certain memory errors—errors of commission, that involve the creation of new content—present a challenge for traditional causal accounts. This is because a strict interpretation of the causal theory holds memory to be a purely preservative capacity. In response to these types of errors, proponents of the causal theory have embraced what are sometimes known as hybrid theories of memory—theories that maintain the causal component of memory but allow for some constructive processes that can generate additional content in memory. Alternatively, we saw that proponents of revisionist accounts of episodic memory took the prevalence and consistency of these types of memory errors as grounds for rethinking the function of episodic memory, altogether.

As I've mentioned several times in the previous two chapters, I think there are good reasons for defending the causal condition, and so of the two main approaches for addressing errors of commission in episodic memory, I think the hybrid views offer the more promising solutions. In order to make this determination, however, it's important to be clear on exactly what the scope of revision proposed by these views entails, and how these proposed revisions address the issue of memory errors. It's also important to consider whether the evidence of these memory errors is sufficient to warrant the scale of functional revision that these views suggest.

In this chapter, I will assess two of the more prominent revisionary accounts found in the literature—De Brigard (2013) and Michaelian (2016b). I will then compare how these views propose resolving the difficulty presented by memory errors with an alternative, hybrid, proposal (Robins 2016b). Ultimately, I will be arguing that the revisionary suggestions provided by De Brigard and Michaelian are unwarranted (as in the prevalence of memory errors does not justify the full-scale revision of episodic memory), and unacceptable (as in these views do not provide a satisfactory solution).

I'll begin the chapter (section 3.1) by looking in detail at De Brigard's episodic hypothetical thinking (2013), and Michaelian's episodic imagination[7] (2016). In many ways, these two views are similar: both argue that the prevalence and consistency of memory errors of commission warrants the revision of episodic memory. Similarly, both views propose memory should be considered part of a larger cognitive system. Despite this, their views diverge on two key issues: the function of episodic memory, and the necessity of a memory trace. Since, as we saw in the previous chapter, the need to explain errors of commission is at the heart of these proposals for re-thinking episodic memory, in section two (3.2) I'll compare how these revisionist views treat memory errors with an account of memory errors from a more traditional, hybrid, view on memory (Robins 2016b). In section three (3.3) I'll evaluate the implications of functionally revising episodic memory, both in terms of the rejection of a memory trace, and in making sense of memory errors. Here, I will argue that neither of these revisions should be accepted. In section four (4.4), I'll provide a summary of the chapter.

3.1 Revisionary views

De Brigard (2013) and Michaelian (2016b) are two of the main proponents of revising episodic memory. Although their views are, broadly, quite similar, each of their views contains nuances that makes lumping them together for the purpose of analysis problematic. In order to provide a better overview of what each thinker is suggesting, I'll deal with their views independently, flagging the areas in which I think the two approaches diverge from each other in important ways.

I'll begin this section by outlining De Brigard's reasons for proposing a functional revision of episodic memory (section 3.1.1.). Then, I will look in detail at how, exactly, he thinks this revision should be applied (section 3.1.1.1). I will also evaluate his stance on a memory trace, and what this means for his view (section 3.1.1.2). I will then go on to present a brief overview

[7] Michaelian (2016a) refers to his view as a "simulationist view". Michaelian (2016b) refer to simulationist views as a classification of views that see episodic memory as part of a larger cognitive system, these include mental time travel (Suddendorf and Corballis, 2007), episodic hypothetical thinking (De Brigard 2014), and episodic imagination (Michaelian 2016a). These are also sometimes referred to as "continuist" views since they conceive of episodic memory as being continuous with other cognitive systems or capacities (see Robins 2020).

of Michaelian's account of *episodic imagination*, paying particular attention to the differences between this view and De Brigard's (section 3.1.2). I will also evaluate his view's commitment to a memory trace (section 3.1.2.1), which, I think, is an important point on which the two accounts diverge.

3.1.1 De Brigard's view

In his 2013 article *Is memory for remembering? Recollection as a form of episodic hypothetical thinking*, De Brigard lays out a case for revising the cognitive function of episodic memory. As a starting point, he takes the prevalence and systematicity of certain types of memory errors—errors of commission—and argues that these are not necessarily errors, as understood by traditional theories of memory, but rather features of a larger cognitive network he calls *episodic hypothetical thinking*.

To make this case, De Brigard considers the prevalence of cases of "misremembering". These include common memory distortions: telescoping effect, boundary distortion, and field to observer effect (De Brigard 2013, 158-160). He also considers the results from the DRM paradigm (Roediger and McDermott 1995), and Loftus-style false memory studies (Loftus 1975). Based on these studies, he argues that cases of misremembering are '*pervasive*' (i.e., they occur frequently). He also points to studies that suggest subjects with healthy memory systems are more prone to misremembering than those with memory impairments (Schacter et al. 1996; Melo et al. 1999; Ciaramelli et al. 2006), from which he concludes that these memory errors are also '*normal*' (i.e., they occur in ordinary healthy individuals). A final observation he makes of these studies is that only certain kinds of information are susceptible to being misremembered, in particular, information that has a '*plausibility*' about it (i.e., is possible to have happened) (Pezdek et all. 1997).

If it is the case that memory errors are both *normal* and *pervasive*, De Brigard argues, then it places proponents of traditional memory theories in an unfavourable position: they have to accept that a healthy episodic memory system frequently and predictably malfunctions. On top of this, he notes, that while the prevalence of malfunctions, in itself, isn't necessarily grounds for

rejecting a given functional interpretations, after all many systems malfunction frequently and predictably without giving reason to re-think their overall function, episodic memory is unlike these other systems, and therefore should not be exempt from functional revision. In support of this claim, he considers biological threat detection systems (meerkat alarms) and human heuristics (availability heuristic) as examples of biological systems that malfunction frequently but nonetheless maintain the same functional assessment (De Brigard 2013, p. 163-4). The episodic memory system, however, De Brigard argues, is like neither of these cases.

Episodic memory is unlike these other biological systems, according to De Brigard, because in the case of meerkat alarm calls and the availability heuristic, there have been environmental changes that increase the rate of false-positives, whereas the memory system hasn't undergone any similar environmental change. To add to this, De Brigard points out that when a system malfunctions, it is typically less beneficial for the organism than when it functions correctly; however, regarding episodic memory, it isn't clear that this is the case. According to De Brigard, memory errors occur frequently and often go overlooked (De Brigard 2013, p. 164-5). What's more, he argues, memory errors might even be beneficial: here, he cites studies that indirectly show a correlation between false recognition and success on problem solving tasks (Howe et all 2010; Dewhurst et al 2011). If, as he suggests, these studies support the claim that memory distortions are adaptively advantageous, then this puts more strain on traditional views that interpret these distortions as malfunctions.

Overall, then, his position seems to be that because memory errors are *normal* and *pervasive*, and are unlike other *normal* and *pervasive* kinds of biological system malfunctions (meerkat alarm malfunctions and human availability malfunctions), this presents a genuine problem for any account of memory that considers memory errors to be *normal* and *pervasive* malfunctions.

3.1.1.1 De Brigard's functional revisions

Considering the difficulty that these types of memory errors pose for the causal theory, De Brigard proposes a functional re-assessment of episodic memory. His reasoning is as follows: just because memory systems do something (remember) this doesn't mean this is what they are

for (function). Rather than simply basing our functional assessment of episodic memory on how it seems to function, he suggests taking a *role function approach*, in which the function of a system is assessed according to its contribution to the organism (Cummins 1975). For this analysis, we need to look at 1) how the mechanisms in the system work, and 2) how the system contributes to the functioning of the system in which it is contained.

De Brigard begins his functional analysis by considering a case from Loftus' false memory paradigms (Loftus 1975). In the case he proposes, a red car fails to stop at a 'yield' sign. When asked by a police officer 30 minutes later if the car failed to stop at the 'stop' sign, the subject misremembers the event as having involved a *stop* sign instead of a yield sign. De Brigard's explanation of these results is that the memory in question involves the reconstruction of dispersed, encoded information, that is assembled in memory as a *gist-like representation* (Schacter and Addis 2007). Importantly, this information is largely incomplete, and so the memory system is responsible for reconstructing an episodic memory from limited information (the colour red, a sign, a car, the sound of the crash, etc.). To do so, according to De Brigard, the memory system must 'fill in the gaps' of missing information. This process is constrained, not only by the encoding process, but by previously experienced, categorically similar, schema-consistent information. In support of this, he points to Bayesian computational models (Anderson and Schooler 1991) that determine the likelihood of a certain item being remembered in a given context to be the product of the likelihood of that item belonging to that context combined with the probability of that item previously featuring in that context. In other words, it's not only the experienced episode, but also categorical similar past experienced episodes and the probability of an item being present in a given scenario that determine the likelihood of recollecting that specific item in that specific scenario.

Returning to the car accident case, this would explain why, when cued by the police officer's verbal utterance "stop sign" (increasing the likelihood that a stop sign was, indeed, present at the scene), the subject is more likely to remember a stop sign (which they've presumably seen more of in the past) than a yield sign (which are less common). This analysis provides a rough sketch of how the relevant mechanisms in episodic memory work to produce a representation of the event, which responds to the first part of the role functional analysis. Accordingly, De Brigard

argues, even though the subject is misremembering a stop sign instead of accurately remembering a yield sign, the mechanisms within the memory system are still functioning the same way as they would in a case of accurately remembering.

Based on his analysis, De Brigard draws three important consequences. *Firstly*, certain cases of misremembering and cases of successful remembering apply the same cognitive mechanisms. Just as the above scenario describes a case where the mechanisms responsible for misremembering are the same as in cases of successful remembering, so too are there cases of successful remembering where the mechanisms involved are working just as they would in cases of misremembering. Accordingly, the memory system is probabilistic, so what's important in a successful case of remembering is that the memory system recalls an optimal representation of an event given a cue (De Brigard 2013, p. 170). A *second* consequence of his analysis is that the memory system deploys mechanisms that are used for purposes other than simply recollecting. The same neural regions, for example, are activated during the perceptual processing as well as during the recollection of a given event (Wheeler et al. 2000). And, *thirdly*, De Brigard concludes that our memory systems use probabilistic strategies to retrieve information.

It is this third consequence that is most relevant to De Brigard's proposal. Tying this back to one of the features he first noted regarding cases of misremembering, that they have a certain *plausibility* about them (i.e. we tend to misremembering things that might have happened), he argues that our memory system is sensitive not only to what has happened, but what could have happened. In explanation of this, he suggests that the same mechanisms underlying episodic memory also underly episodic imagination (De Brigard 2013, p.173-4), and episodic counterfactual thinking (De Brigard 2013, p. 175)—all of which, purportedly, belong to a larger cognitive system of *episodic hypothetical thinking* (De Brigard 2013, p.176-7). This, then, answers the second part of the role function analysis: how the system contributes to the function of the organism in which it is contained. Accordingly, episodic memory does not function to recollect past experiences, but rather the recollection of past experiences available in episodic memory are simply products of a larger cognitive system that generates first person mental

simulations of events that may happen, may have happened, or could happen[8] (De Brigard 2013, p.177).

Thus, De Brigard concludes that a theory of episodic hypothetical thinking provides a better explanation for cases of misremembering. This is a view that effectively proposes revising the traditional, preservative, function of episodic memory from a system that stores information to a system that predicts probable scenarios.

3.1.1.2 De Brigard on a memory trace

From De Brigard's proposed revision, we can see two interconnected, but distinct, issues emerge. The first has to do with the cognitive function of episodic memory: what functional purpose does episodic memory perform? The second issue has to do with the status of a causal connection to a past event, or memory trace.

Regarding the first issue, De Brigard's view clearly proposes a revision of the traditional functional model of episodic memory that takes memory to be a preservative capacity. According to De Brigard, it is a mistake to take remembering (i.e. recollecting past events) as the primary function of the episodic memory system. Instead, remembering should be understood as an operation within a larger cognitive system—episodic hypothetical thinking—that is responsible for simulating probabilistic first-person scenarios.

De Brigard's stance on the second issue, whether or not this process involves a memory trace, is less clear. In his proposed analysis of Loftus' car-accident case of false memory, the explanation

[8] To bolster his theory, De Brigard references various supporting studies. Notably, deficits in autobiographical memory are associated with deficits in the ability to project oneself into the future (Tulving 1983, Klein et all 2002, Hassabis et al 2007). Studies that show parallels between the development of episodic memory and future thinking in children (Atance and O'neill 2001), and adults (Addis et al 2008). And studies of neuroimaging showing the *core brain network* (composed of the hippocampus, the posterior cingulate/retrosplenial cortex, the inferior parietal lobe the medial prefrontal cortex, and the lateral temporal cortex) is involved during autobiographical remembering and future projecting (Schacter et al. 2007)—which is consistent with Addis and Schachter's *constructive episodic simulation hypothesis* (2007). All of these, he argues, support the view that episodic memory is part of this larger cognitive system.

relies heavily on the optimal reconstruction of the event from incomplete fragments of memory. If this is the case, then his account is entirely consistent with the notion of a causal connection via memory trace, for what else are these "fragments of memory" if not memory traces? Most contemporary advocates of the causal theory agree that the process of remembering involves at least some level of reconstruction (Robins 2016b), which may very well resemble the process De Brigard has described.

This position is complicated, however, when De Brigard describes one of the consequences of this analysis—that the same mechanisms involved in cases of misremembering are involved in cases of veridical remembering. His claim is that in both cases, the memory system is functioning properly: "[T]herefore, a successful recall produced by reconstructing the optimal representation of an event given a cue would count as veridical recall even if the sensory details of the original stimulus were not attended or encoded" (De Brigard 2013, p. 172). Here, he is suggesting that it's possible to have a case of veridical memory without any of the original stimulus, and so no causal connection. Effectively, this would be a case of optimally reconstructing the representation of a scenario such that the reconstruction provides an accurate representation without the corresponding experience, similar to imagining an experience as it happened without actually experiencing it. The way to understand this, I assume, is that although memory typically requires a memory trace (as in his analysis of the car-accident example), it doesn't necessarily have to—which would conflict with the causal theorist's account.

This reading is challenged, however, by De Brigard's position in a more recent article (I believe he has revised his position). De Brigard (2020) has argued for the explanatory indispensability of a memory trace. In this article, he re-interprets Martin and Deutscher's (1966) argument for a causal theory account of memory contra the contemporaneous anti-realist view (Malcom 1977). Effectively, De Brigard's argument, here, is that we need to posit a memory trace as an inference to the best explanation, in support of which he provides a novel variation on Martin and Deutscher's original argument, which was focused primarily on analyzing the ordinary concept of *remembering* rather than an empirical hypothesis about the capacity of memory. Considering this, I think it's fair to assume that although De Brigard supports revising the function of episodic memory, he isn't entirely committed to dispensing with the notion of a memory trace.

In the next section, we'll look at Michaelian's episodic imagination—a view that is similar to De Brigard's episodic hypothetical thinking, but more committed to redefining memory without the causal condition.

3.1.2 Michaelian's view

The simulationist revision that Michaelian proposes—episodic imagination—is similar to De Brigard's episodic hypothetical thinking, though, according to Michaelian, it is more focused on developing an account of the process of episodic thinking, whereas De Brigard's episodic hypothetical thinking is focused more on providing an account of its function. Michaelian considers there to be two key differences between their views: firstly, although both views consider episodic memory to be part of a larger cognitive system that encompasses different forms of episodic imagination, De Brigard places a greater emphasis on counter-factual thinking, and so is skeptical about the accuracy of episodic memory, whereas Michaelian takes it to be more *reliable*; and secondly, De Brigard sees the function of episodic memory as providing the 'raw material' for episodic hypothetical thinking, whereas Michaelian takes episodic future thinking to be the primary function (Michaelian 2016b, p. 115-6). I would add a third difference to the list, and that is the commitment to a memory trace—Michaelian's episodic imagination view explicitly makes it clear that episodic memories, even veridical ones, don't require a causal condition.

One worry that accompanies these revisionary views is that once we abandon the notion that memory is a preservative capacity and that it stores information via memory trace, we should also abandon the notion that the representations in episodic memory are reliable. Michaelian points out that constructive memory research typically focuses on cases where the process of memory construction goes wrong, where retrieved representations don't correspond with experienced episodes, such as in cases of confabulation (2016b, p. 90-91). The underlying assumption in this worry is that if memory is understood as a preservative capacity, then it is reliable insofar as it accurately performs this function; if, however, we dispense with the notion that memory performs such a function, that of preserving the representational content of past

experiences, then this may be grounds for doubting that memory is in fact a reliable process. Michaelian, however, argues that this worry is unfounded. Considering the incorporation of testimony (Michaelian 2016b, p.127-48), and the meta-cognitive processes of source monitoring (processes that monitor the reliability of the origin of an episodic representation) (Michaelian 2016b, p.149-168), and process monitoring (processes that determine the process involved in an episodic memory—i.e. a simulation of a past episode, or a future episode) (Michaelian 2016b, p.169-200) to offset the uncertainty introduced by the constructive features of memory, he argues that memory should still be treated as a *reliable* process. Just as visual illusions don't show that vision is an unreliable process, so too do misrememberings not show that episodic memory is an unreliable process (Michaelian 2016b, p. 148).

The reliability of episodic thinking, Michaelian argues, is connected with the quality of consciousness with which we experience episodic simulations. His story for the evolutionary adaptivity of episodic memory follows: the quality of consciousness of episodic imagination distinguishes episodic memory in humans from episodic-like memory in animals (Michaelian 2016b, p. 207; Rowlands 2009). The consciousness of episodic memory can be understood in terms of autonoesis (first-person subjective quality) and chronesthesia (consciousness of the time in which mental time travel occurs) (Tulving 2002, p. 315). These features are too complex to be an evolutionary spandrel (a by-product that wasn't selected for). Rather, Michaelian explains, these features contribute to source monitoring (autonoesis discriminates between self and other forms of mental times travel) and process monitoring (chronesthesia discriminates between past and future forms of episodic simulation), which, in turn contribute to the reliability of episodic memory (Michaelian 2016b, 229-233). According to Michaelian, an account of the function of episodic memory needs to take these features into consideration, and only a future-oriented functional account does so[9].

[9] Michaelian organizes the explanations for episodic memory into three broad groups: past oriented explanations, future oriented explanations, and social oriented explanations. Past oriented explanations maintain that episodic memory represents the past in-order to contribute to future judgments (Michaelian 2016b, p. 219). Social oriented accounts explain the adaptivity of memory as based on social functions: character judgments (Klein et al 2009), epistemic vigilance (Boyer 2009), and social relationships (Skowronski and Sedikides 2007). And, future oriented explanations look to explain episodic memory as contributing to the simulation of future scenarios (Suddendorf and Corballis 2007). Of these three groupings, Michaelian argues, it is only the future oriented accounts that consider both the relationship between episodic memory and other episodic cognitive capacities, and the subjective quality of

3.1.2.1 Michaelian on a memory trace

A third distinction between Michaelian and De Brigard's view that I think is worth addressing is Michaelian's commitment to dispensing with the causal condition of memory, or a memory trace.

Michaelian (2016a) argues that the causal condition of memory needs to be supplemented with a reliability condition in order to better account for the generative features of episodic memory. Michaelian (2016b), then argues that with a reliability condition in place, the causal condition is no longer necessary. On his simulationist account, Michaelian is clear that episodic memory is a sub-species of episodic imagination more generally, and that any episodic simulation draws on different sources of information to construct simulated episodes. "While simulation of a given past episode presumably *often* draws on information originating in the agent's experience of that particular episode, it will rarely draw exclusively on such information, and in principle it need not draw on such information *at all*." (Michaelian 2016b, p.103). Effectively, with the reliability condition in place, Michaelian maintains there is no need to posit a memory trace to explain the accuracy of episodic memory. What's more, Michaelian (2016a) argues, the reliability condition allows for a more inclusive taxonomy of memory errors that effectively succeeds where he sees causal theory and hybrid accounts of memory falling short—in accommodating different species of memory error. Thus, dispensing with the causal condition is a central feature of Michaelian's account, and important for explaining how to make sense of the different species of memory errors.

Since memory errors, and in particular, as we saw in chapter two, errors of commission, are the central motivator for proposing revisions of episodic memory, in the next section we'll look at how these two views approach this issue. We'll also consider Robins' (2016b) hybrid account.

episodic memory (2016b, p. 220). Thus, he considers the most promising explanation for the adaptive selection of episodic memory to be future-oriented.

3.2 Memory errors and revisions

Now that we have an idea of what the two main revisionary views are proposing, we can assess how these views accommodate the different kinds of memory errors.

If you recall from chapter two, we saw that there are two broad classifications for categorizing memory errors: *errors of omission*, that involve leaving out content that was present in the initial experience; and *errors of commission*, that involve including content that was never present in the initial experience (Schacter 2001). It is the latter category, errors of commission, that presents a problem for causal theory accounts, since traditional, causal, models of memory do not recognize any generative (content creating) features in episodic memory.

Revisionists, like De Brigard and Michaelian, think the best way of dealing with these memory errors is a functional revision of episodic memory, such that, rather than label them as memory "errors" according to a causal theory model of memory, they should be understood as deliberate features of a more extensive cognitive system (either episodic hypothetical thinking, or episodic imagination, respectively). More moderate approaches to dealing with errors of commission come from hybrid views, like Robins (2016b). According to these views, we can make room for errors of commission while still maintaining the general framework for memory set out by the causal theory.

In this section, we'll evaluate how these different accounts deal with errors of commission. I'll begin (section 3.2.1) with an overview of De Brigard's approach to memory errors, and raise some concerns with his reasoning. Then (Section 3.2.2) I'll examine Robins' suggestion for classifying the different kinds of memory errors—an approach that she argues takes some of the revisionist pressure off causal theory models. Michaelian specifically rejects Robins' classification, and provides a revisionist alternative, which I'll examine (section 3.2.3). Then (section 3.2.4), I'll challenge Michaelian's alternative view, ultimately suggesting Robins provides the better taxonomy for memory errors of commission.

3.2.1 De Brigard on memory errors

From De Brigard's argument for episodic hypothetical thinking, we saw that two important features he noted regarding errors of commission were that they were *frequent* and *ordinary*. These features are a motivating factor for why he sees the need for a functional revision: if certain kinds of memory errors occur frequently in an otherwise ordinary, functional memory systems, then one explanation might be that these are not, in fact, malfunctions, but rather symptoms of a misdiagnosed function. This is the explanation De Brigard (2013) favours.

For his part, then, De Brigard (2013) is content to group these different cases of memory error together. These include: telescoping effect, boundary distortion, and field to observer effect, as well as the results from the DRM paradigm (Roediger and McDermott 1995), and Loftus-style false memory studies (Loftus 1975). By doing so, he effectively places a greater weight on the preponderance of false memories, which, in turn, provides a stronger motivation for a revision of episodic memory.

As we saw in our analysis of memory errors (chapter two), however, the observed frequency of memory errors in the different studies showed a significant variance. In the DRM paradigm studies, for example, memory errors registered as high as 70% for lure words, whereas in the Loftus false memory paradigm, false memories had a much lower success rate of up to 30%. This discrepancy makes sense, since there is a significant difference in misremembering a thematically related word on a list (as in the DRM paradigm) and coming to possess the false memory of an entirely fictitious event, such as a hot-air balloon ride (as in the Loftus false-memory paradigm). This discrepancy raises the question, then, of whether it's fair to say that *all* of these memory errors occur *frequently*. A further concern should be raised regarding whether the contrived experimental settings that succeed in generating false memories qualify as *ordinary*—perhaps, the subjects are ordinary, but the experiment conditions are certainly irregular. Remembering a long list of thematically relevant words (DRM paradigm), for example, doesn't seem like an ecologically valid task. In the same way, providing subjects with material specifically designed to cue misremembering (i.e false journal entries, doctored pictures, and duplicitous testimony), as in the Loftus paradigm, isn't consistent with the settings in which

memory typically functions. It seems like a bit of a stretch to claim that the results from these experiments are reflective of the *ordinary* nature of memory errors.

In light of the differences between these cases of memory error, an alternative to grouping them together would be to differentiate them. Robins (2016b) proposes just that.

3.2.2 Robins on memory errors

Robins (2016b) argues that there is an importance difference between cases of misremembering (as seen in the DRM paradigm) and false memory (as seen in the Loftus experiments). This difference, she points out, is that the DRM paradigm combines features of preservative memory (in retaining information about the lists) and generative memory (in producing words that didn't appear on the lists). The best way to make sense of this, she argues, is by proposing a *hybrid* account of memory—one that combines preservative features (consistent with a causal connection) and generative features (consistent with some degree of memory construction).

As Robins notes, she takes the difference between misremembering and false-memory to be significant, although she admits that it might turn out to be a matter of degree more than of kind (2016b, p.434). We can see her reasoning if we consider what's happening in the DRM cases: the DRM cases involve falsely remembering a word from a list in which the list, as a whole, is more or less, accurately remembered. Depending on where we choose to set the parameters of what constitutes an episodic memory, these cases could either qualify as false memories, where the falsely remembered word constitutes the extent of the false memory, or cases of memory distortion, where the falsely remembered word only constitutes part of the episodic memory of the entire list. Even if we accept the first of these options, however, that the extent of the false memory is constituted by the falsely remembered word, there is still an important difference between the cases where a word strongly associated with list words (like *spider*) is falsely remembered, and those where a word weakly associated with list words (like *sandwich*) is falsely remembered—and that is that in the former case, in order to explain the high rate of false recollection, we need to refer to the strong association between the falsely remembered word and

the words that actually appeared on the list, and doing so requires referring to a causal connection to the experience of seeing the words on the list.

This important difference is what leads Robins (2016b) to suggest these different memory errors be classified separately. This re-classification, she argues, presents a difficulty for revisionist positions since they are unable to adequately make this distinction because, according to these views, all memories are a species of confabulation. To make the important distinction between misremembering, as in the DRM paradigm cases, and confabulation, as in the false memory cases, she proposes a two-stage model of memory: *retention* of information from the original event and, construction of an *accurate* representation during the retrieval process (Robins 2016b). These two stages correspond with the two features of the folk notion of the function of episodic memory we outlined in chapter 1—namely, that episodic memory functions to preserve the *accuracy* and *authenticity* of past experiences. While Robins' view isn't as strictly preservative as folk assumptions about episodic memory might be, it maintains the causal theory insight.

TABLE 1 | The causalist taxonomy of memory errors. (from Michaelian 2016.b)

	Retention	Accuracy
Successful Remembering	Yes	Yes
Misremembering	Yes	No
Relearning	No	Yes
Falsidical confabulation	No	No

On her two-stage model (table 1, p. 40), we can account for the various species of memory error. *Mismemberings* (such as cases like the DRM paradigm) occur when the retention (or authenticity) condition is met, but the accuracy condition is not. False memories (such as in the Loftus experiments), or confabulations, occur when neither condition is met: the memory is neither accurate, nor is it authentic. Successful cases of episodic memory occur when both conditions are met. And, relearning occurs when the retention condition is not met, but the accuracy condition is.

Reclassifying these cases of misremembering as Robins proposes, challenges the revisionist position in two ways. Firstly, it lessens the preponderance of memory errors that De Brigard suggests is motivation for functionally revising episodic memory. If you recall, one of De Brigard's main inferences from the cases of misremembering was that they were *pervasive* (i.e. they occur frequently). While this might be the case if all of these types of memory errors are grouped together, once they are differentiated, it appears that while some occur frequently (misremembering as in the DRM paradigm), others are much less common (confabulation as in the Loftus paradigm). Importantly, according to Robins' classification, the causal condition is still required to make sense of the more frequent cases of misremembering. Which brings us to the second challenge: Robins' classification provides a framework for a hybrid account of memory—a model that still relies on the concept of a memory trace, but allows for some constructive processes. The implicit challenge, here, is that if such a hybrid view can account for memory errors, and there is still strong evidence to support the idea of a memory trace (reviewed in chapter one), then a complete functional revision of episodic memory becomes unappealing.

3.2.3 Michaelian on memory errors

As we saw, however, Michaelian has argued against the need for a causal condition, or memory trace, suggesting instead that a reliability condition is better suited for making sense of these memory errors.

Michaelian (2016a) disputes Robins' (2016b) two-stage assessment of memory. He argues that her account fails to acknowledge two more important classifications of memory error. These are cases of *veridical* confabulation and cases of *false* relearning.

According to Michaelian, confabulation is typically assumed to imply falsity, which he argues is mistaken (2016a, p. 4). Here, he draws an analogy between veridical hallucination (Lewis 1980) and veridical confabulation—if it's possible to hallucinate something that actually happened, then it's equally possible to confabulate something that actually happened. Although veridicality

is not the norm in either case, we must allow for it. In such cases, we would have *veridical* confabulation.

Just as confabulation implicitly assumes falsidity, which, according to Michaelian, is mistaken, relearning implicitly assumes veridicality, which, according to Michaelian, is equally mistaken (2016a, 5). If we consider the classic diary case of relearning, wherein an individual reads the content of their diary and comes to remember (relearn) specific information, the underlying assumption is that the information in the diary is correct, thus we have veridical relearning. If however, the information in the diary is incorrect, and the subject still comes to relearn it, then we have a case of *falsidical* relearning.

Regarding these two types of errors, Michaelian points out that Robins' two-stage model cannot distinguish between cases of veridical relearning and veridical confabulation, and between cases of falsidical confabulation and falsidical relearning. This is because, according to Robins' two-stage model, veridical relearning and veridical confabulation both consist of memory errors characterized by an absence of the retention condition but the presence of the accuracy condition. And, falsidical relearning and falsidical confabulation both consist of memory errors characterized by an absence of both the retention and the accuracy condition. Accordingly, there would need to be a further classification to adequately accommodate the different kinds of memory errors.

TABLE 2 | The simulationist taxonomy of memory errors. (from Michaelian 2016b)(altered)

	Internality	Reliability	Accuracy
Successful remembering	Yes	Yes	Yes
Misremembering	Yes	Yes	No
Veridical confabulation	Yes	No	Yes
Falsidical confabulation	Yes	No	No
Veridical relearning	No	Yes/No*	Yes
Falsidical relearning	No	Yes/No*	No

*Reliability condition doesn't matter for relearning

Michaelian (2016a) proposes just that: a three-stage model that breaks down into internality, reliability (as opposed to retention), and accuracy (table 2, p. 42). On his view, since a memory trace is not a necessary condition for episodic memory, he replaces Robins' retention condition with a reliability condition. The addition of the internality condition, that determines whether the content of the episodic representation comes from within the subject (as in cases of memory and confabulation) or involves an external prompt (as in cases of relearning), he argues, allows his view to distinguish between cases of veridical relearning and veridical confabulation, and between cases of falsidical relearning and falsidical confabulation. This is because both cases of relearning require an external prompt, and so meet the externality condition, whereas other classifications of memory do not.

Using this model, Michaelian argues, we can: 1) better accommodate the different kinds of memory errors, and 2) dispense with the causal condition, or memory trace (since, instead, we have a reliability condition).

3.2.4 What of Michaelian's externality condition?

The cases that Michaelian suggests are left out of Robins' account are, intuitively, controversial. Typically, we don't think of confabulation as being veridical, nor do we think of relearning to be falsidical. This is because falsity is usually assumed in the definition of confabulation, as is veridicality assumed in the definition of relearning. Michaelian, however, argues that this shouldn't be the case (2016b).

Relearning occurs when an individual doesn't properly retain the memory of a specific event, but upon being provided with a description of that event, comes to possess the corresponding episodic representation. When the description that the individual is provided is veridical, then the individual is engaged in veridical relearning; when the description is falsidical, then they are engaged in falsidical relearning. In effect, the variation on the classic diary example of relearning that Michaelian (2016b) uses to support this distinction is the same process conducted in the Loftus experiments, where subjects are provided with the description of fictitious childhood event, and then come to possess a corresponding (false) episodic memory.

These experiments from the Loftus paradigm are typically classified as cases of confabulation. If we introduce Michaelian's externality condition, however, it seems that they would more appropriately be classified as cases of falsidical relearning. According to Michaelian's taxonomy, true cases of confabulation only occur without an external prompt. This distinction, however, seems somewhat difficult to determine: most memories (and memory errors for that matter) involve some form of external prompting, whether it be contemporaneous or historical. Identifying cases of confabulation that involve no immediate or past external prompt seems highly difficult and, I would argue, unhelpful.

Similarly, cases of veridical confabulation entail, essentially, the same process as falsidical confabulation, except for that they happen to be accurate. Michaelian endorses this suggestion because, presumably, it works well with his idea of mental time travel. I can't, however, help but think of an analogy to the broken clock that is right twice a day. Do we maintain that the clock is broken *differently* when it happens to be correct at midnight and noon, or is it effectively broken in the same way despite the time? Basically, Michaelian is arguing that when confabulation happens to contain some accurate information, it is a different kind of error than when it doesn't. I disagree: confabulation is a memory error, and if it happens to contain some accurate information, then this is purely a matter of chance that does not warrant a distinct classification of memory error.

As Michaelian, himself, points out, the notion of veridical confabulation is a contentious idea to begin with. In support of it, he draws an analogy between veridical hallucinations in perception (Lewis 1980) and veridical confabulation in episodic memory, reasoning that if we accept the possibility of the former, then we should have no problem with the latter. Interestingly, veridical hallucinations in perception aren't exactly a given, either. Burge has argued that there is no such thing as a veridical hallucination in perception (2010. p. 381-83). He reasons that perceptions necessarily entail a demonstrative element that refers to a context-bound particular; and, since veridical hallucinations don't have a referent, they can't qualify as veridical perceptual representations (Burge 2010, p. 381). Although this reasoning doesn't exactly translate into an argument against Michaelian's notion of veridical confabulation—since Michaelian is explicitly

denying a necessary causal connection in the case of episodic memory—it does make the contentiousness at issue clear. Given the context of classifying memory errors, I think, then, that the category of veridical confabulation is, also, unhelpful.

Since both of Michaelian's added memory error classifications—veridical confabulation and falsidical relearning—are problematic, I would assert that Michaelian's externality condition doesn't really contribute to the classification of memory errors, and that it simply expands the domain of episodic memory to include features of episodic hypothetical thinking—namely, hypothetical predictions. If this is the case, then there are no good reasons for favouring Michaelian's taxonomy over Robins, and a good reason—that it introduces unhelpful classifications of memory errors—for rejecting it.

Having examined how these revisionary views propose treating memory errors, we can turn back to the question of a memory trace and the function of episodic memory.

3.3 Memory trace and function

According to a traditional causal theory interpretation, the function of episodic memory is necessarily connected with the notion of a memory trace. On this view, episodic memory functions such that it makes accurate representations of past, first-person experiences consciously available. How it does this is via a memory trace: a stored, encoded representation of past experiences. Once we move away from this traditional framework, as we've seen with De Brigard's and Michaelian's simulations views, this relationship becomes less straightforward: De Brigard wants to revise the function of episodic memory, but still maintain something consistent with the idea of a memory trace; Michaelian, on the other hand, wants to both revise the function of episodic memory, and dispense with the idea of a memory trace. I think, then, it makes sense to look at each of these ideas independently.

First (section 3.3.1), I'll revisit the reasons for why I think it is important that the function of episodic memory includes reference to a memory trace. I will then address Michaelian's view that we should dispense with a memory trace, arguing that this position isn't supported by the

available evidence. Next, (section 3.3.2) I'll evaluate the revisionist agenda, breaking it down into two issues: whether the evidence of memory errors supports the functional revision of episodic memory (section 3.3.2.1), and whether the proposed functional revisions are better suited to accommodate the different kinds of memory errors (section 3.3.2.2). Ultimately, I will argue that the revisionary suggestions provided by De Brigard and Michaelian are unwarranted (as in the prevalence of memory errors does not justify the full-scale revision of episodic memory), and unacceptable (as in these views do not provide a satisfactory alternative).

3.3.1 What about a memory trace?

I've suggested in several parts throughout this paper (sections 1.4.2, 2.3.2) reasons for why it might be problematic to dispense with a memory trace. There is, I think, good reason for preserving a memory trace.

The most obvious, and straightforward reason, to me, is in-order to successfully *represent the past*—i.e. in order to explain the ability of episodic memory to successfully represent episodes from the past. If you recall from chapter 1, this was one of the four reasons Robins (2017, p. 79-80) surveyed. It simply doesn't seem possible that one could consistently (as in not merely a matter of chance) and accurately (as in with a relatively high degree of fidelity) represent past experiences without a causal connection. If episodic memories weren't established in causal connections to past experienced episodes, then they would be no more reliable than clairvoyant premonitions about the future—and we don't typically take these to be all that reliable.

Michaelian, however, does think episodic memories are still reliable despite his rejection of the causal condition, hence his introduction of a "reliability" condition. While we might not consider clairvoyant premonitions about the future to be all that reliable, there are certainly other kinds of predictions we can make that are relatively consistent and accurate. Provided we have enough evidence, then there's no reason we can't consistently and accurately make inferences about the future—this is done all the time when we predict things like the weather, travel times, economic forecasts, etc.

And it stands to reason that, since we presumably have a larger dataset about past experiences, our inferences about them would be all the more accurate than our predictions about the future. Consider, for example, crime scene analysis. Around the turn of the 20th century, fictional characters like Sherlock Holmes and Hercule Poirot were popularized for their uncanny ability to solve crime by collecting evidence and making precise inferences about what occurred. While these fictious characters possess almost supernatural abilities, their popularity represents a growing awareness of the reliability of evidence-based inferences about the past. Modern crime scene analysis has come a long way since the fictious days of inspecting boot-prints with a magnifying glass, and forensic sciences can render complex reconstructions of what occurred absent witness testimony, through things like fingerprints, blood-spatter analysis, character profiling, etc. Based on these kinds of evidence, entire events can be reconstructed, and these reconstructions are generally considered *reliable* representations of what actually happened. Of course, this reliability is entirely contingent on the amount, and quality, of the available evidence.

This, I think, is an important point when considering the issue in terms of episodic memory, since when we are talking about reconstructing an event in episodic memory, the "evidence" in question is constituted effectively by other memories, or memory traces. No matter how good Sherlock Holmes is at solving crime, he can't just sit back in his pipe-smoke filled room and conjure up solutions to crimes without first being debriefed on the crime by Dr. Watson, or visiting the crime scene himself—effectively, there needs to be this causal connection. In the same way, if we expect our episodic memories to be reliable, as Michaelian does, in the absence of a causal connection (or memory trace), as Michaelian suggests, then this reliability is entirely contingent on the reliability of the memories from which it is derived. Ultimately, this derivation must conclude in the information stored from experienced past events (or memory traces).

I don't think Michaelian would entirely disagree with the above point. On his view, episodic memories are simulations that typically draw on information from an agent's past experienced episodes (memory traces), just that they don't necessarily have to (Michaelian 2016, p.103). It is this second part of the claim that is mysterious: how else can we have reliable representations of past experiences without drawing on information from those experiences?

On Michaelian's view, the representation made available in episodic memory (the simulation) draws on a variety of information stored in memory (source information); often, that information is stored during the experience of the specific episode in question (memory traces), though, sometimes, the information used to generate the simulation would be stored during other, similar, experiences, just not that specific episode. Effectively, he's drawing a distinction between the stored information (the memory traces) and the simulations generated from them (the episodic memories) and arguing that there needn't be a perfect 1:1 mapping in order for the latter to constitute a reliable representation of the past.

The difficulty with his position is that the accuracy observed in episodic memory relates not only to the generalized content of the episodic representation, but also to episode-specific content. The reliability of episodic memory isn't simply a feature of the accuracy of general information about past experiences, but of specific details indexed to specific past experienced episodes. That is to say: although some of the observed content in episodic memory can be explained by general inferences, some of it can't. The best way to make sense of these episode-specific details is by positing connections to specific past experienced episodes (or memory traces). Most hybrid theories of memory accept that the representations available in episodic memory contain some combination of generic information (perhaps inferences about past experienced episodes) and specific details (memory traces).

While the precise nature of the relationship between the preservative (i.e. memory traces) and the constructive features of episodic memory is far from being fully understood, we don't need to understand the finer points of this relationship in order to appreciate the necessity of a memory trace. At its bare bones, a memory trace is simply a theoretical posit. This is why many proponents have suggested an inference to the best explanation as offering the strongest support for memory traces in episodic memory (De Brigard 2020, Khalidi 2023). The underlying point remains that the best way of accounting for the episode specific accuracy found in episodic memory is by positing a memory trace. If Michaelian wants to dispense with this explanation, then he'll have to provide a better one, something that, as I've argued, his reliability condition does not achieve.

3.3.2 On revising the function of episodic memory

Turning, now, to the function of episodic memory, our discussion so far has raised several important questions. Firstly, if we refrain from dispensing with the memory trace, do we still need to revise the function of episodic memory. And secondly, if we do revise the function of episodic memory, how do the surveyed revisionary views stack up.

3.3.2.1 De Brigard's call for revision

Although, as we've seen, De Brigard's view isn't committed to the idea of completely dispensing with the causal condition of memory, he is nonetheless a staunch proponent of revision. This is because he sees the traditional causal theory view as incapable of accounting for the frequency and normalcy of memory errors. There are a few claims in his account worth touching on.

Firstly, as I've suggested in the previous section, the frequency and normalcy of errors of commission is debatable. If we group all of these types of memory errors together, as De Brigard seems to suggest, then this does speak to their prevalence (2013). However, as Robins (2016b) has pointed out, there are good reasons to think they should be differentiated. If we follow Robins' suggested classification, then the frequency of memory errors becomes less striking: certain, minor, cases of misremembering (such as in the DRM paradigm) might occur frequently, but other, more extreme, cases of false memory (as in the Loftus paradigm cases) would occur much less often.

The same can be said of the normalcy of these memory errors. It is true, as De Brigard points out, that these cases of memory error (the DRM paradigm and Loftus' false memory studies) occur in ordinary, healthy individuals. The same, however, cannot be said of the circumstances in which they occur. Certain memory errors (such as boundary extension, field-to-observer perspective, and the DRM paradigm) might occur in ordinary situations, but the contrived scenarios required to produce the results found in Loftus' false memory paradigm are far from ordinary, routine, scenarios.

If this is the case, then memory errors are neither as *pervasive*, nor as *normal*, as De Brigard's assessment maintains. This significantly weakens De Brigard's argument for revising episodic memory, especially if we consider that other, hybrid, models are promising candidates for assessing memory errors (Robins 2016b).

Even if we grant De Brigard the first part of his argument, however, that memory errors are *pervasive* and *normal*, I'm still not sure that this necessarily warrants the extreme revision of episodic memory he suggests.

De Brigard's claim that the frequency and ordinary nature of errors of commission warrants a functional revision of the episodic memory system because it is unlike other biological systems (meerkat alarms and the availability heuristic), is debatable. Here the differences he points to—that the memory system hasn't undergone an environmental change, and memory errors aren't problematic (and might even be beneficial)—are not very persuasive.

To say that the environment in which the memory system functions hasn't undergone any significant change seems incorrect. If we assume, for argument's sake, that episodic memory is a preservative capacity (as traditional causal theories maintain), then we assume it is concerned with storing information. The human environment of information storage has developed rapidly, and changed drastically: from language, writing and pictures, to photographs, videos, recordings, etc.). To judge the exact impact of these changes on human cognitive function would be a complicated analysis, but to suggest that they have had no impact at all seems shortsighted. The simple ability to cognitively offload information that is later available for retrieval is a substantial development, and has undoubtedly changed the environment in which episodic memory functions.

One way to imagine this change would be that having evolved the ability to store information through writing, humans abandoned cultivating certain aspects of memory; in turn, having more accurate written (stored) records would make shortcomings in memory more obvious. Evidence suggests that humans in pre-literate societies had extensive memories—without the ability to store information in texts, they would have had few options other than storing it in memory. To

facilitate this, they developed techniques for cultivating memory. The method of Loci technique (or the memory palace), for example, was first described in Cicero's *de Oratore* (55 B.C.) and details a way of cognitively storing information by allocating it to various places within the mental image of a familiar structure, like a house, or palace. There is evidence that similar techniques were employed much earlier, and date to pre-historic times (Kelly 2017). With the proliferation of writing, however, many of these techniques become redundant—why spend such an effort to remember things when it is much easier to just transcribe them? Nowadays, these techniques of remembering are rarely employed outside of memory contests, where they are exhibited more as a novelty than a useful skill. This shift away from using techniques to store information in memory in favour of storing it externally (via text, and other mediums), could certainly explain a change in memory. This would suggest that some of the failures of episodic memory might be due to lack of cultivating this faculty through training, as opposed to inherent failures of the capacity itself.

Another possible effect of this shift to storing information is that humans may have come to be more receptive of external prompts: in the past, humans might not have had the means for reconstructing their memories as frequently from new information as they do today (books, pictures, the internet). As these stores of information have developed, humans in turn may have come to increasingly depend on them for accurate information. If this information turns out to be incorrect, however, this could be responsible for increasing the frequency of false positives in memory.

Historically, new techniques of storing information entailed a certain level of legitimacy: manuscripts took hours of work to transcribe, printing documents were expensive, the technology involved in producing photographs and videos required specialized knowledge, etc. (Madden, A. D., Palimi, J. and Bryson, J. 2005). Cotemporally, however, these techniques have become much more accessible, and so simply storing information in some of these formats may no longer be as legitimizing as it once was. Consider, for example, an altered photograph, like those used in the Loftus-style experiments. At the time (in the late 70s), doctoring a photograph would have required skill and cutting-edge technology. Now, it's common practice for teenagers to alter pictures before they post them online. Considering such significant changes, it's easy to

see how the development of the technologies we use to supplement memory could, in turn, affect how our memory systems function.

Most people use prompts to remember things, they store information in certain ways, and are susceptible to scenarios where information they've stored, or sources they trust as prompts, are tampered with. Given the enormous shift in technology we use to store information, and effectively supplement our memory systems, it's hard to see how you could make the case that the environment in which episodic memory functions hasn't undergone any significant changes.

These are just two possible suggestions, but the point is the same: that the landscape in which memory functions has certainly changed sufficiently to account for memory errors being more frequent and/or apparent. In this case, memory systems might be *just like* other biological systems for which changes in the environment have altered success rates (i.e. human heuristics and meerkat alarm systems). If we accept that there have been changes in the environment in which episodic memory functions, which I think seems fairly obvious, then De Brigard's argument doesn't hold.

The further claim that memory errors aren't problematic and might even be beneficial is also debatable. Certain minor memory errors might not be problematic, but others might be extremely problematic. Undoubtedly, there are cases where misremembering can have grave implications, such as a witness sending an innocent man to jail; and, alternatively, there are cases where accurate memory is extremely advantageous. Considered from an evolutionary perspective, the issue is not whether accuracy is advantageous, but to what extent it's efficient. Obviously, humans don't have the cognitive capacity to accurately record every detail of every past experience and call these to mind on cue. A probable solution to this is a cognitive tradeoff between the accuracy and authenticity of storing information—a possibility I've already suggested in chapter 1.

Taken together, the above points clearly refute De Brigard's argument for revising episodic memory. The evidence does not support that memory errors are as *pervasive* or as *normal* as De Brigard suggests. And, even if they were, the pervasiveness and normalcy of a biological

system's malfunctions is not necessarily grounds for reassessing its function, episodic memory is not an exception to this.

3.3.2.2 Revisionary functional changes

As to what the functional revisions proposed by De Brigard and Michaelian look like, while intriguing, I'm not sure they are convincing.

If you recall, De Brigard proposed a two-part functional analysis of episodic memory. The first part of this was an analysis of the mechanisms involved in episodic memory, the second part of this was an assessment of the functional contribution to the overall organism. Considering the first part of this analysis, De Brigard's account, provided we include the necessary link to the causal condition, is not implausible and is entirely consistent with many hybrid views on the constructive features of episodic memory (Robins 2016b). The second part of this analysis, however, that episodic memory is part of the larger system of episodic hypothetical thinking, seems to be a bit of a stretch.

Both De Brigard and Michaelian share this view that episodic memory is part of a larger cognitive system, either episodic hypothetical thinking or episodic imagination, respectively. Support for this view, however, is unconvincing. The core observation that episodic memory and episodic imagination share the same neural underpinnings (Schachter & Addis 2007), for example, can be explained in many ways, and is consistent with much less revisionary accounts than those covered here. Presumably, for example, when we imagine things, we draw on information from past experienced episodes. In the same way, when we remember things, we draw on that same source of information. It follows that both cognitive faculties—our episodic memory and our imagination—utilize some of the same cognitive processes, which are constituted by the same neural networks. This, however, does not entail the claim that these two cognitive systems are, in fact, part of a single cognitive system. Many cognitive systems share similar neural networks since, in effect, there's limited neural space to work with for a multitude of cognitive functions. The amygdala, for example, is involved in processing negative emotions, such as fear; it is also, however, involved in the perception of odor intensity, sexually arousing

stimuli, trust from faces, biological motion, and sharp contours (Nathan & Del Pinal 2016). These processes certainly entail diverse cognitive systems. The argument that sharing neural structures implies shared cognitive function is essentially a reduction of cognitive function to neural function, and shouldn't be taken as a conclusive point that these cognitive systems are, in fact, part of a single system.

In the next chapter, we'll look more specifically at the function of episodic memory.

3.4 Summary

We began this chapter with an overview of two of the more prominent revisionary views in the episodic memory literature—De Brigard's episodic hypothetical thinking (2013), and Michaelian episodic imagination (2016b). We saw that, while both accounts propose, broadly, similar functional revisions of episodic memory, they differ on several important points.

In section one (3.1), I outlined De Brigard's (section 3.1.1) and Michaelian's (section 3.1.2.) reasoning independently. I also pointed out particular areas on which the two accounts diverge: on whether episodic memory is *reliable*, and on the necessity of a *memory trace*.

Since memory errors are the central motivation behind these views, in section two (3.2) I evaluated the different stances on memory errors. De Brigard treats the different kind of memory errors collectively. This strengthens his argument for revising episodic memory. However, there are good reasons for thinking these memory errors should be differentiated (section 3.2.1). Doing so significantly weakens the motivation for a functional revision of episodic memory. Robins (2016) does just this (section 3.2.2)—she suggests reclassifying memory errors according to two criteria (*retention* and *accuracy*). Reclassifying memory errors, as such, both weakens the revisionists motivation, and provides a more promising, hybrid, framework for addressing memory errors. Michaelian disputes this classification (section 3.2.3), suggesting instead we should replace Robins' *retention* criterion (the causal connection) with a *reliability* criterion. He also thinks we need to add a third criterion—the externality condition. According to Michaelian, this effectively allows us to account for veridical confabulation and falsidical relearning, two

kinds of memory errors that cannot be accounted for on Robins' model. These two added classifications of memory error, however, are controversial, and I argue that they should not be included (section 3.2.4). Without these, there are not reasons for favouring Michaelian's taxonomy over Robins'.

In section three (3.3), we return to the two central issues of these revisionary proposals: the necessity of a memory trace, and the scope of functional revision. Here, I argue that the evidence does not support dispensing with the notion of a memory trace (section 3.3.1)—even Michaelian's view (the only surveyed proponent of dispensing with a memory trace) is unable to provide coherent reasoning. The scope of functional revision can be broken down into two points: whether the evidence of memory supports a functional revision in the first place (section 3.3.2.1) and whether the proposed functional revisions adequately address the difficulty presented by memory errors of commission (section 3.3.2.2). In both cases, I argue that they don't. Ultimately, I argue that the revisionary suggestions provided by De Brigard and Michaelian are unwarranted (as in the prevalence of memory errors does not justify the full-scale revision of episodic memory), and unnecessary (as in these views do not provide an acceptable alternative).

Having established that these functional revisions are unhelpful, we can now go on to look more specifically at the function of episodic memory.

Chapter 4: The Function of Episodic Memory

A central question of this thesis has been the classification of episode memory. Specifically, whether we should classify episodic memory as having a broadly preservative capacity, in line with traditional accounts of memory, or whether this classification should be revised to accommodate episodic memory as part of a larger cognitive system. At the core of this debate, is a question about the function of episodic memory. Before we can properly classify a capacity, like episodic memory, it's important to have an idea of what that capacity actually does: what is its function?

But when we ask about the function of episodic memory, what exactly does that mean? Although there is a general sense of the term function, something like the role, or what it does, that most would likely agree to, the term can have a variety of meanings depending on the context in which it is being applied. Do we mean, for example, how something functions within a specific system, or, rather, what that thing was designed to do, or selected for; are we referring to functioning towards a specific end, or simply at a basic, mechanistic level? These sorts of questions complicate the broader notion of a function. The issue is further complicated when applied to a cognitive concept like episodic memory, which isn't exactly as easy to study as mechanistic systems, or even simpler biological systems. Nonetheless, many of the thinkers we've been discussing make claims about the function of episodic memory and making sense of this is an essential part of the arguments found in this paper.

In this chapter, I'll review the concept of function and outline some contemporary ideas about the function of episodic memory. Many of these proposals look at the function of episodic memory from an etiological perspective, and are consistent with what we've referred to as hybrid theories of memory—theories that maintain episodic memory involves a memory trace, but allow for some constructive processes. In the context of our current discussion, however, I think that a causal-role account of function is a better method of assessing the function of episodic. The main proponent of a causal-role analysis of the function of episodic memory we've seen is De Brigard (2013), who concludes from his analysis that episodic memory is not, in fact, a broadly preservative cognitive capacity. If we incorporate the necessity of a memory trace,

which I've argued for in previous chapters, however, I think a causal role assessment of episodic memory does support the claim that episodic memory functions as a broadly preservative cognitive capacity.

In section one (4.1), I'll begin with a general discussion of functions. What exactly does it mean for something to have a 'function', and how does this relate to a biological-psychological system like episodic memory? I'll also look at two of the dominant ways of assessing functions: etiological accounts of function, that look to explain why something exists, and causal role accounts of functions, that look to explain how something contributes towards the system within which it exists.

In section two (4.2), I'll look more specifically at the function of episodic memory: what does episodic memory do? Here, I'll go over some of the general features of episodic memory and assess their functional implications. I'll also look at three contemporary accounts of the function of episodic memory. All three accounts are consistent with the view that episodic memory is a broadly preservative capacity, though they propose widely different reasons for why episodic memory might be adaptively advantageous.

In section three (4.3), I'll discuss some of the similarities and differences between these accounts in more detail. While each view focusses on a unique attribute of episodic memory, their functional assessments share some common features. In some cases, they also propose different structures for the episodic memory system. Here, I will make the case that these three accounts are compatible, but that none of them offers a proper etiological account.

In section four (4.4), I'll make the case for why I think a causal-role account of function is better suited for our current discussion. Considering this, De Brigard's (2013) causal-role assessment of the function of episodic memory presents a problem, since his analysis concludes that episodic memory does not function as a broadly preservative capacity. I'll argue, however, that if we incorporate the necessity of a memory trace into a causal-role method of functional analysis, this supports the claim that episodic memory functions as a broadly preservative cognitive capacity.

Thus, regardless of which method of functional analysis we use, the function of episodic memory is consistent with being broadly preservative.

4.1 Functions

Throughout this paper, we've been referring to the function of episodic memory as being either one of a 'broadly preservative' capacity, as in the case of traditionalists, or as a part of a more inclusive cognitive system that is responsible for episodic hypothetical thinking, as in the case of simulationists like De Brigard and Michaelian. Although we haven't discussed the topic of function in detail, these are effectively claims about the function of episodic memory—what is it, exactly, that episodic memory does?

To approach this question more precisely, it's worth taking a step back and asking another question: what do we mean to say when we make claims about the function of episodic memory? What does "its function" mean?

Generally, when someone asks about the function of a certain thing, we take it that they mean something like its purpose, or its use—what exactly does 'the thing' do? For the most part, we're using function in this same sense when we talk about the function of episodic memory; however, determining precisely what episodic memory *does* is a complex issue.

For starters, consider a simple example: As I'm writing this, I'm sitting at my computer, the window cracked slightly for fresh air, with my trusty coffee cup perched beside me on a stack of loose papers. If someone who had never seen a coffee cup, perhaps someone from a distant culture that doesn't drink warm caffeinated beverages, were to see me, such a person might be tempted to ask: what is that curious thing perched beside you on that stack of papers (they'd probably be curious about more than just my coffee cup, but bear with me)? In response, I would likely say: "it's my coffee cup, duh". Realizing, however, that that person is probably unfamiliar with what a coffee cup is, I would soften my tone and elaborate that it's a portable vessel designed for consuming hot beverages. My observer might then point out— "but it's empty, why is it sitting there?", to which I would respond that although my coffee cup was designed for

consuming hot beverages, currently it's being used to keep that stack of loose papers from shifting about in the wind. In fact, it can be used for many things other than drinking hot beverages: it could be used for drinking cold beverages, it could be used for watering my plants, to store spare pens and pencils, or as it's currently being employed, as a paperweight. "Ah" my ever-inquisitive observer might say, "so then why would you say it's for consuming hot beverages when it has so many uses?".

In this exchange, my observer raises an interesting point: when something has many roles or serves multiple purposes, how do we determine which of these is its function? In this specific case, the question is easily answered, provided one is familiar with coffee cups. We know that each feature is incorporated in the design for a specific reason: the cup is hollow so it can contain liquids, it has a handle so that hot contents won't burn the handler's fingers, it's made of ceramic to insulate its hot contents, etc. All of these are reasons for why a coffee cup has the features that it does, and of course we know this because humans designed coffee cups for precisely this purpose[10]. But, what about the functions of things that humans didn't' design?

Just as the term function can be used to explain the features of man-made artifacts, like coffee cups, it can also be used to describe biological traits and systems: what is the function of the human eye, for example, or the beaver's tail? While the term 'function' retains a similar meaning in both contexts, determining the function of a biological system isn't quite as straightforward as that of a coffee cup, primarily, because humans didn't design these structures.

4.1.1 Functions in biology

The reason functions are of interest in biology is that they provide us with an explanation—a rationale for why something is the way that it is. As with the coffee cup, once we understand its function, we can make sense of why it has the features that it does. This also gives us a better understanding of when things aren't working as they typically do—cases of malfunction. If the cup is leaking, for example, we know that something isn't working as intended, and we have a

[10] Broadly speaking, philosophers have individuated artifacts in terms of the intentions of the makers, or their current uses, or both factors.

functional understanding of its features which allows us to accurately assess the problem. In the case of a biological system, like episodic memory, however, determining the function becomes more complicated.

Unlike coffee cups, humans didn't design the episodic memory system, or any biological systems for that matter. Of course, we're aware of some of the ways that we've come to use episodic memory: reexperiencing past episodes, remembering specific events, *perhaps* imagining how things might have gone differently, or how they might go in the future. But, just like the coffee cup can have many uses outside of that which it was designed for, so too it's likely that we've come to use our faculty of episodic memory for many purposes other than that which it was originally designed for. Even claiming that a biological system was "designed" in the first place is stretching the analogy since, effectively, these systems weren't *designed* by anyone, but rather came about as the product of natural selection. This is an important distinction in the study of functions that has led many philosophers to differentiate between artifact functions (like the coffee cup) and biological functions (like the human eye)[11] (Garson 2019, p. 31).

So, when we talk about the function of a biological system, like episodic memory, we're really referring to an explanation for why that thing exists. Specifically, Garson argues, we're referring to a causal explanation for how that thing has come to be—effectively, "functions are just condensed causal explanations" (2019, p. 15). Technically, this kind of causal explanation is called an *etiological* explanation, since it is concerned with the causal history of a biological feature. There are several varieties of etiological accounts found in the literature, but the most well-known are the selected effects theories (Neander 1983; Millikan 1984). Although there is some disagreement between these theories as to how exactly these functions should be delineated, they all agree that accounts of biological functions provide causal explanations for biological traits.

What's important for understanding the etiological function of episodic memory, then, is that the function, effectively, explains why we have episodic memory. Since memory is a biological

[11] There are some interesting cases that incorporate features of both artifact and biological function: like traits bred into specific animals or cases of gene selection (Garson 2019, p. 31)

system, presumably this discussion doesn't involve an explanation that appeals to intelligent design, but rather to some process of natural selection. In this sense then, a question of the etiological function of episodic memory amounts to an explanation for why humans have episodic memory: why did they develop episodic memory in the first place, and why has this capacity persisted?

While the etiological method is the dominant way of assessing the functions of biological systems, it is not the only way, nor does it come without criticism. In effect, etiological accounts are at their best when there is a clear line of sight between a trait's evolutionary development and its contemporary role. This is consistent with the underlying assumption that a trait's current function is consistent with why the trait was selected for in the first place, or why is has persisted (continued to be selected for). When there is a discrepancy between a trait's current role, and its evolutionary development, however, an etiological account becomes less intuitively appealing. One noted difficulty for the etiological account of a trait's function, for example, is explaining vestigial traits—traits that no longer perform their original function (Griffiths 1993). In such cases, traits can persist without a function, or be exapted for an altogether different function. Similarly, it's possible for a trait that was never selected for in the first place, a product or an accident, to be exapted for a novel function. While an etiological account can explain the trait's original function, it can't explain the current function, and so we need both to understand the discrepancy. So, the etiological account is limited in that way.

An alternative account of a trait's function would be Cummins-style functional analysis, which looks to explain complex capacities by breaking them down into simpler systems (Cummins 1975). Accordingly, a given trait is assigned a function based on how it contributes towards the system in which it exists; this system, in turn, is assigned a function determined by its contribution to the overriding capacity. On this view, a functional analysis of the coffee cup doesn't need to reference its particular causal history, and it may be sufficient to provide a functional explanation that accounts for a coffee cup as a paperweight, or a watering can, depending on its current use within the system in which it is being employed (i.e. paper storage, or plant care). While this way of ascribing function may not tell us about the evolutionary origin

of a given capacity, it does provide more flexibility in assessing how a given capacity or feature is currently being deployed.

We'll return to this discussion of the different ways of determining a function, and look at how this relates to episodic memory, in section four. For now, let's look at some potential reasons for why humans might have the capacity of episodic memory.

4.2 What does episodic memory do?

Why exactly do we have the capacity of episodic memory? Before the multiple memory systems hypothesis gained traction, which saw cognitive memory systems separated into five distinct systems—procedural, perceptual-representational, primary (working), semantic, and episodic memory—memory was generally thought to be a unitary capacity (Klein et al. 2002). This corresponded with the view that human learning was also a unitary system, and that memory (i.e., the storage and retrieval of information) was an integral part of it.

Accordingly, the same process of encoding, storage, and retrieval, applied to all forms of memory. On this model, at least intuitively, one might think that episodic memory formed somewhat of a foundation. From the vivid representations made available in episodic memory, other, more specified, pieces of memory could be derived: remembering how to ride a bike, for example, might entail the process of calling to mind past experiences of bike riding, or remembering the capital of France might involve calling up the memory of sitting in eighth grade geography class and isolating that particular fact. Over time the vividness of the representations may recede, but important facts and procedures would be retained.

While this conception of how memory works may appeal intuitively, there are several reasons for thinking that this isn't how memory systems actually work. For starters, it's come to be understood that these different memory systems are independent, or "functionally isolable" (Klein et al. 2002). The workings of procedural memory probably have little, if anything, to do with those of episodic memory, and while episodic and semantic memory have more in common, they too come apart in important ways (Tulving 2002). If we accept the multiple memory

systems hypothesis, then it's unlikely that these memory systems constitute a unitary system of which episodic memory is foundational.

Another reason for questioning this model has to do with the evolutionary relationships between these different memory capacities. Since the etiological function of episodic memory would need to appeal to its evolutionary origin, it's important to consider the different time frames of these memory capacities. There is some debate on this subject, with Tulving arguing that episodic memory is evolutionarily recent and uniquely human (2002), while others have argued it has a longer history, and that the core properties of episodic memory can be found in mammals, non-human primates, and certain birds (Allen & Fortin 2013). Despite these disagreements, however, there is a general consensus that episodic memory is evolutionarily more recent than, and presumably developed from, semantic memory (Michaelian 2016b). If this is the case, then it's difficult to make sense of how a more recently evolved cognitive capacity could be functionally prior to its evolutionary ancestor.

If we assume, then, that semantic memory is evolutionarily prior to episodic memory, why did this capacity to store and recollect representations of past experiences evolve? What added benefit does episodic memory afford us? In accordance with a selected effects model of function, an account of the function of episodic memory would need to appeal to certain features of episodic memory that are adaptively advantageous and that would allow for such a cognitive capacity to be selected for and developed over time.

In his earlier attempts at defining episodic memory, Tulving (1972) suggested that episodic memory was responsible for storing episode specific details, what he referred to as the 'where-what-when' of experiences. In such a way, episodic memories would function as an index to past experiences. Unfortunately, this kind of content is easily available in semantic memory. In fact, most of the content that is often associated with episodic memory is just as easily stored through some combination of semantic and procedural memory—in effect, we don't need episodic memories to remember the capital of France or how to ride a bike. This lead Tulving to revise his definition of episodic memory to include *autonoetic consciousness*—the first person, experiential, mental capacity associated with self-awareness over subjective time—which does

seem to be unique to episodic memory (1983). This, however, still leaves open the question of why we have this capacity.

There have been several attempts at accounting for why we might have this capacity of vividly recollecting personal episodes of past experience. Let's look at some contemporary suggestions.

4.2.1 Social decision making—The Scope Hypothesis

One possible explanation is that memory systems evolved to facilitate decision making. Along these lines, Klein, Cosmides, and Tooby (2002) assess the function of multiple memory systems from an adaptationist perspective, focusing specifically on trait judgments. They put forward a view they refer to as the *scope hypothesis*: that episodic memories place boundary conditions on semantic memory generalizations.

Since memory allows organisms to retain information from their experiences, it is inherently linked with the process of learning (Tulving 1995). As we saw in chapter two, psychologists have often used learning paradigms (the DRM paradigm, Loftus false memory paradigm) to study memory. Klein, Cosmides and Tooby suggest the same connection applies between *decision rules* (any mechanism that alters behaviour on the basis of information) and memory. From an adaptationist perspective, it follows that the pressures of natural selection only apply to those features that are expressed, and it is through decision making that organisms expose their learning (Klein, Cosmides, and Tooby 2002, p. 306-7).

In their account, Klein, Cosmides, and Tooby assume a distinction between *inceptive memories* and *derived memories*. Inceptive memories are representations of the world stored as they were encoded at their inception, whereas derived memories are higher-order representations derived from inceptive memories (2002, p. 310). Accordingly, inceptive memories correspond with episodic memories, and derived memories correspond with semantic memories. They theorize that different decision rules require different information from different sources of memory, and in order to efficiently access the appropriate information requires an effective retrieval mechanism, or "search engine" (2002, p. 309). Focusing specifically on *trait judgments*—

judgments about personality traits— they propose that decision making rules in this domain follow a particular retrieval pattern that they refer to as *the scope hypothesis*.

Essentially, people form trait summaries based on a database of experienced past episodes—if person *x* was rude to person *y* in nine out of ten past encounters, then person *y* will associate the trait of being rude with person *x*. These trait summaries are derived memories: they are stored generalizations that can be called upon to expedite the complex dynamics of social interactions. In order for these generalizations to be effective in decision making, however, they require boundary conditions—conditions that mark the extent to which the generalization persists. According to the scope hypothesis, episodic memories provide these boundary conditions. Thus, the effective use of a trait summary required the generalizations of semantic memory coupled with specific boundary conditions—episodic memories.

To see how this works, consider the following example (my example, not theirs). Imagine you have a good friend John whom you know to be particularly *easy-going*—well-liked and accommodating practically to a fault. This describes John perfectly, except for the few times when you've witnessed someone openly disparage religion in his presence. John's demeanor completely changed and he became very out-spoken, to the point of almost being aggressive. In this case your character summary of John might have him indexed as particularly *easy-going*, but you'd recall those specific instances of when that generalization didn't apply—namely in conversations about religion. Accordingly, you'd have a trait summary of John as *easy-going* stored in semantic memory, but calling up that trait summary would trigger boundary conditions: episodic memories of instances where John was not so *easy-going*. This way, you can effectively make decisions in social contexts: say, for example, if you had to decide on a friend to bring with you to a lecture on atheism, you might quickly rule John out as a potential candidate. This is how Klein, Cosmides, and Tooby (2002) envision episodic memories as providing boundary conditions for the scope of generalization.

They tested this hypothesis. In several studies, they had participants rank trait adjectives that applied to themselves, and to a significant other (their mother). They then had participants presented with a series of tasks: *describe* (whether the trait applied to them), *define* (define the

trait), *recall* (recall an instance of that trait) (Klein, Cosmides, & Tooby 2002, p. 316-18). The subjects where then presented with two subsequent tasks relating to a pair of trait adjectives: either the same trait repeated (e.g. *rude* and *rude*), or a trait and it's opposite (e.g. *rude* and *polite*). They found that when the *describe* task preceded the *recall* task, and the trait pair contained opposites, performance on the recall task was significantly faster than when the *define* task preceded the *recall* task, or when they trait pair repeated. These results were concurrent with the scope hypothesis prediction that a personal trait summary would prime recall of boundary conditions (episodic memories of when the trait didn't hold), which makes for a faster performance on the *recall* task (since the episodes were just primed). It was theorised that traits that didn't have summaries (generalizations that aren't well-established in semantic memory) would prime exemplars of the trait, as opposed to boundaries on the generalization (since there's no generalization). This would predict that when the trait adjective was repeated (e.g. *rude* and *rude*), performance time would be faster when the *describe* task preceded the *recall* task. These results were not confirmed in the first study—it appears that in the self-study, all traits had summaries, regardless of how the participants ranked their level of applicability—apparently, "we know ourselves well" (Klein, Cosmides & Tooby, 2002). For this reason, the second study involving the participant's mother was included. In this case, it was theorized that some traits would have well-developed summaries, which primed scope boundaries, and others would not, which would prime exemplars. The study confirmed these predictions.

Though the Klein, Cosmides, and Tooby (2002) study focuses specifically on decision making with regards to trait judgments, they suggest that the scope hypothesis might apply to other categories of decision-making as well. These include script-based learning (Schank 1982), schema-driven memory (Graesser 1982), and object categorization (Nosofsky, Palmeri, & McKinley 1994). It's also worth noting that Klein, Cosmides, and Tooby offer the scope hypothesis as one potential adaptive advantage of episodic memory, and not necessarily as its primary advantage. Still, this gives us an idea of how episodic memory might feature in decision making.

This, then, is one possible suggestion for how episodic memories are adaptively advantageous and, hence, for the function of episodic memory. Boyer (2008) provides another suggestion.

4.2.2 Affective recall

An important feature of memory, in general, or the ability to retain information, is that it affords humans the ability to use past information to influence current decision making—and, if we're going to make claims about the evolution of episodic memory, this feature needs to be taken into account (Suddendorf & Corballis 1997). If we consider that much of the content made available in episodic memory, the 'where-what-when' as Tulving puts it, could just as easily be made available in semantic memory, and that procedural memory accounts for retaining the content of processes, this limits the scope of what kind of learning might be derived from episodic memory. Limiting the scope as such leaves us with what's unique about episodic memory—namely its phenomenological quality. Boyer follows this line of reason and suggests that a possible explanation for why we have episodic memories is affective recall—or, remembering emotions (2008, p. 18-20).

Boyer argues that current functional accounts of episodic memory focus on decision making (Boyer & Barett 2005, Suddendorf and Corballis 2007, Cosmides and Tooby 2000), which can effectively be explained without appealing to any unique features of episodic memory. What's unique about episodic memory, according to Boyer, is the affective component of imagery associated with "re-experiencing the past" (Boyer 2008, p. 16-7). In relation to this, he raises the question of pro-social human behaviour: why is it that humans have tended towards engaging in other-regarding behaviours, when these are most often at odds with individual fitness? In other words, people will behave in such a way that is counter-productive to their short-term self-interest, and favours longer term, social benefits, even to the extent that, in some cases, those benefits won't accrue on an individual level. Boyer argues that these pro-social behaviours are not the direct product of human rationalization, but rather the product of intuitive, often emotional, preferences (Fessler 2001). Accordingly, emotions such as shame, pride, and guilt may provide the impetus for pro-social behaviour where rational self-interest falls short.

It is along these lines, Boyer argues, that emotive characteristics of episodic memory might come into play. What's unique about episodic memory, according to Boyer is the imagistic, emotional quality of re-experiencing episodes. His suggestion, then, is that the human ability to recollect

emotions, what he refers to as *affective recall*, may be responsible for contributing to pro-social human decision making. Whereas rational decision making may guide humans on an opportunistic course of action, recollecting the emotions associated with past experiences—shame, pride, guilt, etc.—might influence humans towards seeking a more pro-social outcome. Importantly, on his account, it is the autonoetic quality of episodic memory that allows us to re-experience past emotions with the same vivacity they were initially felt. In this process of decision making, then, episodic memory would provide a tool for accessing the affective qualities of past experiences.

From an evolutionary standpoint, we can see how this would work out: as human societies became more complex, evolutionary pressures would favour traits that contribute towards group cohesion. These traits might be less geared towards individual fitness and more geared towards group fitness. In this scenario, it's very plausible that cognitive capacities that facilitated pro-social human behaviour, would, over time, be selected for.

In support of this reasoning, he cites various studies on the roll of emotions in decision making (Ursu & Carter 2005). Arguing that even though the semantic content of episodic memory might be susceptible to distortions (Roedigger and McDermott 1996), the recollection of past emotions is much less flexible (Levine & Bluck 2004).

This, then, is another possible explanation for why humans developed episodic memories—to contribute to pro-social decision making. Let's now look at one final possibility.

4.2.3 The Communicative function of episodic memory

A third candidate for the function of episodic memory has to do with its role in human communicative interactions. Mahr and Csibra (2017) propose that episodic memories allow us to exert epistemic authority in the claims we make about the past, most notably in terms of human social engagements. This functional account, they argue, is preferable to other functional accounts of episodic memory because it is best equipped to make sense of the tension between the *preservative* and the *generative* features of episodic memory.

Mahr and Csibra make a distinction between what they call *event memory* and *episodic memory*, which is important for understanding their account (2017, p. 4). On their view, the structure of episodic memory can be understood by making a distinction between the stored information involved in episodic memory (i.e., the memory trace), a "scenario construction system", and episodic memory (2017, p. 2). Roughly, the scenario construction system combines stored information (the memory trace) with relevant semantic information to construct a detailed scenario that extends over time and space—this process, they suggest, might involve a similar Bayesian inference as De Brigard's (2014) account detailed in chapter three. When these constructed scenarios are conceptualized as first-person, past experiences, they constitute *episodic memories*; otherwise, they can still influence behavioural decisions, but merely constitute *event memories.* Accordingly, episodic memory has an inherently metarepresentational format, and entails a particular epistemic attitude—that the event in question was experienced first-hand. As a result of this, *event memories* generate a belief that a given event occurred, and *episodic memories* provide epistemic justification for that belief. Thus, episodic memory is what they call *epistemically generative*—it provides justification for holding a belief (Mahr & Csibra 2017, p. 6).

This epistemic generativity, according to Mahr and Csibra, is an important feature for making sense of the function of episodic memory. On their functional analysis, epistemic generativity allows individuals to represent the reasons for their beliefs, which can be given in testimony. The autonoetic feature of episodic memory, in turn, determines which beliefs the individual has epistemic authority over. As Mahr and Csibra see it, this can partially explain the tension between preservative and generative accounts of memory: how a functional episodic memory system can produce both false and veridical representations (2017, p. 13). Since episodic memory is epistemically generative, it functions to provide justification for one's beliefs. When a person has publicly committed to a certain version of events, they take on epistemic responsibility for the truth of that version of events (2017, p. 10). On the one hand, episodic memory functions in support of pre-existing beliefs, which explains phenomena like recollective my side-bias (Mercier & Sperber 2011), selective remembering, and motivated forgetting (Anderson & Hanslmayr, 2014)—in these cases, episodic memory is more about convincing

others of our beliefs. On the other hand, the complex dynamics of human social communication would lead to an evolutionary "arms race", wherein epistemic vigilance becomes a crucial feature of episodic memory: individuals have an interest both in not being deceived, as well as in not discrediting their epistemic authority by being disingenuous, both of which benefit from a veridical representation of the past (2017, p. 13).

As to why the episodic memory system evolved in such a way, Mahr and Csibra suggest one possible explanation: social commitments. Although referencing past events can have a myriad of important uses, social commitments are unique in that they represent particular events which can have tremendous impacts on the parties involved. Being able to track and reference these communicative interactions, and hold those involved to their commitments, is integral for society to function. What's more, social commitments always entail effects in the future that are contingent on representing them as having taken place in the past (Mahr & Csibra 2017). Thus, as one possible explanation, though not necessarily an exhaustive one, Mahr and Csibra propose that episodic memory functions as a communicative tool for moderating social commitments.

These, then, are three possible functional explanations for the evolution of episodic memory—to provide boundaries on the scope of generalizations, as affective recall for pro-social human behaviour, and to communicate social commitments. In the next section, we'll discuss these views a bit more.

4.3 About these accounts

In the previous section, we looked at three proposals for the function of episodic memory—placing boundary conditions on the scope of generalizations, affective recall, and communicating social commitments. These proposals assess the function of episodic memory primarily from an etiological perspective and are consistent with the claim that episodic memory is a broadly preservative capacity. Despite this, they come up with quite different reasons for thinking why episodic memory might be adaptively advantageous. While each proposal offers a unique reason as to how, or why, episodic memory might have evolved, these ideas share some common features: namely, the relation between episodic memory and the human social domain, and the

role of episodic memory in decision making. There are also some important differences between these accounts, particularly in terms of how they propose structuring the episodic memory system.

4.3.1 The etiological function of episodic memory

All three of the functional analyses of episodic memory outlined in the previous section approach their analyses from an etiological perspective. These views are also consistent with hybrid theories of memory.

Klein, Cosmides, and Tooby explicitly conduct their functional analyses from what they refer to as an *adaptationist* perspective. They aim to treat a cognitive system, like episodic memory, as a naturally selected system and assess its function using the same framework as would be applied to the evolution of a biological system (Klein, Cosmides, Tooby 2002, p. 307-8). Similarly in their functional analyses, both Boyer (2008) and Mahr and Csibra (2017) look to identify unique features of episodic memory that have been adaptively advantageous—features that account for why our capacity of episodic memory might have been selected for and persisted. Both accounts agree that what's unique about episodic memory is the quality of autonoesis, though they think this has been adaptively advantageous for different reasons: affective recall on the one hand (Boyer 2008), and epistemic justification on the other (Mahr & Csibra 2017). It is clear, then, that these three accounts are interested in determining the etiological function of episodic memory[12]—a function that explains why episodic memory was selected for, and why it has persisted.

It's also the case that all three of these proposals are consistent with what we've been referring to has hybrid theories of memory—theories that maintain episodic memory is a broadly preservative cognitive capacity, but contains some generative content. In the cases of Klein, Cosmides, and Tooby (2002) and Boyer (2008) this is straightforward. Klein, Cosmides, and

[12] Mahr and Csibra make it explicit that their account is primarily aimed at the function of episodic memory as a mature system, and that they aren't directly studying its evolutionary history (2017, p. 2). Their account does, however, assess the function of episodic memory as an *adaptive function* (2017, p. 6). A fair reading, then, might be that their functional analysis incorporates both etiological and causal-role components.

Tooby suggest that episodic memories represent specific stored experiences that provide boundary conditions on semantic generalization—this is an inherently preservative process. Boyer (2008) suggests that episodic memories function in support of affective recall—remembering emotions—which is also an inherently preservative process.

In the case of Mahr and Csibra's analysis, the preservative status of episodic memory is less clear. According to Mahr and Csibra, episodic memory has a meta-representational structure: not only does it represent a specific episode (content), but it represents it as having been experienced first-hand (format) (2017, p. 6). Accordingly, event memory is what they refer to as *epistemically preservative*, whereas episodic memory is *epistemically generative* (2017, p. 5). Also, their account allows for significant flexibility between representing what actually happened, and what one has committed to as having happened (2017, p. 12-3). This isn't necessarily a problem, however, since it offers a unique way of making sense of the preservative and generative features observed in episodic memory. What's important on their account is that, as a result of epistemic vigilance, this process tends towards veridical representations—we have a vested interested in representing things as they actually occurred. Considering that the *content* of episodic memory is derived primarily from the stored content of past experienced episodes (memory traces), and the system has tended towards producing veridical representations, then I think it's fair to conclude that this qualifies as a broadly preservative capacity.

We can see, then, that these three accounts assess the function of episodic memory from an etiological perspective, and are consistent with it functioning as a broadly preservative capacity. Despite this, they offer different suggestions as to precisely what that function might be. There are, however, some important similarities in what they propose.

4.3.2 Functional similarities

While the three accounts discussed offer different suggestions for the function of episodic memory, they share some common ideas about the areas in which episodic memory has been advantageous.

One idea that's common in all three of these accounts is the relation between episodic memory and human social dynamics. Klein, Cosmides, and Tooby's (2002) study assesses how episodic memories might provide boundary conditions on the scope of generalizations with regards to trait summaries. Part of their reasoning for this is that complex human social dynamics requires the efficient sort of decision making that the *scope hypothesis* purports to provide. While they suggest this retrieval pattern might exist in other areas—script-based learning (Schank, 1982), schema-driven memory (Graesser 1982), and object categorization (Nosofsky, Palmeri, & McKinley 1994)—they also stress their suspicion that the most important functions of episodic memory have to do with social interactions (Klein, Cosmides, & Tooby 2002, p. 310). Boyer's account of the function of episodic memory argues that *affective recall*, our ability to recollect emotions, is a unique feature of episodic memory, and that this may have evolved because of the way emotions influence pro-social human decision (2008, p. 18-20). Accordingly, the more cohesive and pro-social a society, the more likely members of that society would thrive, evolutionarily speaking. Mahr and Csibra's idea is that episodic memories provide epistemic justification for our beliefs, and that this capacity to reference particular past events is notably relevant for managing social commitments—in a sense, the very fabric upon which human society is built (2017). All three proposals, then, share this notion that episodic memory functions in relation to human social dynamics.

This is an interesting feature of these views since, at least intuitively, episodic memory is a very personal, almost intimate, capacity: it allows us to represent personal past experiences in the privacy of our own minds. Often, the reminiscing of past experiences is associated with solitude and contemplation, as opposed to social endeavours. Despite this, however, if we consider the influence of social dynamics in terms of human evolution, it probably shouldn't be all that surprising. Humans are, after all, social animals. From an evolutionary perspective, probably many of our cognitive capacities evolved in tandem with our evolutionary needs as members of a social group. Certainly, most of our more advanced cognitive faculties relate in some way to social needs and requirements. The ability to conceive of and conceptualize past events that episodic memory provides, while undoubtedly a solitary pastime, is probably equally, if not more, suited for communicating the details of these events, which is an inherently social process.

A similar point might be made about another common theme that these proposals share—decision making. Both Klein, Cosmides and Tooby (2002), and Boyer (2008) explicitly propose that the function of episodic memory pertains to human decision making. While in and of itself, we might think of episodic memory as a capacity more geared towards providing the content upon which decisions are made, as opposed to being part of the decision-making process, from an evolutionary perspective this distinction probably isn't all that important. It is likely that we retain the information that we do precisely because it is, or has been, useful. And a straight-forward gauge of this is how it's been employed by our decision-making processes.

Mahr and Csibra's proposal is not explicitly about human decision making, but rather about providing the ability to communicate epistemic justification for our beliefs. Interestingly, they suggest that *event memory* can guide human behaviour and so likely does influences our decision-making processes. On their account, however, *event memory* is not the same as *episodic memory*, wherein episodic memory requires the further epistemic attitude associated with autonoesis that, on their view, makes it unique.

This raises another interesting feature of these analyses, which is that they suggest different structures for the episodic memory system.

4.3.3 Architecture of episodic memory

Between these views—notably between Klein, Cosmides and Tooby (2002), and Mahr and Csibra (2017)—there is a difference in the way the architecture of episodic memory is presented.

While Klein, Cosmides and Tooby (2002) don't directly discuss the architecture of episodic memory, they make an important distinction between what they refer to as *inceptive* and *derived* memories. Inceptive memories are stored as they were encoded on inception, whereas derived memories are generalizations *derived* from inceptive memories. In the context of the *scope hypothesis*, inceptive memories are like episodic memories and derived memories are a kind of semantic memory. Accordingly, at least in terms of trait summaries, derived memories are generalizations based on episodic memories. This implies that episodic memory has a similar

structure to the 'intuitive' model we outlined at the beginning of section two—that episodic memories provide the foundation for other memories. While we dismissed this model as it relates to the evolution of a unitary memory system, the model might still be applicable to the relation between specific cases of episodic and semantic memories. That is to say, how memory systems evolved doesn't necessarily have to restrict how we cognitively employ them. Even if, for example, semantic memories are evolutionarily prior to episodic memories, it's possible to have other semantic memories, perhaps second or third order semantic memories, derived, in turn, from the content of episodic memories.

This structure of episodic memory is quite different from the model provided by Mahr and Csibra (2017). On Mahr and Csibra's account, they distinguish between *event* memories and *episodic* memories. *Event* memories are the product of a scenario construction system that generates a temporally extended simulation of past events, which is created from stored information (memory traces) and semantic inferences. The way they describe this process is similar to the Bayesian inferencing proposal De Brigard (2013) suggests for episodic hypothetical thinking (chapter 3). *Episodic memories* on their account, however, are not the equivalent of *event memories*, but include the feature of being autonoetic. Thus, episodic memories are event memories that incorporate a specific epistemic attitude—that the episode in question was witnessed, first-hand.

If we consider these different architectures from an evolutionary standpoint, then we can see how this might complicate an etiological assessment of the function of episodic memory. On Mahr and Csibra's account, the content of episodic memory proceeds through three stages: first information is stored (the memory trace), then an event scenario is constructed (event memory), which is the product of the memory trace and other semantic memories, and finally an epistemic attitude is incorporated with the event memory to generate the representation of a first-person experienced past event (episodic memory). From an evolutionary perspective, it is unlikely that these three processes developed in conjunction as the solution to a single challenge. Each of these stages is a highly complex process in its own right, and probably developed over an extensive period of time. Through this process, it's probable that selection pressures also changed and developed. More likely, then, each part of this process was selected for over time

because it was adaptively advantageous for slightly different reasons, or for similar reasons in different contexts.

Having these different structural architectures complicates how to go about determining the function of episodic memory.

4.3.4 Mutually inclusive functions: capacities?

The difficulty with ascribing functional assessments when such assessments entail different structural breakdowns of the episodic memory system, is that these accounts run the risk of talking past one another. In effect they can focus on, and in this case I would argue they are focussing on, different features of episodic memory in their structural analysis. The upside of this discrepancy, however, is that it doesn't put these accounts directly in contention. As we discussed previously, artifacts, as well as biological systems, can perform a variety of different functions in different contexts. These functions, however, are not proper functions in the *etiological* sense, they aren't necessarily the reason the thing exists, or has persisted, but rather they are *capacities* in which it can, or does, function.

Consider how the different functional analyses we've outlined focus on different features of episodic memory. Providing boundary conditions on the scope of generalizations, for example, likely doesn't require the same level of meta-representation as it does for episodic memory to qualify as epistemically generative. According to Mahr and Csibra (2017) *event memory* is sufficient for informing behavioural decisions, and so Klein, Cosmides, and Tooby's (2002) functional assessment might more aptly be describing the function of what Mahr and Csibra have labelled *event memory*, a sub-system that is constitutive of, but not identical with, episodic memory. Although Boyer doesn't propose a detailed architecture for the episodic memory system, he seems to maintain that *affective recall* is strictly preservative, and so according to Mahr and Csibra's (2017) model of episodic memory, a basic memory trace might be sufficient to perform such a function. It's easy enough to make the case that each of these accounts is referring to a different feature of episodic memory. The difficulty this presents, then, is that,

depending on how we carve up the episodic memory system, these functional analyses might be explaining different aspects of episodic memory.

The upside, however, of this discrepancy is that these accounts might not be mutually exclusive. If we return, for a moment, to our earlier discussion of coffee cups, you'll recall that a coffee cup can function in many different capacities. By focussing on specific features of the coffee cup's design, assessments of the coffee cup's function might generate quite different conclusions: by focusing on the handle, for example, one might determine that the coffee cup was designed to be held, or by focussing on the coffee cups weight, one might determine that the coffee cup was designed not to be tipped over. These, in turn, could contribute towards the coffee cup functioning to water plants, or as a paperweight. These functions, of course, are not mutually exclusive—they are both consistent with the design of a coffee cup, though neither provides a proper etiological account of why the coffee cup exists as it does.

In the same way, assessments of the function of episodic memory that focus on different features will likely conclude that episodic memory has different functions. If we focus on the emotive quality of episodic memories, we might determine that it functions in support of *affective recall*; if we focus on the epistemic qualities of episodic memory, we might conclude that it is *epistemically generative*; or, if we focus on the instance specific details, we might conclude that it works in conjunction with semantic generalizations. And these functions would not be mutually exclusive, either—episodic memory could function in all three of these different capacities. In effect, given the evolutionary history of a system like episodic memory, it's quite likely that such functions would have a high degree of compatibility. It's possible that a system that adapted for one function, can be exapted for another. These functions wouldn't necessarily have to be synchronic, either, in order to be jointly responsible for the current features of episodic memory. This would be consistent with the authours' own assessments of their functional accounts—none claim to be providing an exhaustive etiological account of the function of episodic memory[13].

[13] Klein Cosmides and Tooby (2002, p. 324), think the *scope hypothesis* is only one potential function of episodic memory; Boyer (2008, p. 20) suggests affective recall is just one possible way episodic memory functions; and Mahr and Csibra (2017, p. 13-4) leave open the possibility of episodic memory contributing to more than just social commitments.

The difference, however, between the functional capacities of episodic memory and those of the coffee cup is that we can't easily determine the relationship between these capacities and the proper etiological function of episodic memory. In the case of the coffee cup, the different features that are associated with different functions all come together within the design of a coffee cup. This is because coffee cups were designed for a specific purpose—drinking coffee (maybe they were originally designed for drinking tea, but the point's the same). Because coffee cups were designed, as such, there is a consistency between the concept of a coffee cup and its structural instantiation: the concept of a coffee cup is neatly constituted by its various features, all of which contribute towards the functional end of drinking a warm beverage. It is unlikely, however, that our episodic memory system has the same degree of consistency between concept and instantiation. The concept of episodic memory is a human cognitive concept that describes a capacity to represent past, experienced episodes, which is instantiated by neural systems that humans didn't design, but rather came about as the products of natural selection. Because of this, there is an inherent distinction between the cognitive capacity we recognize as episodic memory and the neural systems that instantiate it.

It would be unlikely to determine that all of the features that constitute the episodic memory system on a neural level fit neatly within a cognitive functional hierarchy that recognizes episodic memory; or, that these neural systems don't participate in a multitude of other cognitive functions. This is simply a feature of the relationship between cognitive kinds and their neural representation. This, however, makes charting the evolutionary development of a capacity like episodic memory quite complex. Our capacity of episodic memory may have developed from a variety of other cognitive capacities, all of which were likely selected for the different advantages they provided.

That is not to say that there is no merit in studying the etiological function of cognitive capacities, nor that episodic memory doesn't have a proper etiological function, but to draw attention to the interesting distinction between the evolution of a cognitive system, like episodic memory, and how we currently employ it. Part of what we're interested in when we discuss the function of episodic memory is not strictly how it came to be, and what evolutionary pressures

led to its development, but rather what does it do—how does it currently contribute towards our cognitive architecture?

4.4 Broadly preservative function

In the previous section, we discussed some similarities and differences between different functional accounts of episodic memory. One interesting feature that came out of this discussion is that these views are not mutually exclusive: it's very possible that the episodic memory system functions in all three of these capacities (and likely a few others, as well). Despite this, it's unlikely that any of these functional accounts describes the proper etiological function of episodic memory—more likely, they describe how some of the features of our episodic memory system have been, and continue to be, adaptively advantageous. When we ask a question like whether episodic memory is a broadly preservative, or a generative, cognitive capacity, however—that is to say is it more concerned with the storing of information, or the creating of it—we're not just interested in explaining the origin of our episodic memory system. We're also interested in determining what episodic memory does: what role does it currently fulfill in our cognitive architecture.

4.4.1 Biological vs. cognitive

So far in this discussion of function we've been considering episodic memory primarily as a biological system, which ultimately it is, but it's also a cognitive capacity. And while cognitive capacities are certainly biological systems in the broader sense, they are unique in that they are systems dedicated to processing information. Accordingly, the kind of information they process, and how they go about processing it, are important features for differentiating our cognitive systems.

If you recall our discussion from chapter one (section 1.2.1), declarative memory (i.e. cognitive memory that can be expressed in terms of propositional content) is typically broken down into semantic memory (memory of facts and propositions) and episodic memory (first-person representations of past experienced events). While 'memory' in general might be too broad of a

concept to categorize as a cognitive kind, there are good reasons (surveyed in chapter one) for thinking that declarative memory, and its sub-categories of semantic and episodic memory, do qualify for the status of cognitive kindhood.

One way of interpreting the function of a cognitive system, like episodic memory, is by giving an etiological account of its function. In this case, the function of episodic memory amounts to an explanation for why episodic memory exists: why was episodic memory selected for in the first place, and why has it persisted? This functional analysis appeals to the features of episodic memory that have been adaptively advantageous and looks to trace their evolutionary development. As we saw in the previous section, however, in the case of a complex system, like episodic memory, this process becomes complicated. It's likely that our capacity for episodic memory developed from a variety of other cognitive capacities, each of which came to be selected for because it was advantageous for different reasons. Focussing specifically on these reasons is important for establishing possible explanations for why, and how, our capacity of episodic memory developed, but it doesn't paint a complete picture of what it does, and how we currently use it.

An alternative approach to assessing the function of a cognitive system, like episodic memory, would be a Cummins-style causal role account of its function. This kind of functional analysis looks specifically at the role that a system or trait performs in the system within which it exists (Cummins 1975). In the case of episodic memory, then, a Cummins-style causal role account of its function is concerned primarily with determining how our episodic memory contributes towards our cognitive architecture. This is different from an etiological account of the function of episodic memory because it is focussed specifically on what episodic memory does, as opposed to why episodic memory evolved in the first place.

If we consider episodic memory as a cognitive system, then that means that it is a system responsible for the processing, storage, and retrieval of information. How this information is available to us, and how we use it has likely undergone significant developments from when it was initially selected. As humans have evolved, the environments in which our cognitive capacities function have also shifted, and as a result, the kind of information we store, and how

we use that information, has changed as well. As with many human features that evolved for specific reasons but have since been adopted into the complex process of human social evolution, there may be some discrepancies between the evolutionary trajectory of episodic memory and how we currently employ it.

Since cognitive systems are delineated according to the kind of information they process, and how they make that information available to us, when we are considering the function of episodic memory as a cognitive system, what we really want to know about is how it currently fulfills that role. What kind of information does episodic memory make available to us: is it veridical or confabulatory? And, how does it process this information: is it preservative, or generative? In short, is episodic memory actually for remembering? This notion of the function of episodic memory, I think, is more relevant to our current discussion, and I would suggest a Cummins-style, structural account of function is better suited for assessing this (Cummins 1975).

4.4.2 Different functional analyses of episodic memory

The accounts we looked at in section two all approach the function of episodic memory from an etiological perspective. They all look to explain why our capacity of episodic memory has been adaptively advantageous. While each account focusses on a different aspect of episodic memory, all three are consistent with its function being broadly preservative: whether it's preserving the recollections of emotional states (Boyer 2008), episode specific event details (Klein, Cosmides and Tooby 2002), or providing the content upon which one's beliefs are created and maintained (Mahr & Csibra 2017), the common denominator is that episodic memory is a capacity that allows us to store the content of past experiences in-order to generating veridical representations of those experiences.

In contrast, the functional analysis provided by De Brigard, which we reviewed in chapter three, proceeds as a Cummins-style, causal role structural analysis. Instead of focussing on a specific feature of episodic memory that has been adaptively advantageous, he looks to assess the episodic memory system on two levels: 1) how the mechanisms in the system work, and 2) how the system contributes towards the larger system in which it is contained. To these he concludes

that (1) the episodic memory system reconstructs dispersed, encoded information, that is assembled in memory as a *gist-like* representation, in-order to (2) generate probable representations of past experiences. Importantly, these "probable representations" are not necessarily representations of actual experiences, and so episodic memory is not so much a broadly preservative capacity as it is a capacity aimed at generating probable scenarios.

This presents a problem if we think the function of episodic memory is broadly preservative. It also suggests a significant discrepancy with the etiological accounts surveyed in section two. De Brigard's causal role functional analyses (2013), however, precedes his more recent position on the *indispensability of a memory trace* (2020). In his initial functional analysis, De Brigard was not committed to the necessity of a memory trace. As I've argued in previous chapters, however, which is consistent with De Brigard's more recent position (2020), a memory trace is a necessary component of episodic memory. If we take this into consideration, then I think a Cummins-style causal role account of the function of episodic memory is consistent with the notion that episodic memory is a broadly preservative cognitive capacity.

4.4.3 A broadly preservative cognitive function

If we take as an example the episodic memory structure suggested by Mahr and Csibra (2017), which entails three stages from memory trace, to event memory, and finally episodic memory, then, accordingly, the memory trace itself might be the only purely preservative capacity in this process. Without fully committing to this structure, I think it gives us a useful way of parsing out the cognitive capacities involved in episodic memory, and one that is largely consistent with De Brigard's own model (section 3.1.1.2). Of these capacities, it would appear that, by definition, the memory trace is a preservative cognitive capacity.

I've argued in earlier chapters (sections 1.4.2, 2.3.2, 3.3), there are good reasons for accepting that (1) memory traces exist, and (2) that they are an integral part of episodic memory. The first of these points is fairly uncontroversial and was argued extensively in chapter one. The second point is more controversial, with De Brigard (2014) and Michaelian (2016b) arguing that memory traces were not essential to episodic memory. Despite this, De Brigard has revised his

view to argue for the "explanatory indispensability of the memory trace" (2021), which, of the two, leaves Michaelian (2016b), as the only advocate of episodic memory without a memory trace. In chapter three, I outlined several reasons for why I think Michaelian's position is unsatisfactory—ultimately, I don't think we can make sense of episodic memory without acknowledging a memory trace (section 3.3.1).

So, then, if we accept that the memory trace is a preservative cognitive capacity, that it exists, and that it is an integral part of episodic memory, does this mean that episodic memory should qualify as a broadly preservative cognitive capacity?

Framed as such, this might seem to amount to a merely semantic distinction: whether having an essentially preservative cognitive component (the memory trace) is enough to qualify a cognitive capacity (episodic memory) as being broadly preservative. Here, one might argue that simply having a preservative capacity as a constitutive part is insufficient for considering the cognitive system on the whole as being a broadly preservative capacity. Imagination, or hypothetical thinking, for example, likely contain preservative components, but we wouldn't want to consider these to be broadly preservative cognitive capacities. The necessary relation between episodic memory and a memory trace, however, entails a specific causal connection to a specific, experienced past episode. This is importantly different from the kinds of stored content that might feature in imagination or hypothetical thinking in that the content of the episodic memory refers explicitly to the content of the memory trace—they are referring to the same actual experience. This can't be the case for either imagination or hypothetical thinking, since they aren't representing actual experiences. In these cases, the stored content is being used to represent possible, or hypothetical, occurrences, but not actual ones.

The important difference, I think, is a feature of how we employ these different cognitive capacities. In the case of episodic memory, we use this capacity to represent past experienced episodes. This is not to say it's infallible. There are a multitude of well documented memory errors discussed in chapter two; however, for a cognitive system to err in a given capacity is not the same as saying it isn't used, or useful, in that capacity. To make the latter claim we would need to show memory errors to be so pervasive that episodic memory isn't a reliable capacity for

representing past experienced episodes. I don't think any of the evidence presented so far in the dissertation, and we discussed memory errors in detail in chapter two, supports this claim. On the contrary, I think it shows that we have this cognitive capacity to represent past experienced episodes with relative accuracy, which we recognize as being episodic memory.

Imagination, or hypothetical thinking, on the other hand, are not cognitive capacities we use for representing past experienced episodes. These cognitive capacities are used for representing hypothetical scenarios: personal or impersonal situations that might have happened in the past, might be happening in the present, or might happen in the future. Even if these cognitive systems (episodic memory and imagination/hypothetical thinking) share common, or even identical, neural underpinnings, there is an important distinction that needs to be recognized, which is that they function in completely different capacities. Being unable to draw a fundamental distinction between episodic memory and hypothetical thinking would be exceedingly problematic, since it would entail an inability to differentiate reality, or what actually happened, from fiction, or what didn't.

In the cases of memory errors discussed in chapter two, this might be precisely what is happening. Loftus' false memory paradigm (1997) for example might constitute cases of confusing episodic memories with imagination, or hypothetical thinking. Even in these cases, however, where the situation is contrived to blur the distinction between reality and fiction (false testimony, altered photographs, etc.) the vast majority of subjects still get it right (~70%). This, in itself, not to mention the extensive supporting evidence as to the reliability of episodic memory, supports the view that humans are quite successful at making a distinction between what actually happened and what didn't. And, in order to make such a distinction, we must be using our cognitive capacities of episodic memory and imagination, or hypothetical thinking, in different ways: one to represent past experienced episodes, and the other to represent hypothetical scenarios. If this is the case, then these systems must have relevantly different outputs.

Given this, that the memory trace is a purely preservative cognitive capacity, that it exists and is essential to episodic memory, wherein episodic memory represents past experienced episodes

with relative success, I think a Cummins-style causal role structural analysis is consistent with episodic memory functioning as a broadly preservative capacity. Considering De Brigard's two-part causal role assessment: 1) how do the mechanisms in the system work and 2) how does this system contribute towards the larger system within which it exists (De Brigard 2013). We're left with something to the effect of: the episodic memory system uses (1) a memory trace and other semantic inferences to (2) generate a first-person representation of a past, experienced, episode.

A Cummins-style causal role structural analysis of episodic memory, then, is consistent with the etiological accounts of the function of episodic memory surveyed earlier in this chapter. According to both methods, the function of episodic memory qualifies as broadly preservative. Episodic memory is "broadly" preservative, because it does incorporate other semantic information, which is consistent with the constructive features of memory, and with hybrid views—if it were a *strictly* preservative cognitive capacity, it wouldn't allow for any content generation, like the *archival* views detailed in chapter one. It is "preservative" because the content it represents is necessarily connected to information stored from the specific experience being represented. I think it's correct, then, to claim that episodic memory is a broadly preservative cognitive capacity.

4.5 Summary

We began this chapter by looking at the concept of a function, and going over how this might relate to episodic memory. In section one (4.1), we saw that there are various ways of assessing functions. In biology, functions typically explain the reason why a trait or system was selected for and has persisted (section 4.1.2). This causal explanation for traits and systems is known as an *etiological* account of function. While this is the dominant method for assessing the function of traits and systems, it is not the only method. Cummins style, causal-role functional analysis offers an alternative method of determining functions based on how the trait or system fulfills a particular roll within a system, or context, within which it exists.

In section two (4.2), we discussed applying the concept of function to a system like episodic memory. One way of doing this is by focussing on features that have been *adaptively*

advantageous. In the case of episodic memory, one feature that is unique, and a possible candidate, is autonoetic consciousness—the first-person, experiential quality typically associated with episodic memory. We then looked at three possible explanations for why episodic memory might be adaptively advantageous. Klein, Cosmides, and Tooby (2002) suggest episodic memory might have evolved for decision making, in particular for providing boundary conditions on semantic generalizations in order to facilitate social decisions (section 4.2.1). Boyer (2008) suggest that the autonoetic quality of episodic memory is relevant for *affective recall* (section 4.2.2). In this case episodic memory allows us to vividly recall past emotions which, in turn, contributes towards pro-social decision making. Mahr and Csibra (2017) provide a third suggestion (section 4.2.3). They argue that episodic memory is e*pistemically generative*. It allows us to create and maintain our beliefs about past experiences, a function that is particularly useful for moderating social commitments.

In section three (4.3), we reviewed these proposals in detail. The three proposals all approach the function of episodic memory from an etiological perspective (section 4.3.1)—they all look to account for why episodic memory has been adaptively advantageous. They are also all consistent with episodic memory being a broadly preservative capacity. Despite proposing different reasons for episodic memory's adaptivity, all three accounts share some similarities in the areas where they think episodic memory might be advantageous: particularly regarding the *social domain* and *decision-making processes* (section 4.3.2). One complexity that arises from these three accounts, however, is that they propose different structures for episodic memory (section 4.3.3). While this creates a problem for assessing the etiological function of episodic memory, it does allow for these accounts to be compatible. Here, I've made the case that these functions more accurately describe capacities in which episodic memory functions, as opposed to a proper etiological function (section 4.3.4). Etiological functions can be difficult to determine, particularly with regards to complex cognitive systems like episodic memory. For the purposes of our discussion, however, what we really want to know about is how we currently employ our episodic memory system, and for this I think a Cummins-style structural analysis is better suited.

In section four (4.4), I make the case for why the function of episodic memory qualifies as being *broadly preservative*. One important point to consider regarding episodic memory is that it is a

cognitive capacity (section 4.4.1). Even though cognitive capacities are ultimately also biological systems, they are unique in that they process information. Accordingly, when we talk about the function of episodic memory as a cognitive system, what we really want to know about is how it currently fulfills that role: what type of information does it process? Given this, I think for our current discussion, the causal role method is more appropriate (section 4.4.2). Despite this the only causal role account of the function of episodic memory we've seen was given by De Brigard in chapter three (section 3.1.1.1) and concluded that episodic memory is not a broadly preservative cognitive capacity. If, however, we incorporate the necessity of a memory trace into a causal role account of the function of episodic memory, this changes that. Here (section 4.4.3), I argue that a causal role account of the function of episodic memory is consistent with episodic memory being a broadly preservative cognitive capacity. Thus, regardless of which method of functional analysis we use, it should be clear the function of episodic memory is broadly preservative. I conclude this chapter, and my dissertation, on this point: that episodic memory is a broadly preservative capacity.

Bibliography

Allen, T. A., & Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences of the United States of America, 110 Suppl 2*(Suppl 2), 10379–10386.

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in cognitive sciences, 18*(6), 279–292.

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(4), E420–E429

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408.

Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 113*(4), 518–540.

Bernecker, S. (2010). Memory: A philosophical study. Oxford: Oxford University Press

Boyer, P., & Barrett, H. C. (2005). Domain Specificity and Intuitive Ontology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 96–118). John Wiley & Sons, Inc.

Bratman, Michael. (2000). Reflection, Planning, and Temporally Extended Agency. Philosophical Review. 109. 35-61.

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

Butler, J. (1736). *The analogy of religion: Natural and revealed, to the Constitution and course of nature. to which are added two brief dissertations: I. Of Personal Identity. II. of the nature of virtue. by Joseph Butler* . Printed for John and Paul Knapton.

Ciaramelli, E., Ghetti, S., Frattarelli, M., & Làdavas, E. (2006). When true memory availability promotes false memory: evidence from confabulating patients. *Neuropsychologia*, *44*(10), 1866–1877.

Cosmides, L., & Tooby, J. (2000). Consider the source: The evolution of adaptations fordecoupling and metarepresentation. *Metarepresentations*, 53–116.

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159–180

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, *72*(20), 741.

De Brigard, F. (2014a), The Nature of Memory Traces. Philosophy Compass, 9: 402-414.

De Brigard, F. (2014b). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155-185.

De Brigard, Felipe. (2020). The Explanatory Indispensability of Memory Traces in advance. The Harvard Review of Philosophy. 27.

Dewhurst, S. A., Thorley, C., Hammond, E. R., & Ormerod, T. C. (2011). Convergent, but not divergent, thinking predicts susceptibility to associative memory illusions. *Personality and Individual Differences*, *51*(1), 73–76

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency*. Oxford University Press.

Fessler, D. M. T. (2001). Emotions and cost–benefit assessment: The role of shame and self-esteem in risk taking. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 191–214). The MIT Press.

Gerrans, P., & Kennett, J. (2010). Neurosentimentalism and Moral Agency. *Mind, 119*(475), 585–614.

Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. *Psychology of Learning and Motivation*, 59–109.

Griffiths, P. E. (1993). Functional Analysis and Proper Functions. *The British Journal for the Philosophy of Science, 44*(3), 409–422.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Hanley, J. R., & Kay, J. (1998). Proper name anomia for the names of people: Functionaly dissociable impairments? *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 34*(1), 155–158.

Howe, M. L., Garner, S. R., Dewhurst, S. A., & Ball, L. J. (2010). Can false memories prime problem solutions? *Cognition*, *117*(2), 176–181.

Hume, D. (1978). Treatise of human nature (L. A. Selby-Bigge, Ed.; 2nd ed.). Oxford UniversityPress.

Irish, M., Addis, D. R., Hodges, J. R., & Piguet, O. (2012). Exploring the content and quality of episodic future simulations in semantic dementia. *Neuropsychologia, 50*(14), 3488–3495. Kapur N. (1999). Syndromes of retrograde amnesia: a conceptual and empirical synthesis. *Psychological bulletin*, *125*(6), 800–825.

Kelly, L. (2016). *The memory code: The traditional Aboriginal memory technique that unlocks the secrets of Stonehenge, Easter Island and ancient monuments the world over*. Allen & Unwin.

Khalidi, M. A. (2023). *Cognitive ontology: Taxonomic practices in the mind-brain sciences*. Cambridge University Press.

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review, 109*(2), 306–329.

Koutstaal, W., Schacter, D.L., Johnson, M.K. *et al.* Facilitation and impairment of event memory produced by photograph review. *Memory & Cognition* 27, 478–493 (1999).

Lewis, D. (1987). Veridical hallucination and Prosthetic Vision. *Philosophical Papers Volume II*, 273–290.

Levine, L. J., & Bluck, S. (2004). How Emotions Fade: Valence, Appraisals, and the Emotional Impact of Remembered Events. In S. P. Shohov (Ed.), *Advances in Psychology Research*, Vol. 30, pp. 3–20). Nova Science Publishers.

Locke, J. 1632-1704. Two Treatises on Civil Government. London :G. Routledge and sons, 1887.

Loftus, E. (1997). Creating False Memories. Scientific American. 277. 70-5.

Loftus, E. (2003). Our changeable memories: Legal and practical implications. Nature Reviews Neuroscience, 4(3), 231-234.

Madden, A. D., Bryson, J., & Palimi, J. (2006). Information behavior in pre-literate societies. *New Directions in Human Information Behavior*, 33–53.

Mahr, J., & Csibra, G. (2017). Why do we remember? The communicative function of episodic memory. *The Behavioral and brain sciences*, *41*, 1–93. Advance online publication.

Malcom, N. (1977). Memory and Mind. Philosophy 53 (204): 270-272.

Martin, C. B., & Deutscher, M. (1966). Remembering. Philosophical Review, 75, 161–196.

McMahan, J. (2002). The Ethics of Killing: Problems at the Margins of Life. New York, US: OUP Usa

Melo, B., Winocur, G., & Moscovitch, M. (1999). False recall and false recognition: An examination of the effects of selective and combined lesions to the medial temporal lobe/diencephalon and frontal lobe structures. *Cognitive Neuropsychology, 16*(3-5), 343–359.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*, *24*(3), 323-342.

Michaelian, K. (2016a). Confabulation, Misremembering, Relearning: The Simulation Theory of Memory and Unsuccessful Remembering. *Frontiers in psychology*, *7*, 1857.

Michaelian, K. (2016b). *Mental Time Travel: Episodic memory and our knowledge of the personal past*. The MIT Press.

Millikan, R. G. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Nathan, M. J., & Del Pinal, G. (2016). Mapping the mind: bridge laws and the psycho-neural interface. *Synthese*, *193*(2), 637-657.

Neander, K. 1983. *Abnormal Psychobiology*. Dissertation, LA Trobe University, Bundoora Australia.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Parfit, D. (1984). Reasons and Persons. Oxford, GB: Oxford University Press.

Pezdek, K., Blandon-Gitlin, I., & Gabbay, P. (2006). Imagination and memory: Does imagining implausible events lead to false autobiographical memories? *Psychonomic Bulletin & Review, 13*(5), 764–769.

Reid, T. &Hamilton, W., Walker., ed. (1850). *Essays on the Intellectual Powers of Man.* Cambridge, J. Bartlett

Robins, S. (2016a). Representing the past: Memory traces and the causal theory of memory. Philosophical Studies, 173(11), 2993-3013.

Robins, S. K. (2016b). Misremembering. *Philosophical Psychology, 29*(3), 432-447.

Robins, S. (2017). Memory traces. In Bernecker, S., & Michaelian, K. (Eds.). The Routledge Handbook of Philosophy of Memory. Routledge.

Rowlands, M. (2009). Memory. In *Routledge Companion to Philosophy of Psuchology,* ed. P. Calvo and J, Symons, 336-345. London: Routledge.

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), 803-814.

Russell, B. (1921). *The analysis of mind.* Allen & Unwin.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Houghton Mifflin.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of Memory and Language, 35*(2), 319–334.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine

the future: the prospective brain. Nature Reviews Neuroscience, 8(9), 657-661.Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press.

Shoemaker, S. (1970). Persons and Their Pasts. *American Philosophical Quarterly, 7*(4), 269–285.

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.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans?. *The Behavioral and brain sciences*, *30*(3), 299–351.

Szpunar K. K. (2010). Episodic Future Thought: An Emerging Concept. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *5*(2), 142–162.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, W. Donaldson, *Organization of memory.* Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. Clarendon Press.

Tulving, E., Schacter, D. L., McLachlan, D. R., & Moscovitch, M. (1988). Priming of semantic autobiographical knowledge: A case study of retrograde amnesia. *Brain and Cognition, 8*(1), 3–20.

Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review, 2*(3), 387–390.

Ursu, S., & Carter, C. S. (2005). Outcome representations, counterfactual comparisons and the human orbitofrontal cortex: Implications for neuroimaging studies of decision-making. *Cognitive Brain Research, 23*(1), 51–60.

Wade, K.A., Garry, M., Don Read, J. *et al.* A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review* **9**, 597–603 (2002).

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11125–11129.

Wheeler, M. A., & McMillan, C. T. (2001). Focal retrograde amnesia and the episodic-semantic distinction. *Cognitive, affective & behavioral neuroscience*, *1*(1), 22–36.