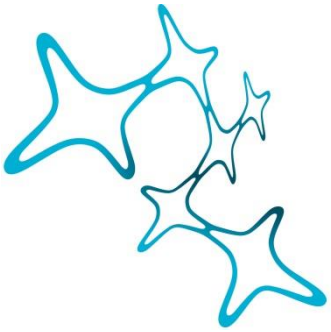
NEUROPHILOSOPHY AND ETHICS OF FALSE MEMORIES AND FALSE BELIEFS

Urim Retkoceri



Graduate School of
Systemic Neurosciences

LMU Munich

Dissertation
at the Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

July, 2020

Supervisor
Prof. Dr. Stephan Sellmaier
Research Center for Neurophilosophy and Ethics of Neurosciences
Ludwig-Maximilians-Universität München

First Reviewer: Prof. Dr. Stephan Sellmaier
Second Reviewer: Prof. Dr. Sven Bernecker
External Reviewer: Prof. Dr. Sarah Robins

Date of submission: July 22, 2020
Date of defense: November 27, 2020

Acknowledgements

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

Too many have helped me get here to name them all. Otherwise this thesis would have twice the number of pages. I wish to thank every single one of them and keep them in my memory.

An erster Stelle danke ich Stephan Sellmaier, der mich seit meiner Masterarbeit unterstützt und in der Welt der Philosophie begleitet hat. Als Biologiestudent, der von Philosophie fasziniert war, schien eine Dissertation, die hauptsächlich philosophischer Natur ist, in weiter Ferne. Es war eine Überraschung zu sehen, wie offen er für ein so interdisziplinäres Projekt war, und wie leicht und locker unsere Diskussionen verliefen. Sein Ansatz wie man Philosophie betreibt, auf der einen Seite äußerst genau und abstrakt, aber gleichzeitig auf der anderen Seite nie abgehoben und immer nah an der gelebten Wirklichkeit, wird meine eigene Art zu philosophieren hoffentlich dauerhaft prägen.

Mein Dank gebührt auch Christian Leibold, der ebenfalls seit meiner Masterarbeit ohne zu zögern bereit war, mir als damaligen Naturwissenschaftler, der in seinem Denken immer philosophischer wurde, stets weiterhin Sichtweisen aus der Naturwissenschaft nahezulegen. Seine Herangehensweise an philosophische Themen war immer äußerst originell und hilfreich.
Gleichermaßen danke ich Sven Bernecker, der sich enthusiastisch dazu bereit erklärt hat meine Promotion zu betreuen und diese Arbeit zu bewerten. Es gibt wohl keine geeignetere Person auf dem Gebiet der Philosophie der Erinnerungen, die mich dabei betreuen konnte. Es war äußerst bemerkenswert im Laufe meiner eigenen Entwicklung in der Philosophie der Erinnerungen, seinen philosophischen Werdegang auf dem gleichen Gebiet, von seinen ersten Papern, über die buchlangen Auseinandersetzungen bis hin zu seiner Ansicht über genau das Thema dieser Arbeit zu sehen. Dadurch konnte ich nicht nur etwas über die Philosophie der Erinnerungen lernen, sondern auch darüber, wie man Philosophie selbst weiterentwickelt.

Special thanks go to Sarah Robins, who agreed to review this thesis. I was excited when she agreed to review it, since I had admired her approach to the topic of the philosophy of memory at various conferences over the years, and rarely found someone who was as open and enthusiastic about looking at a topic from a truly interdisciplinary perspective.

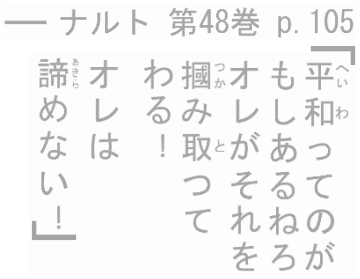
I would like to thank the Graduate School of Systemic Neurosciences for financial, social and intellectual support, as well as all members of the Research Center for Neurophilosophy and Ethics of Neurosciences at the LMU Munich.

Natürlich vergesse ich nicht meine Freunde Rimas und Vasilij, die mich seit Jahren unterstützt haben und mit denen ich immer intensive und ehrliche Diskussionen zu jedem Thema führen konnte.

Ohne Lisa hätte ich diese Arbeit niemals beginnen können. Danke für alles was wir erleben und lernen konnten. Manche Erinnerungen sind zu kostbar als dass man sie jemals vergessen sollte.

Thank you Natasha for sharing so much of your life with me and for reminding me through your curiosity and inquisitiveness of our thirst and love for life. I hope that some day we will get to know more about mixing colors to get violet, because there is so much left to discuss, explore and get to know better in the future

Në fund, falënderoj nënën dhe babën që më ndihmuan, dhe Skendin, si të vetmin, i cili ishte aq i interesuar për gjëra jashtë zakonshme. Sidomos nëna më ka ndihmuar aq sa ka mujt, dhe shpesh më shumë se ka mujt! Natyrisht falënderimet i takojnë Zotit.



to Natasha

Abstract

Memory is incredibly important for human life. Because of this importance, memory has been studied for as long as philosophy and science have existed. But, there are many questions left to be answered and many more left to be discovered.

In this thesis I mainly focus on the phenomenon of false memory and its implications, and aim to give an interdisciplinary analysis of it. To understand what false memory is, or what it could be, I first take a look at memory in general. For this, I outline different ideas of what kinds of memories there are and introduce three contemporary theories of what it means to remember: the causal theory of memory, the simulation theory of memory and the hybrid theory of memory. After briefly pointing out some possible differences between false memory and confabulation, I describe how those three theories of memory characterize false memory. Having set the stage, three chapters follow which deal with different topics that developed out of what is still missing in the contemporary study of memory and false memory.

First, I develop a conceptual account of what it means to falsely remember how to do something. I characterize genuine remembering how as the performance of an act for which a specific ability has been acquired which is necessary to perform that act. False remembering how on the other hand is, crudely put, characterized as the performance of an act that was not tried to be performed but for which the relevant ability has still been acquired.

Second, I take a look at what it might mean to remember emotions. For this I look at two components often attributed to emotions, certain physiological or behavioral responses (which I call implicit emotions) and certain conscious experiences (which I call explicit emotions). I describe what it might mean to remember each of these parts and bring them together in a framework that also includes other aspects that are often attributed to emotions. A brief suggestion of what it might mean to falsely remember emotions follows.

Third, I tread into the intersection between memory, false memory and moral responsibility. Here I answer the question if you are morally responsibility for the veracity of your memories, that is, if you are morally responsible to ensure that your memories stand in the alleged relation to their purported content. I argue that this can be the case under certain conditions, but only if the moral responsibility is derived from something other than ensuring the veracity of the memory in question.

Finally, I discuss how the ideas presented in this thesis could be developed further to understand memory and false memory better. There are still many intriguing questions left to be answered in the interdisciplinary study of false memories, and these questions will guide where we go from here.

Contents

Introduction

Memory is fundamental to life as a human being. It is impossible to even just imagine a fulfilling life without at least some kind of memory. Experiences we treasure in our hearts would be lost and loved ones would be strangers. It is precisely because memory is so important that it deserves to be studied.

Given this importance and ubiquity of memory, it is no surprise that it has piqued the interest of philosophers, scientist and many others for millennia. While I wish I could give an overview of the history of the study of memory[1], such an endeavor here would either be too short to do its significance justice or too long to be fit into a single manuscript. The same goes for the contemporary study of memory with its ever-increasing diversity.

Thus, unfortunately, I will have to confine myself to a selection of specific bits and pieces the study of memory has to offer and try to both, give an overview of the current state of memory and false memory research, and try to focus this writing on its main topic: the neurophilosophy and ethics of false memories and false beliefs.

Outline – What to expect

This thesis is mainly concerned with the phenomenon of false memory and its implications, studied from a philosophical perspective.

To make sense of why certain aspects of false memories are investigated in this thesis, I will first outline different ideas of what kinds of memories there are and introduce three contemporary theories of what it means to remember. After a few remarks about the differences between false memories and confabulations, I will then describe how three contemporary memory theories define false memories, and mention a few things that are still missing in general in the study of memory and false memory.

Three chapters follow which will deal with different topics that developed out of what is still missing in the contemporary philosophical theories of memory and false memory. First, a conceptual account of what procedural and false procedural memory is will be developed. This account will be compared to two current philosophical accounts of false declarative memory. Second, the answer of what it might mean to remember an emotion will be investigated and answered partly. Since the concept of emotion is complex and vastly different ideas of what emotions are exist, only two major components which are often intuitively ascribed to emotions will be analyzed in relation to memory and false memory. Lastly, the still quite unexplored issue of the interplay between false memory and moral responsibility will be studied. Concretely, the question if you are morally responsible to make sure that what you seem to remember is the way you remember it will be answered.

Finally, a concise discussion of how the contributions made in this paper could be developed further in the future will round up the paper, and I will conclude by showing that there are still a lot of interesting questions left to be answered in the interdisciplinary study of false memories.

[1] For an overview of the history of memory research from a scientific perspective see (Tulving, 2007), and from a philosophical perspective see (Bernecker & Michaelian, 2017; Nikulin, 2015).

1.1) Memory

At first it might seem obvious what memory is. You remember things all the time and you have lots of different memories. Yet, when trying to distinguish memory from other phenomena such as perception and imagination, and trying to define it precisely, it seems to become more and more elusive. On top of that, it might not even be clear what should even count as memory and what shouldn't. Philosophers and scientists alike have diverging opinions about this question, and for every definition or taxonomy out there, someone will find it lacking in some way. Yet, there seems to be at least some kind of loose consensus which many contemporary researchers seem to at least implicitly agree on most of the time or take as the default starting point. Thus, I will start by outlining often mentioned kinds, types or forms of memory and how they are usually thought to relate to one another.

1.1.1) Kinds of Memory

There are different ways or grounds on which kinds of memories can be identified and distinguished. Recently, Werning and Cheng (2017) have proposed a systematization for taxonomies of memory.[2] According to this systematization, memory taxonomies classify memories in a scalar (e.g. by time span into ultra-short term, short term and long term memory), natural-kind (i.e. by sets of properties that reflect the natural world) or hierarchical manner. Here I will focus on a hierarchical way of classifying memories since it is the one that seems to be most prominent in the contemporary interdisciplinary and philosophical memory debate. One of the most influential contemporary memory taxonomies, and the one used here, was initially introduced by Squire and Zola-Morgan (1988), and has since then been adapted or refined over time (Squire, 2004).

Declarative and nondeclarative memory

The taxonomy according to Squire and Zola-Morgan (1988) and its later versions (Squire, 2004) first distinguish between those kinds of memories which can, at least in principle, be verbally expressed and consciously recalled, called *declarative memories*, and memories which generally cannot be expressed verbally but are expressed through performance, called *nondeclarative memories*. An example of recalling declarative memory would be remembering that a bike usually has two wheels, while an example of recalling nondeclarative memory would be remembering how to actually ride a bike. While you can easily write down that a bike has two wheels and someone who reads it would normally be able to learn and remember it, writing down how to actually ride a bike, meaning how you move your muscles and body, seems rather difficult (and as e.g. Curran (2001) points out, people who can ride bikes often describe what they do inaccurately), not to mention the question of if it is even possible to learn to ride a bike just by reading such a description.

Yet, intuitively it might still not seem necessary to distinguish declarative memory from nondeclarative memory. Both are forms of memory which can, in principle, be acquired through learning, can be improved with practice and can be forgotten. Remembering that a bike usually has two wheels seems different from remembering how to actually ride it. But, then again, remembering that a bike usually has two wheels is also different from

[2] I will put linguistic approaches of classifying kinds of memory aside (Bernecker, 2008; Hacker, 2013).

remembering that a bike usually has pedals. The key question is whether or not there is sufficient reason to distinguish declarative memory from nondeclarative memory.

In contrast to most previous (philosophical) classifications of memory, which mostly relied on intuition and everyday observations, the taxonomy by Squire and Zola-Morgan (1988) was the result of bringing together different strands of neuroscientific observations and experimentation, with a focus on physiological evidence. In scientific research, such a distinction gained momentum when it was discovered that humans who suffered from certain forms of amnesia showed no statistically relevant improvement of performance on tasks that required mostly declarative memory, but, at the same time, could improve their performance on tasks that mostly required nondeclarative forms of memory (Cohen & Squire, 1980; Milner, Corkin, & Teuber, 1968; Milner, Squire, & Kandel, 1998). Interestingly, these amnesiacs did not recall participating in the experiment or training, or that their respective performance had improved, even though their skills had in fact improved (comparable to non-amnesiacs) (Milner, Squire, & Kandel, 1998; Squire, Cohen, & Zouzounis, 1984). These and similar findings led scientists to conclude that declarative memories are distinct from nondeclarative memories (Cohen & Squire, 1980; Squire, Cohen, & Zouzounis, 1984; Squire & Zola-Morgan, 1988). Follow-up studies using different techniques, such as neuroimaging studies (Grafton, et al., 1992), strengthened this conclusion further, and introduced additional differences between declarative and nondeclarative forms of memories.

Declarative memory: Semantic and episodic memory

Another influential distinction was made earlier by Tulving (1972), but, as he notes himself, similar distinctions have been suggested by others before.[3] According to Tulving (1972) what is now referred to as declarative memory can be subdivided into memories of personally experienced events or episodes which (usually) have a temporal-spatial dimension[4], called *episodic memories*, and (initially) memories that are necessary for the use of language (such as symbols or concepts) called *semantic memories*. However, nowadays it is more common to describe semantic memory as the memory of propositions or facts (Tulving, 2016) or as general knowledge (Tulving, 1985). An example of retrieving episodic memory would be when I remember the experience when I was trying to learn how to ride a bike at the age of ten at the nearby parking lot, fell off and scraped my elbow. In contrast, remembering that bikes are generally used for transportation around the world would be an example of retrieving semantic memory.

Thus, in contrast to episodic memories, semantic memories can be independent of an episode you have personally experienced. As Tulving (1972) notes, naturally you have in some way learned that which is now a semantic memory, meaning there has been an episode of learning. However, semantic memory only refers to the general content of what

[3] Strictly speaking, Tulving (1972; 2005) makes a distinction between memory *systems* and not certain kinds of memory or memory contents which can arise from use of such systems. However, the details of the differences between memory systems, memory kinds, memory contents and other senses of how *memory* is used in these contexts will not play a substantial role here, and I will mostly look at different kinds of memories. For different senses of the term *memory* or *remembering* in these debates see e.g. (Roediger, Dudai, & Fitzpatrick, 2007; Tulving, 2000; Werning & Cheng, 2017).

[4] The aspect of a temporal-spatial dimension is also why episodic memory is often described as involving *mental time travel* (Tulving, 2005).

was learned, not to any experiential aspect of the episode itself. Consequently, concerning semantic memory, it is possible that you remember something that you have not experienced yourself, such as remembering that Socrates died by drinking hemlock even though you did not experience that event yourself.

The distinction between episodic and semantic memory did not seem to be clear-cut even from the beginning. Tulving later (1985) additionally distinguished episodic and semantic memories by the kind of consciousness involved. According to this idea, semantic memory involves a state of knowing, or *noetic* consciousness, while episodic memory involves a state of self-knowing, or *autonoetic* consciousness. Noetic consciousness is described as enabling a subject to be aware not only of presently occurring events and objects, but also absent ones and of symbolic representations. In the case of semantic memory this means for example being able to remember what generally speaking a bike tour looks like. In contrast, autonoetic consciousness is characterized by being aware of events as part of your own experience. In the case of episodic memory this means for example being able to remember *your* last bike tour through the mountains as personally experienced. Tulving (1985) based this distinction mainly on people who had semantic memories but because of a certain kind of amnesia were not able to recall any episodic memories and did not possess autonoetic consciousness.

While this distinction between semantic and episodic memory is not without its fair share of criticism (Werning & Cheng, 2017), it remains quite popular and is often, at least implicitly, assumed in interdisciplinary and philosophical memory debates.

Nondeclarative memory: an umbrella term for the rest

As the name suggests, nondeclarative memory is primarily defined in a negative sense as memory that cannot, even not in principal, be 'declared' or verbally expressed. However, Squire and Zola-Morgan (1988) and others (Squire, 2009) usually define it as the kind of memory which is expressed through performance without requiring any conscious experience. It has been hypothesized that nondeclarative forms of memory cannot be expressed verbally because of this lack of direct access to consciousness (Knowlton, Siegel, & Moody, 2017). But even such a definition is so broad that the term nondeclarative memory can at best be considered an umbrella term for vastly different phenomena. For example, Squire (2004) argued based on neurophysiological differences that nondeclarative memory encompasses procedural memory, priming and perceptual learning, conditioning or associative learning (acquired responses through conditioning), and nonassociative learning (i.e. changed response strength to a stimulus after exposure to it). Apart from procedural memory, the other kinds of nondeclarative memories will not be discussed here further.

Since procedural memory will be investigated in detail later, a short description here will suffice. Procedural memory usually refers to the what in everyday speech is called *skills*, such as riding a bike. In scientific research it is usually characterized as "the memory system in charge of encoding, storing, and retrieving the procedures that underlie motor, verbal, and cognitive skills" (Beaunieux, et al., 2006, p. 521). As the quote already indicates, often three different kinds of procedural memory are distinguished: perceptual, motor and cognitive (Knowlton, Siegel, & Moody, 2017). Actually riding a bike is an example of a perceptual-motor procedure, while quickly solving difficult situations in chess is an example

of a cognitive procedure. However, most complex actions are thought to require all three of them (Adams, 1987).

Kinds of Memory in Philosophy

The interdisciplinary study of memory mostly follows some form of the taxonomy of memory presented above. Traditionally however, philosophers used different taxonomies. Here I will sketch one classification of memory that has been influential in the philosophy of memory for some time, but since, as mentioned, the interdisciplinary debate mostly uses some form of the taxonomy described above, I will sketch the one traditionally used in philosophy only very briefly.

What scientists usually refer to with semantic memory, philosophers employed the concept of *propositional memory* (or *factual memory*), which (as the name suggests) is memory of propositions (p), usually expressed in the form of *x remembers that p*. Saying *I remember that a bike usually has two wheels* would be an obvious example of propositional memory (Debus, 2017; Malcolm, 1963).

Similar to, but still somewhat different than, episodic memory is the notion of *experiential memory* (also called *recollective*, *perceptual* or *personal memory*), which is memory of personal experiences (Debus, 2017; Malcolm, 1963). Intuitively there seems to be a meaningful difference between the sentences *I remember that I got married* (an example of propositional memory) and *I remember getting married* (an example of experiential memory). While the former seems to be just as true whether I remember that I got married because I experienced it or because I read it somewhere, saying that the latter would just as well be true whether I experienced it or just read it somewhere might seem at least somewhat counterintuitive. In contrast to propositional memory, experiential memory, as it is usually employed, seems to require some kind of phenomenological dimension or mental imagery which you get from having personally experienced something yourself (Debus, 2017; Malcolm, 1963). It does not seem unnatural to say that if I *experientially* remember my wedding, I can in some way 'see' my wife in her wedding dress or 'play back' her voice in my mind when we exchanged vows.

Lastly, much like the scientist's procedural memory, philosophers use *habit memory* (or *practical memory*) (Debus, 2017; Bernecker, 2011) to refer to the type of memory when remembering how to do something. Just like in the case of procedural memory, philosophers would say that remembering how to ride a bike would be a realization of habit memory. Much of the discussion about habit memory has taken place in the debate about *knowledge that* vs. *knowledge how* (Bengson & Moffett, 2011; Fantl, 2017), but details concerning this aspect will be described later.

Although the phenomena covered by the traditional philosophical classification of memory are not entirely overlapping with those covered by the usual scientific one described before, the three kinds of memories from either of them seem similar enough to the corresponding one in the other classification to often be used interchangeably.

1.1.2) Philosophical Theories of Memory

While the philosophical study of memory reaches as far back as philosophy in general (Nikulin, 2015; Bernecker & Michaelian, 2017), the contemporary debate is usually seen as having started with the paper *Remembering* by Martin and Deutscher (1966) around the

middle of the twentieth century. Broadly speaking, three main contenders are currently being developed in the philosophy of memory. The oldest strand can be described as a continuation of Martin and Deutscher's tradition, most extensively defended by Sven Bernecker (2010), which is usually labeled as *causal theory of memory*. In rather direct opposition to it are accounts that in some way negate the necessity of the causal part of the causal theories of memory and instead see memory and imagination as functions of the same system, often referred to as *simulation theories of memory*.[5] The most elaborate account of this type has been defended by Kourken Michaelian (2016a) for some time. A somewhat middle position is taken by Sarah Robins (2016) whose theory shares characteristics with both of the aforementioned accounts, and which she calls the *hybrid theory of memory* (but note that (Michaelian & Robins, 2018) use the same term differently).

This is not to say that there are not others who have contributed to the debate in the contemporary philosophy of memory with their own versions or theories (Cheng & Werning, 2016; Debus, 2017; Debus, 2010; De Brigard, 2014; Michaelian & Robins, 2018). However, since the philosophical accounts of false memories presented later arose from these three accounts, I will sketch only them in the following. Additionally, since this work's focus is on *false* memory, I will only give an overview of these three theories of memories, and reserve some relevant details for the description of the corresponding false memory accounts.

The Causal Theory of Memory

Currently the most detailed formulation of a causal theory of memory is being defended in numerous works by Sven Bernecker. Among the three contenders presented here, it is the one that has been worked out most extensively, including book length investigations into various details (2008; 2010), and later defended or applied to different related problems in the philosophy of memory (2017a; 2017b).

In its basic form the causal theory of memory argues that it is a necessary condition that any memory causally depends on the past in a specific way. Naturally, a description like this is too vague in many respects to just capture memory or remembering, but it is the main point the causal theory of memory and the simulation theory of memory presented later disagree on. There are many important details to be filled in, most importantly what the causal dependence must be like, but here I will focus mostly on those aspects that differentiate the causal theory from the simulation theory or the hybrid theory, and on those aspects that will be important for this thesis.

Bernecker and others defend the idea of so-called *memory traces* to establish causal dependence on the past. Memory traces have been conceived in many different ways (Bernecker, 2001; Bernecker, 2010; Bernecker, 2017b; Debus, 2017; Martin & Deutscher, 1966; Robins, 2017), but in causal accounts generally seem to express the idea of 'that which carries the causal connection'. Causal dependence is seen as being realized by a causal connection to the past, and memory traces as establishing this connection. I will not get into details surrounding memory traces, since they are not that important for later chapters presented here, but, according to Bernecker, memory traces are necessary to explain memory and remembering, and other explanations of how a connection to the past or a causal dependence could be realized are less convincing (2001; 2010, pp. 104-127).

[5] Although it should be noted that some versions of the simulation theory do not negate causal dependence on past experiences (De Brigard, 2014).

Here I will describe one central condition that characterize the causal connection and I will illustrate this condition by using the example of an episodic memory of my getting married (and I, as an optimist, assume in this case that is a singular event). According to this version of the causal theory of memory, it has to be the case that if I now remember getting married, this remembering must depend on the past, namely on my (having represented) getting married in the past. In the case of genuine remembering this means that had I not gotten married, then now I would not remember getting married (counterfactual condition); since there would be no corresponding past, any alleged memory cannot causally depend on it. However, Bernecker argues that such a simple counterfactual condition does not suffice to warrant the appropriate kind of causal connection. Stated so simply, it is possible that the counterfactual condition is true, but that we would not say that there is a causal relation between the two (Bernecker, 2010, p. 125; Martin & Deutscher, 1966). What is important is that the counterfactual condition expresses a *causal* dependence. It should not only be true that 'had I not gotten married, I would now not remember getting married' but it should also be true that 'having gotten married is the cause of my currently remembering getting married'.

There are two further essential characteristics that are described in Bernecker's causal theory of memory that will be important for this thesis, namely that a memory must be true/factive and that it must be authentic for it to be a genuine memory (2010; 2017a; 2017b). The need for discussion of these conditions arises from the idea that instead of simply reproducing past representations (often called the *archival view*), remembering is the construction of representations of the past, which can often lead to alterations (often called the *constructive view*) (Bernecker, 2017a). Thus, the question follows if and in what way a current memory can differ from the object of that memory, i.e. what the memory is directed at. According to the truth condition, memory has to be true/factive, meaning it has to correspond to objective reality (Bernecker, 2017a; Bernecker, 2017b). Thus, I cannot genuinely remember getting married if I never got married, since then my alleged memory would not be true. Furthermore, memory must be authentic, meaning what you are remembering now has to be identical or sufficiently similar to what you had subjectively experienced at the event you are remembering.[6] If I am now trying to remember my wedding and I perceived my wife's wedding dress as violet on our wedding day, then for my memory to be *authentic* I must now remember it as violet as well and not as, for example, white. While the actual color of the dress (white) would be important for the truth of a memory, the color that I perceived at the time (violet) is important for the authenticity of a memory. According to Bernecker, for a memory to be a genuine memory it has to be both, true/factive and authentic (2017a; 2017b; 2010, pp. 36-39, pp. 213-239).

The Simulation Theory of Memory

The most worked out version of a simulation theory of memory has been formulated by Kourken Michaelian in various works, with the main focus on episodic memory (2016a; 2016b; 2020). The main contrasting point to the causal theory is that Michaelian explicitly denies any specific *necessary* causal dependence of a memory on the personal experience of

[6] Put concisely, what you are remembering is sufficiently similar if it is relevantly entailed by the original representation (remembering eating eggs instead of scrambled eggs), or if is inferred from it (remembering eating something non-vegan instead of scrambled eggs) (Bernecker, 2017a).

what is remembered (Michaelian, 2020).[7] Just like it is possible to imagine a future or counterfactual experience without there being a causal dependence on that future or counterfactual experience, this version of the simulation theory argues it is possible to remember a personal past experience without there being a specific causal dependence on that personal past experience.[8] This similarity between imagination and episodic memory and various other empirical findings are taken as the motivation to suggest that imagination and episodic memory are expressions of the same kind produced by the same cognitive system, termed the *episodic construction system* (Michaelian, 2016a; 2016b).

Yet, intuitively imagination and memory seem quite different. Hence, the simulation theory had to find a way to describe what the difference consists in. Michaelian has worked the details out in a number of different works (2016a; 2016b; 2020), but here I will mostly focus on those aspects that differentiate the simulation theory from the causal theory, and memory from false memory.

This version of the simulation theory draws on the idea that remembering is a constructive and not a reproductive process. Imagination and memory are seen as expressions of the same episodic construction system that recombines information from past experiences to construct, or simulate, representations of possible episodes or events (Michaelian, 2016a). However, Michaelian argues, episodic remembering does not necessitate that information from the personal past episode that is remembered is used for the simulation (even though it might frequently be the case). Instead, it is possible to remember something by combining various information of past sources and episodes, none of which has to be the personal episode that is being remembered. Applied to memories of my wedding this means that I can remember my wedding without tapping into any information from the experience I have made during my wedding. Instead I could draw on sources from other weddings, other experiences with the people that were at my wedding, generalized schemas of wedding dresses and so on. Thus, since any specific connection to the to be remembered personal past episode is not necessary, a causal connection to it is not necessary (Michaelian, 2016a). Likewise, things like memory traces which are used to explain causal connections are not necessary as well. However, it should be noted that while this simulation theory states that a causal connection is not *necessary*, it does not deny that often it is the case that such a connection exists (Michaelian, 2016a).

So far episodic remembering sounds like simply imaging the personal past, which still seems quite counterintuitive. Michaelian, however, suggests that further essential characteristics are needed, which will also be important for distinguishing memory from false memory. I will only describe two of those further here, namely reliability and accuracy, since the other (internality) is not crucial for this thesis. The episodic construction system has to be "properly functioning" (Michaelian, 2016a, p. 105) which in a nutshell means that it has to be able to reliably differentiate between simulating past from future episodes, and

[7] For now, I will simply confine myself here to episodes that are, or at least seemingly are, from your own personal past, and not from something you have not personally experienced (Michaelian, 2016a).

[8] This, and other, similarities between imagination and memory have led some to suggest that the two are on a single continuum of the functioning of the same underlying cognitive faculty and are only different in degree. Hence those who hold such views are sometimes called *continuists*, in contrast to those who are sometimes called *discontinuists* and think that memory and imagination are distinctly different in kind, cf. (Michaelian, Perrin, & Sant'Anna, 2020).

actual from counterfactual episodes. Importantly, the episodic memory has to be the result of reliable processes employed by the episodic construction system. A process is reliable, Michaelian argues, if employing it tends to mostly produces accurate representations of the past (2016a; 2016b). This is in contrast to an unreliable or malfunctioning episodic construction system or processes, which will be discussed later in the false memory section of the introduction. A simulation or representations of a personal past episode that is the result of unreliable processes employed by an episodic construction system, even if accurate, would then not be an instant of genuine remembering, since any accuracy would be coincidental. To count as memories, my representations of my wedding have to be the result of reliable processes of my properly functioning episodic construction system, which means that I have to generally speaking be able to construct accurate representations of past events in that way and must not most of the time make up inaccurate things of the past when allegedly remembering by employing those processes. According to Michaelian (2016a; 2016b; 2020), for an episodic memory to be a genuine memory it has to be an accurate representations of a personal past event that is the product of a reliable process employed by a properly functioning episodic construction system.

The Hybrid Theory of Memory

Initially, Robins (2016) develops her theory of memory more so in passing while testing how well two different accounts of memory fare in explaining certain memory errors. As mentioned before, recently the idea has gained popularity that, instead of simply storing and retrieving particular past representations (usually called the *archival view*, which aligns well with the main aspects of the causal theory of memory), in memory and remembering representations about past experiences are reconstructed from different sources (usually called the *constructive view*, which aligns well with the main aspects of the simulation theory of memory). Her hybrid theory of memory can be seen as the construction of a new idea that preserves aspects from both of these accounts.

Since Robins' theory mainly aimed at describing and classifying different memory errors and not necessarily giving a fully-fledged theory of memory, and since I will deal with false memories and memory errors later, a brief sketch how her theory characterizes genuine memories will suffice here, while details of her theory will be described later.

According to Robins' initial account (2016), genuine remembering is characterized by two main features, retention and accuracy. Retention is taken from the archival theory of memory and describes that information from the experience of a particular past episodes is stored. For example, if I remember my wedding some information from my experience of my wedding is retained. This means that when I remember my wedding, information from that particular wedding experience (and not just any default or random wedding) has been retained and is used for remembering. Accuracy, on the other hand, is taken to have mainly to do with retrieval and relates more to the constructive theory of memory. When remembering, a representation of the past is constructed, which, in the case of genuine remembering, has to be accurate. Thus, if I remember my wedding, I should remember my wife's wedding dress as white, since it was white. According to Robins (2016), for a memory to be a genuine memory information during the particular past experience has to be retained and later, during retrieval using that information, an accurate representation of the past experience has to be constructed.

In an updated version, Robins (2020) refines her account of remembering. First, a so-called target condition is added, which, crudely put, in this context just expresses the idea of 'that which is intended or taken to be represented actually existed'. In contrast to target is the content which expresses the idea of 'that which is actually represented'. In the example of my wedding this simply means that if I remember my wedding, my wedding must have actually taken place. Second, accuracy mostly stays the same but Robins (2020) notes that it could be differentiated similar to the way Bernecker (2017a; 2017b) does by distinguishing between truth and authenticity. Third, the simple retention condition is replaced by a causal history condition, which expresses the notion that the current representation and the remembered event or experience are causally connected in an appropriate way using a memory trace. For example, my current representation of my wedding must be produced by the memory trace formed by my experiencing the wedding at the time, if it is to count as genuine remembering.

1.1.3) What is still missing?

Thinking about the kinds of memories mentioned in the beginning, but also just your everyday experiences, you might think that some important things are missing in the contemporary philosophy of memory debate. This is unsurprising since memory interacts with many other different mental capacities and is crucial in a variety of different ways. Here I will focus on two things the contemporary philosophy of memory debate is largely missing, but which will be investigated in this thesis to fill the void at least a bit. Later, when describing what is still missing in false memory research, I describe one more topic.

Philosophical theories of nondeclarative memory

The three theories of memory described above mostly deal with declarative memory. Michaelian explicitly focuses on mostly episodic memories (2016a; 2016b), while Robins (2016) theory might more broadly aim at declarative memory. Bernecker (2017a) does mention procedural memory, but the theory he develops cannot simply be applied to it, since, for example, it does not seem clear what authenticity or accuracy would be in the case of procedural memory. This is not to say that these theories could not be applied to nondeclarative forms of memory, and, as I show later, I think they can, but in their current form they are best described as theories of declarative memory at the most.

However, it would be wrong to say that philosophers have not thought about nondeclarative memories. As was mentioned before, philosophers do acknowledge what they often refer to as habit or practical memory, but much of the discussion about those memories has taken place in the debate about *knowledge that* vs. *knowledge how* (Bengson & Moffett, 2011; Fantl, 2017). As the name suggests these debates seem to focus mostly on knowledge, not memory. But, as I will show later, philosophers have largely failed to distinguish between knowing how and remembering how in these debates, which means that there has been active philosophical research in for example procedural memory, but it has been confounded with other notions such as knowledge. One account that comes close to a theory of procedural memory is the one outlined by Katherine Hawley (2003), which I will discuss in more detail later.

Philosophical theories of memory of emotions

As I have mentioned above, the current philosophical theories of memory mainly concern themselves with declarative memories. If that is the case, you might wonder if emotions would be covered by those theories. Philosophers and scientists alike might give quite different answers to this question, which largely stems from the fact that they often use vastly different notions of what emotions are (Bard, 1928; Deonna & Teroni, 2012; Dixon, 2012; Goldie, 2009; Izard, 2010; James, 1983a; Lazarus, 1991; Mulligan & Scherer, 2012; Nussbaum, 2001; Schachter & Singer, 1962). In fact, as I elaborate later, the question of what it would even mean to remember an emotion is contentious and unclear.

Contemporary research into various aspects of the relation between emotions and memory is blooming, both from a scientific (Arun, Kandel, & Rayman, 2019; Nader, Schafe, & Le Doux, 2000; Ramirez, et al., 2013) as well as philosophical side (Arcangeli & Dokic, 2018; Debus, 2007; Gerrans, 2018). However, apart from a few exceptions (Arcangeli & Dokic, 2018; Debus, 2007; LeDoux, 1992), it is usually implicitly assumed whether or not emotions themselves can be remembered and what this means, and there has yet to be an extensive account covering this topic. Yet, since emotions, memories and their interplay arguably play an important role in health and disease, it seems that such an account is long overdue. That is why later in this thesis I will make an attempt to spark a debate by proposing one way in which the question of whether emotions can be remembered can be answered.

1.2) False Memory

Intuitively a false memory is usually understood as an alleged memory of something that either did not happen at all or that did not happen in the way you seem to remember it (Dalla Barba, 2002, p. 28; Pezdek & Lam, 2007). This can range from rather small and mundane alterations where you mistakenly take yourself to remember where exactly you left your keys, up to people being convinced to remember in detail entire events that, however, were actually implanted in them by suggestive questioning and which have never occurred to them, such as getting lost at a shopping mall as a child (Loftus & Pickrell, 1995; Loftus, 2005).

False memories gained public attention towards the end of the twentieth century when it turned out that alleged memories of sexual abuse were in fact (unintentional) fabrications, possibly induced by suggestive styles of questioning or therapy which sought to ‘uncover’ those alleged memories (Brainerd & Reyna, 2005). Naturally, the idea that people might remember things incorrectly has been known from everyday experience before, but only in the last decades has a systematic study of false memories been flourishing (Brainerd & Reyna, 2005; Loftus, 2005). These studies have uncovered that false memories are quite frequent and in fact so frequent that they cannot be described as a pathological manifestation but have to be seen as something that, to a degree, occurs in normally functioning memory systems (Schnider, 2018, pp. 174-184). Some go even so far as to claim that, because remembering is a reconstructive process, every memory is false to a degree (Bernstein & Loftus, 2009). While it is not obviously true that such a claim follows from the reconstructive nature of memory, it is undeniable that false memories exist. Yet, as I will show in the following, the idea that a false memory simply is the memory of an event that did not occur or did not occur in the way you seemingly remember it has faced some serious objections from current philosophical theories on memory errors and false memories.

1.2.1) *False Memory* or *Confabulation*?

If you take a look at the current philosophical theories outlined below which deal with what I have so far called *false memories*, you will see that the term usually used is *confabulation*.[9] While the particular details of the use of the terms *false memory* and *confabulation* are not crucial for the main points of this thesis, I nonetheless feel obliged to say at least a few words, especially since the title of this thesis includes the term *false memories*. However, similar to how there is no generally accepted definition or classification of memory, there is no generally accepted definition of what a false memory or confabulation is (Robins, 2019).

Perhaps the most obvious difference between false memory and confabulation is their use in everyday language. People in everyday situations usually understand what is meant when you say *false memory* or *remembered falsely*, while *confabulation* or *confabulated* is less common when referring to everyday phenomena. This difference in everyday language use is also indicative of how the concepts themselves are often understood differently. One important difference is that confabulation is frequently used to describe alleged memories that do not reflect that which they claim and are the result of a defective cognitive process or system usually found in pathological cases. False memory, on the other hand, is often simply taken as a more general concept that also occurs in everyday situations of healthy individuals (Schnider, 2018, pp. 174-184). In fact, the phenomenon of confabulation was initially used to describe fabrications made in alcoholics suffering from Korsakoff syndrome, even though usually other terms (such as *illusion of memory*) have been used as well (Schnider, 2018; Robins, 2019). Naturally, the debate is more complex with disagreement about the use of the terms and some using *confabulation* to explicitly refer to non-pathological cases as well (Bortolotti & Cox, 2009), but for the purposes of this thesis the differentiation mentioned will suffice.

Thus, *confabulation* is *usually* associated with a pathological context and is *usually* used to describe phenomena resulting from memory disorders, while *false memory* is a more generic term simply meaning to express that what is seemingly remembered did not actually happen (in the way seemingly remembered). Since I am *not only* interested in pathological cases but also in cases where there seems to be something wrong with a memory of healthy individuals in everyday situations (particularly when it comes to the moral responsibility), I will use the term *false memory* throughout the text, unless talking specifically about pathological cases.[10] However, when outlining the three philosophical theories of false memories in the next subchapter, I will primarily use their respective terminology, especially since it seems that they are starting to use the term *confabulation*

[9] *Confabulation* as such does not have to be limited to cases involving memory. However, if confabulations are due to failure of a memory system, they are often described as be *mnestic* or *mnemonic* (Bernecker, 2017b; Robins, 2020; Schnider, 2018). Since I am mostly concerned with memory in this thesis, I will confine myself to confabulations related to memory only, and I will use the term *confabulation* to mean *mnemonic confabulation*.

[10] Arguably, since some of the philosophical theories of false memory go a long way to make the point that not all confabulations are false, in the sense of being inaccurate, my using the term false memory throughout will lead to the counterintuitive result that some false memories are not false in the literal sense of the word. However, since such incidents are quite unlikely outside of pathological contexts, I think using *false memory* generally and *confabulation* when referring to pathological cases specifically seems like a fair compromise.

more liberally to refer to non-pathological cases as well and defining confabulation not on whether they are true or false (Michaelian, 2020; Robins, 2020).

1.2.2) Philosophical Theories of False Memory

The everyday notion of false memory as something that did not happen (in the way seemingly remembered) might be precise enough for everyday use, but it does not hold up to philosophical scrutiny. All of the three accounts of false memories presented here would not characterize false memories, or confabulations, in that way. Robins' (2016) initial notion comes closest to the everyday understanding of false memory, but differentiates memory errors in a much more extensive way. Bernecker (2017b) and Michaelian (2016b; 2020) on the other hand completely depart from the everyday notion of false memory and explicitly argue that false memories, or confabulations, can be completely true. I will outline the different accounts in the chronological order they appeared since the one that follows usually makes reference to the preceding ones.

The Hybrid Theory of False Memory

As mentioned before, Robins' theory (2016) combines aspect from the archival view of memory and the constructive view of memory. However, this is done primarily to explain different sorts of memory errors, and not primarily to give an account of genuine remembering. Initially, she differentiated memory errors from one another and from genuine memory by appealing to two different aspects, retention (of information from the experience of a particular past episode) and accuracy (of the representation of the past constructed at retrieval). The following four combinations (see Table 1) are possible by looking at only those two dimensions:

Table 1: Robins' (2016) initial classification of remembering and memory errors.

	Retention	Accuracy
Remembering	Yes	Yes
Misremembering	Yes	No
Confabulation	No	No
Relearning	No	Yes

In a more recent version of her account, Robins (2020) has refined her classification of memory errors. Hence the following updated classification (see Table 2) is given:

Table 2: Robins' (2020) updated version of a classification of remembering and memory errors.

	Causal History	Accuracy	Target
Remembering	Yes	Yes	Yes
Misremembering	Yes	No	Yes
Confabulation	No (absent)	Yes/No	Yes/No
Relearning	No (deviant)	Yes	Yes

Since I have discussed genuine remembering earlier, and since relearning is not relevant in this thesis, I will only describe misremembering and confabulation as described by the updated classification further.

Misremembering is the type of memory error that occurs when you seem to remember a targeted event, i.e. an event that actually existed, but the representation of that experience is nevertheless inaccurate, i.e. there is a mismatch between content and target, between what you intended to represent and what you represented. This can for example be the case if you represent details of an actual event inaccurately or you mix in details of another actual event. Misremembering is illustrated and backed up by findings made according to the so-called Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995; Robins, 2016). Crudely put, in those studies subjects are first presented with a list of words, and later presented with another list of words. The later list contains some words which were on the original list (list words) and some which were not on the original list (non-list words). Among the non-list words were words that are in some way related to the list words (such as *doctor* being related to *hospital* or *nurse*) and others which were in no obvious relation to the list words (such as *apple* in the same example). Interestingly, related non-list words are more often falsely recognized than non-related non-list word (Deese, 1959; Roediger & McDermott, 1995; Robins, 2016). According to Robins, to explain this pattern of recognition, some form of retention of information or existence of a target from the original list has to be assumed, otherwise it does not seem plausible why related non-list words would be falsely recognized more often than non-related non-list words. Thus, something from the original experience is retained or the original experience is targeted, but the representation that is being constructed is not accurate since those words were not in the original list. Robins terms this specific type of memory error *misremembering*. Applied to the example of my wedding, I could remember that my wife was wearing earrings shaped in the form of the number nine but instead of remembering them accurately as being violet, I remember them as being silver. I successfully remember the target which actually existed, meaning the wedding and the shape of my wife's earrings, but the color is represented inaccurately.

Confabulation, in contrast to misremembering, is defined by there being no relation between your seeming to remember an alleged event or experience and the alleged event or experience (i.e. no causal connection or history). This is the case either because the event or experience never occurred (target never existed) or any accurate representation is coincidental. This is also an important difference between Robins' initial classification and her updated version, since in the updated version it is possible that a confabulation is accurate, even though any accuracy would only be coincidental. These include the prototypical cases often thought of when thinking of false memory that were described in studies where people were through suggestive questioning made to believe having experienced events which actually never occurred to them (Loftus & Pickrell, 1995; Loftus, 2005). In the example of my wedding, this could for example be the case if I seem to remember how the flower girl tripped over my wife's wedding dress, when in reality we didn't have a flower girl at our wedding at all.

The classic idea of false memories as memories of events that did not occur (in the way remembered) would capture the notion of confabulation but would inevitably confound it with the notion of misremembering. Thus, the concept of false memory would be too vague to differentiate between what Robins (2016; 2020) describes as misremembering and confabulation.

The Simulation Theory of False Memory

Michaelian (2016b) picked up on Robins' (2016) idea of classifying memory errors, but reconstructs the approach in such a way as to be described by his simulation theory of memory. According to Michaelian (2016a; 2016b), neither retention from nor a causal connection to a particular past event is needed for genuine remembering. Hence, he suggests a classification of memory errors based on the following three criteria/conditions (like before, relearning is excluded here and internality simply assumed) (see Table 3):

Table 3: Michaelian's (2016b) classification of remembering and memory errors.

	Reliability	Accuracy	Internality
Remembering	Yes	Yes	Yes
Misremembering	Yes	No	Yes
Veridical Confabulation	No	Yes	Yes
Falsidical Confabulation	No	No	Yes
Veridical Relearning	Yes/No	Yes	No
Falsidical Relearning	Yes/No	No	No

Misremembering occurs when reliable processes, i.e. processes that tend to mostly produce accurate representations of the past, are employed by an episodic construction system to produce a representation of the past which happens to be inaccurate. Applied to the aforementioned cases in the DRM studies, a properly functioning episodic construction system employing a reliable process would be expected to misidentify related non-list words, because such words (e.g. *doctor*) are highly likely to have been on the list as well given that they are related to the relevant list words (e.g. *hospital*, *nurse*). Yet, since the related non-list words were not on the list, the simulation or representation created is inaccurate and therefore does not qualify as genuine remembering. Reliable processes were employed, but in this case, it happened that the produced representation was inaccurate.

Confabulation, in contrast, is defined by using unreliable processes, that is processes that tend to mostly produce *inaccurate* representations of the past, employed by an episodic construction system to produce representations of the past. Since it is still possible that unreliable processes produce accurate representations of a past event by coincidence, confabulations need not be inaccurate or false. Thus, if an episodic construction system employing unreliable processes produces an inaccurate simulation of the past, the confabulation is falsidical. If, by chance, the episodic construction system employing unreliable processes produces an accurate simulation of the past, the confabulation is veridical. What defines confabulation then is that an unreliable process was used to construct a representation of a past event. In the case of trying to remember my wedding this could mean that, if at some point in my old age my cognitive faculties start to fail me, I might simply make things up when trying to remember personal experiences long gone in the past. What I make up may be accurate by coincidence. When, at that age and in that condition, I am asked to remember my wedding, I might just piece together information from movies I have recently watched. Naturally, it is quite unlikely that I will by chance confabulate that my wife's earrings were violet, but, given that wedding dresses are usually white, it does not seem implausible that I take myself to remember that her dress was white. Yet, since my episodic construction system constructed the representations by employing

unreliable processes, even if I get the wedding dress color right, this would be a confabulation, since I got it right only by coincidence.[11]

The Causal Theory of False Memory

Bernecker's (2017b) theory of false memory chronologically followed those by Robins and Michaelian described above, but has not been updated yet. As mentioned before, the main point of the causal theory of memory was that a memory has to be, in an appropriate way, causally connected to a past representation of that which is seemingly remembered. This appropriate causal connection, or the lack thereof, is also the main distinguishing criterion between genuine memory and confabulation.

According to Bernecker (2017b), for a memory to be genuine it has to be both true/factual (correspond to objective reality) and authentic (be at least sufficiently similar to the initial perception of reality). However, it can be the case that even a confabulation is true and authentic, in which case it would be a veridical confabulation. Thus, while truth and authenticity are necessary conditions for genuine remembering, they are not sufficient to distinguish genuine memories from confabulations. What is needed, Bernecker (2017b) argues, is that what you seem to remember is not simply true and authentic by coincidence. Confabulations lack a proper causal history, meaning they do not causally depend, in an appropriate way, on the initial representation of that which is seemingly remembered. This causal connection is appropriate if it fulfills the counterfactual condition mentioned before, i.e. your current state of seeming to remember counterfactually depends on your initial representation of the relevant past experience.

Bernecker concisely describes the counterfactual condition in the following way "if the past representation had been different, it would have caused a different state of seeming to remember to match the different past representation" (Bernecker, 2017b, p. 9). Not meeting the counterfactual condition can, for example, be the case if a past representation has never even existed. Applied to my wedding, I would be confabulating memories of my wedding if I had never gotten married in the first place. This is a confabulation because the counterfactual condition is not fulfilled; since I never got married there is no past representation of my wedding that could be different. However, it could also be the case that a past representation existed but my state of seeming to remember just does not counterfactually depend on it. If, for example due to a childhood brain injury, I always seem to remember earrings any bride wears as violet, then even if my wife has worn violet earrings at our wedding, and even if I have perceived them as violet at the time, it could still be that I am not genuinely remembering the color of my wife's earrings at our wedding. What is important is that the counterfactual condition is met: had I perceived my wife's earrings in a different color than violet (e.g. as red), that representation of my wedding would have caused a different apparent memory now that matches the initial representation (i.e. I would now seem to remember them as red). Thus, even if I take myself

[11] In an updated version of his account of false memory, Michaelian (2020) adds meta-level cognition, i.e. judgments or cognitive processes about cognitive processes, to his classification. Simply put this often means that you make a judgment about whether what you seem to remember might be a case of genuine remembering or a memory error. While the issue of attitudes about your memory will play a role later on (but not as fine-grained as in Michaelian's account), in order not to complicate matters further, I will not elaborate on it here.

to remember the color of my wife's earrings at our wedding, and even if my current representation is true and authentic, this would simply be a coincidence unless the counterfactual condition is met.

In a nutshell, according to Bernecker (2017b), what defines confabulation is not that it is inaccurate, since veridical confabulations exist, but that the state of seeming to remember is not appropriately causally connected to the relevant past representation, i.e. that it does not counterfactually depend on the relevant past representation.

1.2.3) What is still missing?

As I have pointed out at the end of the last subchapter (1.1.3), there is a lot left to be discovered in the philosophical and interdisciplinary study of memory. Since the more systematic study of false memory or memory errors in general is, compared to the study of memory in general, relatively new, many questions remain unanswered. It can hardly be denied that different ways of how memory can go wrong exist and there seems to be a widespread consensus that memory errors in general are a common occurrence. Many interesting interactions and implications that follow from the new insights made by the emerging systematic investigation of false memory have yet to be uncovered and studied, and I will name two that will be examined in this thesis.

False Memory and Moral Responsibility

The ethics of memory so far have mainly been concerned with issues such as duties to remember (Blustein, 2017), memory enhancement or deletion (Kolber, 2006; Liao, 2017), forgetting (Bernecker, 2018) or consent (Craver & Rosenbaum, 2018). The issue of false memories has mostly been overlooked so far. Yet, since more and more evidence is suggesting that memory errors are common, the question arises what implications this has for moral responsibility when it comes to memory. Specifically, assuming you are, or should be, aware that your memories occasionally are false, are you then morally responsible to make sure that what you seem to remember was the way you seem to remember it? In the following I will propose that you can be morally responsible in that way, under certain conditions. This will hopefully serve as a starting point of a lively debate of the interaction between moral responsibility and false memory and other types of memory errors.

False Memory for Different Kinds of Memory

While both Robins (2016; 2020) and Michaelian (2016b; 2020) coming from a philosophical side, and, for example, Bortolotti and Cox (2009) or Schnider (2018) coming from a scientific side, propose classifications or taxonomies of memory errors, they mostly focus on memory errors pertaining to episodic memory, or do not investigate how the different types of memory described in the popular memory taxonomies could be false. This is hardly surprising since intuitively it does not seem clear what, for example, false nondeclarative memory would be. Yet, as you know from everyday experience there are many ways in which procedural memory, such as the execution of well-learned skills, can go wrong. Furthermore, it is not obvious if and how the theories of false memory described above would be applied to nondeclarative forms of memory.

Thus, additionally to suggesting a philosophical analysis of how procedural memory could be understood, I will in the following also distinguish different ways in which procedural memory can go wrong. This leads to an account of what false procedural memory

could be, which will be compared to possible applications of two of the false memory theories outlined above.

Taken together, in the following three chapters I will try to find some of the things that are still missing. I will begin by looking at the question of what false procedural memory might be. Afterwards, I will try and answer, at least in part, the question of what it might mean to remember an emotion. Lastly, I will suggest that and when you are morally responsible for the veracity of your memories. Finally, I will conclude the thesis by pointing out not only what of that which was missing was found, but also where we could go from here.

2) False Procedural Memory

(The whole content of this chapter, excluding this note, is part of an accepted manuscript of an article published by Taylor & Francis in the journal *Philosophical Psychology*.)

Lately it seems a number of philosophical memory theories are incorporating false memory phenomena into their conceptual frameworks. At the same time, scientific research is extending its analysis of false memories to nondeclarative forms of memory. However, both sides have paid little attention to the notion of false procedural memory. Yet, from everyday experience as well as from psychological investigation we are aware of different ways procedural memory goes wrong. Here, I characterize the conceptual foundation of false procedural memory. First, I distinguish remembering-how from knowing-how by proposing that remembering-how requires the performance of an act. Accordingly, genuine remembering-how is characterized as the performance of an act for which a respective ability has been acquired that is instrumental in the execution of said act. False remembering-how is identified as a kind of error where a subject acquires the ability to perform a certain act, which is then correctly executed, but is not what the subject tried to perform. This framework of false procedural memory is delineated from notions of interference and crosstalk. A comparison with current philosophical theories of false memory and analysis showing the relevance for current psychological research and everyday life concludes the paper.

Introduction

There is a more or less intuitive understanding of what a false memory is. Roughly put, a false memory is often understood to be a memory of something that never actually occurred (Pezdek & Lam, 2007).[12] This definition is usually applied to forms of memory which can, in principle, be verbally expressed and consciously recalled, typically referred to as *declarative memory* (Loftus & Pickrell, 1995; Squire, 2004). Even though it might also be common to talk of remembering how to do something (such as remembering how to ride a bike), which is often referred to as *procedural memory* or *habit memory* (Bergson, 1991; Cohen & Eichenbaum, 1993; Cohen & Squire, 1980), there does not seem to be an intuitive notion for *falsely* remembering how to do something.

While research into false declarative memory has been established in the last decades (Laney & Loftus, 2013; Loftus, 2005), research into if and how memories which are expressed through performance, typically referred to as *nondeclarative memory* (Squire, 2004) (including procedural memory), can be false is still in its infancy.[13] More and more literature is being published on false nondeclarative memory, such as in priming studies (McDermott, 1997; McKone & Murphy, 2000; Schacter, Gallo, & Kensinger, 2011). However,

[12] Recently, serious objections have been voiced against this definition, which (leaving aside specific differences between confabulations and false memories) state that (mnestic) confabulations can also be veridical (Bernecker, 2017b; Michaelian, 2016).

[13] In this paper I mostly adhere to the common taxonomy of memory proposed by Larry Squire (2004) as a guideline, but the exact taxonomic relations do not play a substantive role here. In this taxonomy, nondeclarative memory encompasses procedural memory, priming and perceptual learning, learned responses through conditioning (associative learning), and changed response strength to a stimulus after exposure to that stimulus (nonassociative learning).

the notion of false procedural memory seems to have gained little to no attention. This might not be surprising at first, since there is no intuitive, or otherwise established, notion of what a false procedural memory might be. At the same time, procedural memory arguably plays a crucial role in the life of humans and non-human animals, and has been researched extensively. Furthermore, in ordinary language we also acknowledge cases in which we feel procedural memory goes wrong in one way or the other, which usually are all simply subsumed under the terms *mistake* or *error*. Yet, as is elaborated later, psychological investigation does distinguish between different types of errors, and such research is being critically applied to different fields, such as designing training schedules in ways as to maximize training effects (Brydges, Carnahan, Backstein, & Dubrowski, 2007; Porter & Magill, 2010) or constructing machines so that potential for errors is reduced to avoid accidents (Sharit, 2012; Wickens, Hollands, Banbury, & Parasuraman, 2016). Thus, the question of what false procedural memory could be and how it can be defined will be investigated here.

Outline and Objectives

I will first motivate the investigation of false procedural memory by an everyday example of false remembering-how and distinguish it from other kinds of errors. A brief discussion of the differences to the closely related phenomena of interference and crosstalk will follow, in which I lay out that false remembering-how differs from each of them by being more specific and concrete, which makes it possible to distinguish false remembering-how from other errors concerning memory. After introducing the popular notions of procedural memory or habit memory employed by contemporary scientists and philosophers, I will point out differences between knowing-how and remembering-how. Having set the stage, I will propose a characterization of different cases surrounding remembering-how. It is here that I characterize false remembering-how as a specific kind of error in which (broadly speaking) a subject acquires the ability to perform a certain act which is then correctly executed but it is not what the subject tried to do. Another kind of error, unsuccessful remembering-how, is similar to false remembering-how, but differs from it in that what the subject does instead of what they tried to do is nothing for which the subject acquired a corresponding ability. After considering some objections that might arise, a look at how relevant such a notion of false procedural memory might be for the philosophical debate, psychological research and everyday life will conclude the paper.

2.1) What Kind of Phenomenon is being investigated here?

Back when corded telephones were still popular and numbers had to be dialed one digit at a time, there were certain telephone numbers that I dialed daily. While at first I had to explicitly recall each digit in order to dial the correct number, after a while this procedure became so automated that at a certain point I had difficulty explicitly recalling what the phone numbers actually were. I simply wanted to call a certain person and my fingers seemed as if they were moving on their own. However, it sometimes happened that I wanted to call a good friend of mine but my parents would answer, at which point I realized that I had dialed the wrong number. In ordinary language this might simply be called a mistake. Yet, in my defense, I did do at least something right, namely correctly dial my

parents' telephone number. What went wrong was not that I performed a certain movement procedure incorrectly, but that the correctly performed procedure was not the one I tried to perform. This is different from another kind of mistake in which I dialed the incorrect digit sequence and was either connected to a complete stranger, or was subjected to some version of *the number you have dialed is not available*.

Now that corded phones are becoming a rarity and telephone numbers are saved under specific names into phones, these incidents are facing extinction. Yet, what stayed the same is that it sometimes happens that certain learned movement procedures are executed correctly but it is not what was tried to be performed. Since this kind of error does not usually have substantial consequences in everyday life, it shares the same term with other types of errors. Yet, distinguishing these cases of what actually went wrong continues to be and should be of great interest for psychological research.

2.1.1) Early Investigations: Errors, Mistakes and (Freudian) Slips

Although the psychology of errors is interesting in its own right, only a few important points can be presented here. Since the kinds of errors or mistakes described above occur every once in a while in everyday life, it is not surprising that they caught the attention of early scientists and philosophers. One remnant that shows this is that the specific kind of error we make by saying something we actually did not want to say has its own name, *Freudian slip*. While Sigmund Freud saw this as an outburst of repressed impulses (Freud, 1917; Reason, 1990), later theories tried to explain errors like these in a different light. Generally, the modern literature on errors differentiates between *mistakes* on the one side, and *slips* or *lapses* on the other. According to the standard definitions, mistakes are characterized by failure of an intended action to lead to a desired goal due to an inadequate plan (planning failure), while slips and lapses are actions that are not performed as planned (execution failure) (Reason, 1990). However, the traditional literature on errors often overlooked developments in procedural memory research, or mentioned memory primarily with regard to lapses as a forgetting phenomenon. I will mention two notions that did build bridges between research on errors and memory research, crosstalk and interference, but show why they are inadequate for the specific phenomenon portrayed here.

2.1.2) Crosstalk

One notion that was used to explain this kind of mistake was the idea of crosstalk "in which elements of one task either bias responses in the other or actually migrate from one activity to the other" (Reason, 1990, p. 29). Crosstalk seemed especially important in explanations of diminished performance when trying to perform two tasks simultaneously or in close succession (often called *dual-task performance*) (Kinsbourne, 1981), but is also used for the study of single tasks as well (Collins & Hay, 1994; Terzis, 2001). However, crosstalk is too broad and abstract to specifically capture just the kind of mistake of falsely remembering-how that is aimed to be characterized here. Under the heading of crosstalk the example described above where I mistakenly called my parents instead of my friend could either be due to the correct performance of an unintended movement procedure (error on the level of procedural memory), or due to the correct performance of an intended but, given the context, ultimately inappropriate movement procedure (error on the level of declarative memory). Concretely, to use an example put forth by William James that is often cited in this

context (James, 1983b, p. 119), we might imagine that someone goes to their bedroom with the intention of changing their clothes for dinner but ends up going to bed. James explains this in the manner that environmental cues can lead to an automatic response (1983b). Undressing in the evening in one's bedroom is what one usually does in order to go to bed. Again, using crosstalk at least two explanations of this phenomenon would be valid. It might be an error concerning procedural memory: wanting to change clothes but somewhat absent-mindedly performing a different movement procedure. But, it might also be an error on the level of declarative memory: wanting to change clothes but finding oneself undressing in the evening in one's bedroom and concluding that what one really wanted to do was to go to bed (as this is what one usually does in those circumstances), and as a result correctly performing the movement procedures needed to undress and go to bed. Both are adequate explanations using the notion of crosstalk as either an unwanted but similar procedure 'takes over' the intended one, or environmental cues bias one's attention in a way that suggests that what one really wanted to do was something else which is then performed correctly.

One reason for this broad and rather unspecific applicability might be that crosstalk is more of an abstract metaphor from engineering for interference (Kinsbourne, 1981; Pashler, 1994) with quite different characterizations than a well-defined psychological phenomenon. While certain ideas, such as similarity between the crosstalking elements, are usually shared by different uses or notions of crosstalk, their more specific realizations vary widely depending on the underlying theoretical assumptions, such as ranging from actual physical similarity (Terzis, 2001) to overlap of abstract functional patterns (Collins & Hay, 1994; Kinsbourne, 1981).

2.1.3) Interference

A different explanation coming from cognitive psychology is the idea of interference. Classic interference theory states that interference occurs when one learned thing hinders the recall of another learned (or to be learned) thing (Underwood, 1957; Wohldmann, Healy, & Bourne, 2008). While this has been researched quite extensively, also for motor memory (Koedijker, Oudejans, & Beek, 2010; Panzer, Wilde, & Shea, 2006; Wohldmann, Healy, & Bourne, 2008), it ultimately leaves open what concretely the result of the interference is. In its traditional form, interference occurs when one memory or instance of learning leads to failure of recall of a specific memory, or decreased performance in a specific memory task. This brings us back to the ordinary language notion of mistake and explains part of the mechanism behind it, but does not specify if the error results in correctly performing a learned movement procedure instead of an intended movement procedure (dialing my parents' number when wanting to dial my friend's telephone number), or simply in incorrect performance (dialing digits that correspond to a stranger's telephone number or to no assigned number at all). Both would lead to decreased performance in a specific memory task. Making a distinction between these two outcomes, and the respective types of error, as will be done in the following, can help improve future studies on interference and application of their results in general.

Consequently, while the phenomenon of interest, which I term false remembering-how, has, in some form or another, partly been described by earlier theories, they are too unspecific or abstract to capture the particular differences between false remembering-how

and other procedural memory errors. The notion of (action) slips comes closest to what I have in mind but is generally not connected to the idea of procedural or long-term memory. This is why I will, after specifying what is usually meant by procedural or habit memory, lay out characterizations that distinguish genuine remembering-how, lucky shots in procedural memory, unsuccessful remembering-how and false remembering-how.

2.2) What is Procedural Memory?

Scientists and philosophers have often used different classifications of memory, and one type of memory in one classification rarely corresponds perfectly to any type of memory in another classification. For example, the philosopher's experiential memory usually describes something different from, albeit similar to, the scientist's episodic memory (Bernecker, 2011). Procedural memory in science and habit memory (or *practical memory*) (Bernecker, 2011) in philosophy are both used to refer to the type of memory that is employed when recalling learned behavior, such as remembering how to actually ride a bike. In everyday speech this kind of memory is often described with terms such as *muscle memory* or *skills*, stressing action-oriented aspects. There seems to be little to no meaning difference between procedural memory and habit memory, and the use of procedural memory is becoming more widespread in contemporary philosophy (Robins, 2017) as well. Accordingly, I will generally use procedural memory to refer to both, the scientist's procedural memory as well as to the philosopher's habit memory.

While there is no explicitly agreed upon definition of procedural memory, in scientific research it is typically understood as "the memory system in charge of encoding, storing, and retrieving the procedures that underlie motor, verbal, and cognitive skills" (Beaunieux, et al., 2006, p. 521). Put differently, "procedural knowledge refers to 'knowing how' to do something (such as riding a bicycle) and represents one's knowledge of procedures that is gained through experience" (Knowlton, Siegel, & Moody, 2017, p. 295). This last definition already classifies the acquired skill as a type of knowledge described as knowing how which is represented in some way. I will not assume whether or not procedural memory is a type of memory (or knowledge) that represents something. For now, it is important to note that definitions of procedural memory usually focus on skills or procedures which are learned, acquired, retained, represented or in another way made (in principle) accessible to a subject or system.[14]

In scientific research, these procedures or skills are often distinguished into three different kinds: perceptual, motor and cognitive (Knowlton, Siegel, & Moody, 2017), even though most, if not all, complex actions seem to require all three forms (Adams, 1987). Perceptual and motor skills are sometimes subsumed under, or seen as, a single sensori-motor or perceptual-motor skill (Rosenbaum, Carlson, & Gilmore, 2001). Given the purpose of this paper, the differences between these forms of procedural memory are negligible and

[14] The ways the term (*procedural*) *memory* (and also *remembering*) are used are generally inconsistent across the literature. It is sometimes used to denote a system, a mental state, a form of knowledge, a change of behavior or a change of internal representations, among other uses. For discussions related to the different uses of memory and remembering see (Werning & Cheng, 2017), and (Roediger, Dudai, & Fitzpatrick, 2007) for the uses in science more generally. Similar inconsistencies can be found related to the terms *know(ing)-how* and *remember(ing)-how* as well, which is why relevant differences between them are elaborated later in the paper.

the main focus is on (sensori-)motor skills or procedures, such as remembering how to ride a bike.

Habits are sometimes classified as procedural memory as well (Knowlton, Siegel, & Moody, 2017), and unsurprisingly as habit memory in philosophy (Bergson, 1991). But, since habits closely resemble learning through association of stimulus and response in conditioning, they could arguably be seen as a type of conditioning (which, given the taxonomy referred to above, would be different from procedural memory) (Squire, 2004). The term *habit memory* is sometimes also used in science to refer to acquired habits (Milner, Squire, & Kandel, 1998).

2.3) Knowing-how vs. Remembering-how

The knowing that vs. knowing-how debate is much too voluminous to be described in-detail here, but it is still important since a lot of the discussion about knowledge or memories of skills and procedures took place there; cf. (Fantl, 2017) for an overview, or (Bengson & Moffett, 2011) for a more comprehensive debate. That is why this section focuses on two things that are of relevance here: First, the thesis that knowing-how and remembering-how are different is introduced. Second, the idea that much of the talk about knowing-how in philosophical literature refers to remembering-how is laid out. This distinction will help answer the question of how remembering-how to do something and falsely remembering-how to do something might be defined.

Knowing-how is different from Remembering-how

Even though one of the most prominent questions in the knowing that vs. knowing-how debate is whether or not knowing-how is in fact a type of knowing that, cf. (Bengson & Moffett, 2011), it is not discussed here further (at least not explicitly). The differences between knowing-how and remembering-how have received much less attention. While it might not only be limited to procedural memory, I tend towards seeing knowing-how as referring to a subject's knowledge state (that is, what a subject knows), and remembering-how as referring to a subject's acts (that is, what a subject does). This is not an entirely novel view of memory, as it has been proposed in the past (Moscovitch, 2007; Russell, 2005), and some contemporary philosophers also seem to share the notion of action-orientedness. For example, Dorothea Debus states that procedural memories "usually manifest themselves in *physical* actions, such as someone's actually riding a bike, and are *not* usually elements of a subject's *mental* life" (Debus, 2017, p. 74) stressing that manifestation of procedural memory is realized through action, and not (just) possession of a mental state.

If this distinction holds, it means that I can know-how to do something, for example know-how to ride a bike, without currently performing any (bodily) act, such as riding a bike.[15] Thus, to truly say that I know-how to ride a bike does not imply that I am, at the

[15] I use the terms knowing-how and remembering-how in the hyphenated form to refer to cases captured by the terms procedural memory or habit memory. Sometimes knowing/remembering how are used in a different sense (such as saying that someone today knows how Caesar died) which is not the main concern here. Similarly, procedural memory is sometimes expressed through the use of other words than knowing or remembering (such as saying that someone can play the piano well; though context is quite important here) (Ryle, 2009a).

respective time, riding a bike. However, to truly say that I remember-how to ride a bike does, given that procedural memory manifests itself in certain acts, presuppose that I perform said act at the respective time. In this respect, someone can know-how to do something without remembering-how to do it. When exactly someone remembers-how to do something is elaborated later.

I propose that this distinction might help clear up some of the confusion and cases proposed in the knowing that vs. knowing-how debate. Jason Stanley and Timothy Williamson's piano player who lost their hands (2001) could, under this framework, be said to *know*-how to play the piano but not *remember*-how to play it, which might express that he has learned it, and has not forgotten it (Ryle, 2009a), but does not perform the act in question. I am inclined to agree with Gilbert Ryle in that exercising knowing-how can take multiple forms (2009b). The handless piano player could still exercise his knowledge-how to play by, for example, correcting their students when they make a mistake. But, these exercises of knowledge-how are, under this framework, not instances of remembering-how, since the handless piano player cannot actually play the piano anymore. In some of the cases in the knowing that vs. knowing-how debate, knowing-how is used in the sense of *knowing-how*, while in other cases it is used in the sense of *remembering-how*. Distinguishing these cases might help resolve some of the issues surrounding that debate.

2.4) Procedural Memory and False Procedural Memory

The following account of procedural memory is quite similar to Katherine Hawley's notion of knowledge-how (2003), in which she argues that knowledge-how is defined by warranted counterfactual success. Similar to Hawley, I do not necessarily presuppose that knowing-how is or is not made up of knowing that. However, it seems that the following characterizations are most compatible with an anti-intellectualists ability account of knowing-how, which (roughly) states that knowing-how is not simply a form of (or not entirely dependent on) knowing-that but instead is defined by having an ability to do what a subject knows-how to do (Fantl, 2017). Since the main focus here is on *remembering*-how, and since, as stated above, I take remembering-how to be different from knowing-how, remembering-how will require its own account. In her definition of knowledge-how, Hawley states that for someone to know-how to do something they (among other conditions) would have to succeed under certain counterfactual conditions if they tried to perform an action (Hawley, 2003). Similarly, I take it that success under certain conditions is a necessary condition for remembering-how.

The general account is given next to an example case of remembering-how. I have chosen the performance of a backflip because it is a rather clear-cut, discrete instance of when the action is performed successfully (usually, when someone jumps and rotates their body roughly 360° backwards in the air, and in the end lands on their feet again).

2.4.1) Not Remembering-how

Since I propose that remembering-how is an act, someone does not remember-how to do something when they do not perform the relevant act given the appropriate circumstances:

Case 1 — Not remembering-how:

C1: x does not remember-how to φ at time t(2) if:	C1: x does not remember-how to do a backflip at time t(2) if:
P1: x does not φ at time t(2) given the appropriate circumstances	P1: x does not do a backflip at time t(2) given the appropriate circumstances

It might be claimed that this implies that a master pianist who lost their hands cannot remember-how to play the piano. Similarly, it might be objected that, for example, someone who attempts a backflip but due to an unusually strong gust of wind is blown away from their course, would fail to perform the backflip and thus, according to my definition, would not remember-how to do a backflip. To counteract such objections and focus on the relevant phenomena, it is proposed that someone only remembers-how to do something if they perform the relevant act, or a sufficiently similar act, under appropriate counterfactual conditions. What conditions are deemed appropriate and what kind of act is sufficiently similar to the act in question will strongly depend on the context and will rarely have clear-cut criteria. Hawley gives the example that knowing-how to drive a car might mean something else in the USA (driving an automatic car might suffice) than in the UK (driving a manual shift car might be necessary) (2003). Since I take remembering-how to be an act, someone does not remember-how to do something if they do not perform that act given the appropriate circumstances.

2.4.2) Remembering-how

To distinguish remembering-how from cases in which someone out of sheer coincidence or luck performs an act, two further conditions are added. First, the remembered act has to have been learned, or in some way acquired, before; that is, the subject will have to have acquired the ability to perform the act in question. Second, had the subject not acquired the ability to perform the act in question, they would not have performed the act at the respective time. This also seems to necessitate that the ability needs to have been acquired at a time *before* the performance in question, to avoid cases such as backtracking counterfactuals (Lewis, 1979). Consequently, remembering-how to perform an act will be defined in the following way:

Case 2 — Remembering-how (procedural memory):

C2: x remembers-how to φ at time t(2) if and only if:	C2: x remembers-how to do a backflip at time t(2) if and only if:
P1: x φs at t(2) given the appropriate circumstances P2: x has acquired the ability to φ given the appropriate circumstances at a time(span) t(1) which preceded t(2)	P1: x does a backflip at t(2) given the appropriate circumstances P2: x has acquired the ability to do a backflip given the appropriate circumstances at a time(span) t(1) which preceded t(2)
P3: x would not have φed at t(2) given the appropriate circumstances if P2 had been false	P3: x would not have done a backflip at t(2) given the appropriate circumstances if P2 had been false

2.4.3) Lucky Shots

Successful remembering-how should be distinguished from cases in which someone just out of luck performs an act which can be learned and remembered. These include cases of beginner's luck, such as trying to do a backflip and coincidentally succeeding. In these cases we might want the person to repeat the act to make sure that it was not in fact just a fluke. However, repeating or failing to repeat the act in order to make sure seems to be more of an epistemological question, which expresses knowing whether the person in question remembers how to perform an act or was just lucky. This is independent of the person actually remembering-how to perform the act, which means actually *performing* it because they learned it. If someone out of luck performs a certain act, it is characterized as follows:

Case 3 — Lucky Shot

C3: x does not remember-how to φ but by chance φs at time t(2) if and only if:	C3: x does not remember-how to do a backflip but by chance does a backflip at time t(2) if and only if:
P1: x φs at time t(2) given the appropriate circumstances P2: x has not acquired the ability to φ given the appropriate circumstances at a time(span) t(1) which preceded t(2)	P1: x does a backflip at time t(2) given the appropriate circumstances P2: x has not acquired the ability to do a backflip given the appropriate circumstances at a time(span) t(1) which preceded t(2)

Here it might be objected that if someone performs the act in question, such as doing a backflip, they also know-how or have the ability to do it (Fantl, 2017). In this view, not having the ability to perform an act and performing an act might contradict each other. A similar objection was raised in the knowing that vs. knowing-how debate and, in my opinion, answered quite well by Alva Noë (2005; Fantl, 2017). He distinguished between being able to do something and having the ability to do something. While someone who has never attempted a feat might still *be able* to pull it off out of sheer luck, they would still not *have the relevant ability*, meaning they did not learn how to perform the task before the time in question. However, to remember-how to do something does seem to presuppose that

someone has learned it before. Otherwise, if someone were to ask what a novice can remember-how to do, it would have to be an almost endless list of skills which they might simply perform out of sheer luck, cf. (Fantl, 2017; Glick, 2011).

It is conceivable to construct a case similar to the Lucky Shot, where the subject has acquired the ability to perform a certain act but the acquired ability plays no instrumental role in the subject's performing the respective act. This might seem highly improbable at first, as it might be implicitly assumed that nothing intervening has occurred in the meantime. However, it might be objected that the subject has acquired the ability to perform a certain act, then has (for example, due to brain damage) lost this ability, and then performs the respective act in the same fashion as someone who had never acquired the respective ability the first place. This might very well be called a lucky shot, even though it does not correspond to the characterization of a Lucky Shot given above. Ultimately, the essential characteristic in lucky shots seems to be that the possible acquisition of a certain ability plays no instrumental role in the performance of the corresponding act. However, as cases where one acquired a relevant ability and lost it but then coincidentally succeeds (or similar cases) seem highly improbable, they might be more distracting than helpful in the endeavor set out here.

2.4.4) Unsuccessful Remembering-how

Apart from luck shots, there also appear to be cases in which someone tries to remember-how to perform an act but fails to do so. There seem to be two kinds of distinct cases which in natural language are both called *error* or *mistake*:[16] One in which something that is tried to be remembered is incorrectly executed (unsuccessful remembering-how), and another in which something learned is executed correctly but it is the 'wrong' thing, meaning something that the subject did not try to perform (false remembering-how). While it is conceivable to try and characterize errors through other means than intention, such as deviation from general practice, it seems that intention is the most intuitive and generally accepted angle under which errors or mistakes are viewed (at least in the psychological literature on errors); for others cf. (Reason, 1990). Reason even goes so far as to claim that "the notions of intention and error are inseparable" (1990, p. 5).

In unsuccessful remembering-how it does seem necessary that some kind of act, or restraint of act, is executed that is different from the act that was learned and intended. Otherwise, it would either be no act at all (not doing anything, including not restraining one's acts), or intended successful remembering-how (correctly performing a learned act one tried to perform). To distinguish unsuccessful remembering-how from false remembering-how to do something, the unintendedly performed act must not be something that the subject has acquired the ability to do:

[16] I will disregard the distinction sometimes made in psychological research between errors, mistakes, slips and lapses as it is too comprehensive to be considered here; cf. (Reason, 1990) for an overview.

Case 4 — Making a mistake in remembering-how (unsuccessful remembering-how):

C4: x makes a mistake in remembering-how to φ at time t(2) if and only if:	C4: x makes a mistake in remembering-how to do a backflip at time t(2) if and only if:
P1: x does not φ at time t(2) given the appropriate circumstances	P1: x does not do a backflip at time t(2) given the appropriate circumstances
P2: x has acquired the ability to φ given the appropriate circumstances at a time(span) t(1) which preceded t(2)	P2: x has acquired the ability to do a backflip given the appropriate circumstances at a time(span) t(1) which preceded t(2)
P3: x does something other than φing at time t(2) which x has not acquired the ability to perform before t(2)	P3: x does something other than a backflip at time t(2) which x has not acquired the ability to perform (e.g. falling on x' head) before t(2)
P4: x tried to φ at time t(2)	P4: x tried to do a backflip at time t(2)

2.4.5) False Remembering-how

In the second type of mistake, something is correctly remembered but it is not what was tried to be performed. All conditions for successful remembering-how are met (compare P1, P2 and P3 in Case 2 to P3, P4 and P5 in Case 5). If it were left out that the subject tried to perform an act (P6), it would, given the definition in Case 2, have to be concluded that this simply is a case of successful remembering-how. And in fact in a way it is successful remembering-how, but intuitively it still seems incorrect to say that everything went fine when, instead of what was tried to be performed, a different act was performed. As mentioned in the telephone number example in the beginning, imagine someone who does not explicitly recall certain telephone numbers but has dialed these numbers so often that they simply move their fingers in a specific fashion without explicitly recalling the exact number. If that person wanted to dial the telephone number of a good friend but instead dialed their parents' number, we might be reluctant to say that it was simply a case of successfully remembering-how to move their fingers in order to dial their parents' phone number. They did correctly remember-how to dial their parents' phone number, but it was not what they wanted to remember-how to do. Consequently, it seems that falsely remembering-how to do something can be characterized in the following way:[17]

[17] This particular example, doing a front flip when trying to do a backflip, might seem highly unlikely in isolation, but it could very well occur in a more complex dance routine.

Case 5 – False remembering-how (false procedural memory)

C5: x falsely remembers-how to φ at time t(2) if and only if:	C5: x falsely remembers-how to do a backflip at time t(2) if and only if:
P1: x does not φ at time t(2) given the appropriate circumstances	P1: x does not do a backflip at time t(2) given the appropriate circumstances
P2: x has acquired the ability to φ given the appropriate circumstances at a time(span) t(1) which preceded t(2)	P2: x has acquired the ability to do a backflip given the appropriate circumstances at a time(span) t(1) which preceded t(2)
P3: x does something other than φing, e.g. ψing, at time t(2)	P3: x does something other than a backflip, e.g. a front flip, at time t(2)
P4: x has acquired the ability to do something other than φing, e.g. ψing, given the appropriate circumstances at a time(span) which preceded t(2)	P4: x has acquired the ability to do something other than a backflip, e.g. a front flip, given the appropriate circumstances at a time(span) which preceded t(2)
P5: x would not have done that other thing, i.e. ψing, if P4 had been false	P5: x would not have done that other thing, i.e. a front flip, if P4 had been false
P6: x tried to φ at time t(2)	P6: x tried to do a backflip at time t(2)

This characterization also excludes combinations of cases, such as not remembering and a lucky shot after another. For example, if someone who knows-how to do backflips but does not know-how to do front flips were to try and do a backflip but, as luck would have it, were to unfortunately slip and do a front flip instead, they did a front flip by sheer chance, not by remembering-how or falsely remembering-how to do a front flip. Falsely remembering-how to do a front flip requires that the subject first acquired the ability to perform a front flip and (among other things) that they would not have done a front flip, had they not acquired the ability. Furthermore, *unfortunately slip* sounds rather like wanting to express an accident than any kind of learned behavior, and, therefore, it seems to be a lucky shot even from a rather intuitive standpoint.

After considering some of the possible objections that might arise against the presented account, I will compare the presented notion of false procedural memory with contemporary theories of false declarative memory, and sketch the relevance such an account might have on current scientific research.

2.5) Possible Objections

Given that the literature in the knowing that vs. knowing-how debate is quite extensive, it is not surprising that some of the issues might apply to the account of remembering-how presented here. Some possible objections will be considered in the following.

2.5.1) The presented Account presupposes Anti-Intellectualism

Since the acquisition of a certain ability is a central part in my characterization of remembering-how, it might be argued that I presuppose an anti-intellectualist ability account. This would, however, be an overstatement. Clearly, as stated, a certain ability has to be acquired in order for someone to remember-how to perform something. But, whether or not this ability is propositional in nature is not presupposed. Hawley states: "we know

how to do things which we've never actually attempted. For example, if you've paid due attention to the safety demonstrations on planes, then you know how to put on a lifejacket, even if you've never tried it" (2003, p. 20). I do not make any claims about whether or not the acquisition of a skill could occur solely through observation or imagination. Empirically, it has been suggested that in some cases it is possible to acquire a motor skill (Helene & Xavier, 2006) or transfer improvements in one task to a new but similar task via mental practice alone (such as reading mirror-inverted words, or typing digits on a keyboard quickly with your left hand after learning to do it with your right hand) (Wohldmann, Healy, & Bourne, 2008). Although still debated, it has been proposed that physical practice differs from mental practice in that, among other things (Nyberg, Eriksson, Larsson, & Marklund, 2006), the former enhances what is termed *effector-dependent representation* as well as *effector-independent representation*, while the latter enhances effector-independent representation only (with *effectors* describing the body parts which are used to execute a task) (Wohldmann, Healy, & Bourne, 2008). I do take it that acquisition of many motor skills usually referred to with terms such as *procedural memory* need a lot of physical practice, but I leave it open whether this is a necessary criterion in every case and with that whether or not relevant abilities are nonpropositional in nature.

2.5.2) Remembering-how, like Knowing-how, needs an Intention

One issue in the knowing that vs. knowing-how debate is whether or not the known act needs to be an intentional one. Hawley points out that "success cannot amount to knowledge-how unless intentional action is involved. We don't describe ourselves as knowing how to produce white blood cells" (2003, p. 26). I agree that a human subject cannot normally remember-how to, for example, produce white blood cells.[18] But, this might not be a reason to make intention a *necessary* condition for successful remembering-how. A premise presented here is that *x ϕs at t(2) given the appropriate circumstances*. Normally, producing white blood cells is not something the subject x does. It might be something that happens in x' body, but it is not an act of x; compare Noë's critique (2005) of Stanley and Williamson's claim that one does not know-how to digest food (2001). Consequently, it is not something x can remember-how, and an intention condition is not necessary. Similarly, many of the things we *do* remember-how to do, especially those we are very skilled at, can be executed without an intention. Take for example a black-belt martial artist who is unexpectedly thrown to the ground. If they have ample time, they might very well form an intention and remember-how to fall correctly so as to minimize damage. However, given the usually very short time they have to react, it is more than likely that, due to their years of practice and automatization of falling techniques, they will remember-how to correctly fall without first forming an intention or elaborate contemplation which falling technique would be the most adequate given the current position of their body to the ground. In other words, remembering-how can be initiated by intention, but it does not have to be.

[18] This is not to say that someone with a vast knowledge of human physiology cannot know how white blood cells are produced in human bodies, and use this knowledge in a way that leads to an increase in the amount of white blood cells in their body (through training at certain altitudes or medication).

2.5.3) Warrant is a necessary Condition for Remembering-how

As mentioned before, in Hawley's account of knowledge-how warrant is a necessary condition. However, warrant is not included in the presented account here. One might therefore ask why knowledge-how requires warrant, and remembering-how does not. Hawley gives the example of Susie, who tries to annoy Joe by smoking. Joe, however, is not annoyed because Susie is smoking, but because she taps on her cigarette box in an idiosyncratic way, which she does every time she smokes (2003). We might be reluctant to say that Susie knows-how to annoy Joe, because she has a false belief about it and lacks understanding. In the same vein we might say that Susie does not remember-how to annoy Joe. However, given the definition of remembering-how presented here, I would have to say that she remembers-how to annoy Joe. And in fact, I do hold that she remembers-how to annoy Joe. The fact that she holds a false belief about what it is that annoys Joe, might exclude her from knowing that tapping her cigarette box annoys Joe, but does not exclude her from knowing-how or remembering-how to annoy Joe.

If it were the case that understanding how one succeeds was a necessary condition for knowing-how or remembering-how, then false beliefs about, for example, riding a bike might exclude proficient bicyclists who have a false belief about bike riding from being ascribed as knowing-how or remembering-how to ride a bike (Fantl, 2017). But, empirical observation suggests that most proficient bike riders do have false beliefs about how to ride a bike (such as claiming to lean left when noticing that they are falling to the right) (Curran, 2001), and we might be reluctant to say that someone who has been riding bikes all their life with great success does not know-how to ride a bike because they hold false beliefs about how they actually do it. While I do agree with Hawley that success in knowing-how (or remembering-how) must be non-accidental to count as knowing-how (or remembering-how) it seems to me that being mistaken or simply not having any (explicit) belief *about* how one does something (such as it often seems to be the case in amnesiacs) (Milner, Squire, & Kandel, 1998; Squire, Cohen, & Zouzounis, 1984), does not exclude someone from actually knowing-how or remembering-how to do it. The presented account here assures that remembering-how is non-accidental through the necessary and instrumental counterfactual requirement of having acquired the ability to perform the act.

2.6) Comparison with Philosophical Theories of False Declarative Memory

There are a number of contemporary theories that define false declarative memories or, strictly speaking, mnestic confabulations.[19] Two of the most recent and, in my opinion, most elaborate are the simulation-theoretical account of memory most explicitly formulated by Kourken Michaelian (2016b) and the causal account of mnestic confabulation by Sven

[19] Depending on the view, some make a distinction between false memories and mnestic confabulations. Bernecker, for example, characterizes confabulation as a kind of error "that occurs when patients produce stories that fill in gaps in their memories" (Bernecker, 2017b, p. 1). If these confabulations are due to failure of a memory system, they are said to be *mnestic* (Schnider, 2018) or *mnemonic* (Bernecker, 2017b) (terms used synonymously). According to this view, a mnestic confabulation is a false memory if what is being confabulated is false. For a more in-depth discussion about types of confabulation from a psychological perspective see (Schnider, 2018).

Bernecker (2017b), with which I will compare the account of false remembering-how presented here.

2.6.1) A Reliabilist Alternative of (False) Procedural Memory

Michaelian's account omits necessary counterfactual connections altogether and instead relies on the reliability of a memory system (2016b). Simply put, under his view a subject confabulates if their remembering is the result of an unreliable memory system. Omitting the necessity of any counterfactual connection is incompatible with the account I have presented here, since the acquisition of an ability in the past is necessary and instrumental in remembering-how and in false remembering-how. However, it would be possible to replace the counterfactual conditions and adept them to a reliabilist account similar to Michaelian's simulation-theoretic account if one wanted to. Concretely, this would mean that x remembers-how to φ if and only if i) x φs at a specific time given the appropriate circumstances, and ii) x' φ-ing is the result of x' reliably working procedural memory system. False remembering-how would be defined analogously (trying to φ but ending up ψing when usually reliably performing φ or ψ). Therefore, it seems to be possible to keep important aspects of (false) remembering-how as presented here (such as conditions differentiating false remembering-how from remembering-how), whether one chooses counterfactual or reliablist conditions, which enables integration of essential features of (false) remembering-how into different prominent theoretical frameworks. On the other hand, lucky shots would be quite difficult to include. Veridical confabulation in declarative memory might correspond to a lucky shot where an unreliably working procedural memory system leads to the performance of an act usually ascribed to a reliably working memory system (unreliable performance could, for example, be due to lack of training in a novice).

2.6.2) Comparison with a Causal Theory of Confabulation

Since the account presented here relies on necessary counterfactual conditions it seems that it shares a few similarities with Bernecker's account (2017b), which also makes use of counterfactual conditions. However, a direct application of how Bernecker defines mnestic confabulations would not yield what I termed false remembering-how or false procedural memory. Under his account, a mnestic confabulation differs from genuine memory by not fulfilling the counterfactual condition; roughly put: lacking counterfactual dependence on certain past representation is the defining factor of mnestic confabulations. This lack can be realized in two ways. Either there actually is no past representation but the mnestic confabulation suggests there is one. Or, there is a respective past representation but the mnestic confabulation does not depend on it and any match is coincidental (Bernecker, 2017b). This would come closest to the Lucky Shot case described here, where there either is no acquisition of an ability in the past or possible acquisition of an ability is not instrumental in the performance of an act. Yet, it would seem quite unnatural and counterintuitive to speak of false memory when someone out of luck succeeded in performing an act. There might not be a case in Bernecker's account that corresponds well to what I have termed *false remembering-how* as intention plays no role in his definition of memory or mnestic confabulation. However, as I have stated in the beginning, we do differentiate types of mistakes from lucky shots; a distinction which seems to rely on some form of intention.

Thus, direct application of Bernecker's account would likely not distinguish these types of mistakes, which was one of the objectives set out in this paper.

Furthermore, Bernecker's account and my account of false remembering-how differ in whether false memories are to be considered memories at all. He does not consider false memories or mnestic confabulations as a form of memory (at least in the case of declarative memory), since if an apparent memory is false it does not represent objective reality (one of the necessary conditions of his account for a mental state to qualify as a memory), or if the apparent memory is true but does not depend on the respective past representation appropriately it does not fulfill the causal condition (also a necessary condition for a mental state to qualify as a memory). However, I do see false procedural memories as defined here as memories since they fulfill all the conditions I have set for genuine remembering-how.

In the next chapter, the value of the presented account of false procedural memory for scientific research is laid out.

2.7) Relevance for Scientific Research

At first, such a characterization of false procedural memory might seem overly complicated to describe something all that common. But as the telephone number example above illustrated, this phenomenon might occur often but has simply not been given a memory term like it is the case in false declarative memory. On the other hand, psychological research does investigate similar phenomena to what has been described here as false procedural memory. Interference theory, where one learned thing hinders the recall of another learned (or to be learned) thing (Underwood, 1957; Wohldmann, Healy, & Bourne, 2008), is a good example as it has been researched quite extensively, including procedural motor memory (Koedijker, Oudejans, & Beek, 2010; Panzer, Wilde, & Shea, 2006; Wohldmann, Healy, & Bourne, 2008).

2.7.1) Interference Theory and False Procedural Memory

Interference theory has traditionally been used to explain forgetting phenomena, but in contemporary research it is quite common to describe failure of memory recall or decrease in task performance due to another memory as *interference*, but increase in task performance or speed of acquisition due to another memory as (positive) *transfer* or *facilitation* (Künzell & Lukas, 2011; Panzer, Wilde, & Shea, 2006; Wohldmann, Healy, & Bourne, 2008). Compared to other types of nondeclarative memory (for example, false implicit memory in priming) (McKone & Murphy, 2000), *false procedural memory* is, however, not an established term in these explanations. Further distinguishing different cases of interference, as is proposed here with the notion of false procedural memory when appropriate, could help refine studies on interference by differentiating between distinct outcomes of the interference. Concretely, this could mean to not only analyze whether error rate increases or speed of acquisition decreases, but also disentangling the types of errors made, and thus distinguishing interference effects and circumstances that result in incorrect performance of an intended act from those that result in successful recall of something that was learned but which was not tried to be recalled. This is proposed in order to separate conceptually similar memory phenomena, and thereby improve studies on interference and their respective application.

Connecting the research field on errors with interference theory in this way could enable researchers to forge links between false procedural memory to research in other more popular types of false memory phenomena, such as false declarative memory, but also highlight differences between them. For example, the parallels of false procedural memory as described here to false memory in priming studies could be investigated. In classical priming studies concerning false memory (Roediger & McDermott, 1995), it is usually the case that participants are presented with a list of words which are somehow thought to be related to another (for example, *web*, *insect*, *crawl*, *poison*), and later asked to recognize or recall which words were on that list. Typically, participants 'remember' a false lure, meaning an item which was never presented but is somehow related to the presented items (such as *spider*) (Roediger & McDermott, 1995). Concerning relatedness, falsely 'remembering' the lure, which was never on the list, is still different from falsely recalling a random item which also was not presented (such as *window*). Given that this might be explained in terms of spreading activation (Roediger, Balota, & Watson, 2001), meaning the list items activate and thus increase the probability of recall of the lure, parallels to false procedural memory, interference or transfer can be proposed. Usually, task similarity plays a major role in the occurrence of interference or transfer (although the details seem less straightforward) (Bock, Schneider, & Bloomberg, 2001; Kremer, Spittle, & Malseed, 2011; Wood & Ging, 1991). It has been hypothesized that facilitation/transfer of one motor skill (such as skateboarding) to a similar one (such as snowboarding) also is due to activation of shared neuronal networks (Künzell & Lukas, 2011), and similar neuroanatomical features between tasks or sets of neurons which rhythmically fire could be a source of interference (Walter & Swinnen, 1994). Whether or not false procedural memory is indeed caused by a similar or the same type of mechanism (such as spreading activation) as in the recall of a false lure in priming studies remains to be seen, but this could be a promising starting point for unifying different phenomena under the notion of false memory.

2.7.2) Application and Development of False Procedural Memory Research

Research on interference and types of errors both continue to be successfully applied in many different areas. For example, interference theory has become a standard in motor learning textbooks (Edwards, 2010) and, as mentioned in the beginning, trying to minimize interference effects often is a relevant part of training schedule or machine design. Further distinguishing the kinds of errors as I have proposed here can help improve upon such applications, as it can make a crucial difference if someone simply performs something incorrectly, or correctly does something else then they actually tried to perform.

Imagine a complex machine such as an aircraft cockpit with multiple types of controls such as buttons to be pressed, switches to be flipped or knobs to be turned. In such circumstances it can, for example, happen that the person operating the machine tries to push a button down but instead performs a switch-flipping movement (as the two finger movements are similar) leading to the activation of an adjacent switch instead of the button. This would be distinct from incorrectly pushing the button (by, for example, applying insufficient force). Given the lack of standardization in different types of aircrafts, machine designers should pay attention to this distinction and employ respective countermeasures for each type of error, such as giving unambiguous, immediate and specific visual feedback only after each activation of a button, switch or knob (Wickens, Hollands, Banbury, &

Parasuraman, 2016). The development and application of countermeasures could be improved with a better understanding of different interference effects, namely trying to predict which circumstances are likely to lead to incorrect execution and which are likely to lead to false remembering-how.

Factors and circumstances influencing procedural memory errors

Developing the research on false procedural memory and interference in general further could facilitate understanding of factors which give rise to different types of errors, and thus improve future applications. There has already been some research (mostly in laboratory settings) investigating factors that influence how likely procedural memory errors are to occur, or what circumstances further/hinder certain interference effects. Only a few of those factors or circumstances can be mentioned here as examples. Generally, one might roughly distinguish aspects that are 'internal' to the subject (such as age, attention regulation, expertise or health) and those that are 'external' to it (such as retrieval cues, task similarity or distracting elements). However, it should be noted that usually the distinction between incorrect execution, forgetting phenomena and false remembering-how is not drawn in these studies, and more research is needed. Nevertheless, the following could be promising starting points that could be applied to the investigation of false procedural remembering as described here.

Concerning internal factors, there is evidence that diseases such as Parkinson's disease or Alzheimer's disease make subjects more susceptible than healthy controls to procedural memory errors where a competing response is executed instead of an adequate or intended response due to distracting elements (Wang, et al., 2013; Wylie, et al., 2009). Absent-mindedness or lapses in attention also seem to increase the probability for failures in execution of skilled actions in general (Cheyne, Carriere, & Smilek, 2006) and have also been described in relation to the execution of unintended movements (Reason, 1990). Regarding interference more specifically, some studies suggest that extensive practice of a movement procedure can shield against interference effects from similar movement procedures which were learned later, but is has also been noted that results across the literature are mixed on this point (Ghilardi, Moisello, Silvestri, Ghez, & Krakauer, 2009; Panzer & Shea, 2008). Furthermore, having to ignore stimuli during actions in the past seems to slow down subsequent reaction times towards those same stimuli when they function as action cues and not as distracting or irrelevant stimuli anymore, and increase error rate (an effect often referred to as (*behavioral*) *negative priming*) (Frings, Schneider, & Fox, 2015; Mayr & Buchner, 2007). This finding lends itself to the possible explanation that inhibition of action towards those stimuli was encoded in memory, which later interferes with required action towards the same stimuli (Grison, Tipper, & Hewitt, 2005), but other (complementary) explanations suggesting, for example, conflict between two responses have been offered as well (Frings, Schneider, & Fox, 2015; Mayr & Buchner, 2006; Mayr & Buchner, 2007).

Among external factors, the amount of time a subject has to respond in a given task influences how likely they will execute an unintended or (relative to the task) inadequate response (when distracting elements are present) with less amount of time available leading to more errors (Betsch, Haberstroh, Molter, & Glöckner, 2004; Heitz, 2014; Wylie, et al., 2009). Yet, others warn this effect might be due to lapses in attention as a result of confusing the goals of the task (Seli, Cheyne, & Smilek, 2012). In a relatively elaborate

experiment, Clark, Parakh, Smilek & Roy (2012) investigated the effects of different types of cues on error rates in routine tasks. Concisely put, they found that distracting elements that are presented at locations where a cue for action is usually expected led to a relatively high error rate and concluded that, due to the expectedness of the cue location, insufficient attention was allocated to checking of the cue information and inhibition of routine task performance. This might be a reason to suggest that cues of similar actions should be placed spatially apart (to lower error rates). Yet, the study did not distinguish between different kinds of procedural memory errors, but only whether or not performance led to successful task completion or not, making application of results as suggested difficult.

Consequently, many different factors influence how likely certain procedural memory errors are, and research investigating them further seems promising. Yet, as the examples above show, many studies confound different kinds of procedural memory errors and are often limited to the lab and oversimplified tasks. Holding these procedural memory errors apart as proposed here and investigation of tasks closer to everyday acts outside the lab will be crucial for future research and application.

Conclusion

In this paper, I have tried to best make sense of what *false procedural memory* might be. While seeing remembering-how as an act rather than just an ability might be counter-intuitive at first, it fits well with current neuroscientific conceptions of remembering, and might help clear up some of the confusion surrounding the knowing that vs. knowing-how debate. At the same time, cases characterized by false procedural memory surely are part of everyday experience, as error research suggests. Thus, they might be more intuitively accessible than it seems at first, even though in ordinary language we usually call false remembering-how as well as unsuccessful remembering-how *mistakes* or *errors* without differentiating them further. Yet, as research on interference and its practical applications suggest, delineating different types of errors can be crucial. While I have opted for a counterfactual theory, it is possible to adept the presented account to a reliabilist notion, enabling integration into different philosophical theories. The decision between a counterfactual and a reliabilist version should be made on a case-to-case basis, depending on the surrounding theoretical assumptions. The characterization of false procedural memory presented here can help clarify the similarities and difference between different forms of false memory phenomena, and could with that further research and the establishment of a taxonomy of false memory in general.

3) Remembering Emotions

(The content of this chapter, excluding this note and subchapter 3.5, is part of a manuscript submitted to and under review at Springer in the journal *Biology & Philosophy*.)

Memories and emotions are both vital parts of everyday life, yet crucial interactions between the two have scarcely been explored. While there has been considerable research into how emotions can influence how well things are remembered, whether or not emotions themselves can be remembered is still a largely uncharted area of research. Philosophers and scientist alike have diverging views on this question, which seems to stem, at least in part, from different accounts on the nature of emotions. Here, I try to answer this question in a way that takes an intuitive notion of emotion and includes both scientific as well as philosophical aspects of both emotions and memory. To do this, I first distinguish between implicit emotions, understood as certain physiological and behavioral responses, and explicit emotions, understood as certain conscious subjective experiences. Next, I show how each of these components of emotions can be remembered. Finally, I bring implicit and explicit emotions, and the ways of remembering each of them, together into an explanation that also includes aspects often ascribed to emotions such as cognition or appraisal. This interdisciplinary endeavor aims to serve as a starting point on what it could mean to remember emotions, and in doing so tries to build a bridge between scientific research and philosophical investigation of the memory of emotions.

Introduction

People often experience emotions when remembering events. Yet, it seems unclear in what ways a currently experienced emotion can be related to the remembering of an event. Put more specifically, is it possible to remember an emotion *per se*, that is, in a similar way someone remembers something like a visual experience?

Philosophers and scientists alike have diverging opinions on this question (Debus, 2007; James, 1983a; LeDoux, 1992; Christianson & Safer, 1996; Ribot, 1897; Ross, 1991; Titchener, 1895; Levine, Safer, & Lench, 2006), but a satisfactory answer is still lacking. At the same time, both memory and emotions arguably play a vital role in everyday life. This question gets even more complicated primarily because of the many different accounts surrounding the concept of emotions, and the lack of consensus on this matter – both, between and within different academic disciplines (Dixon, 2012; Izard, 2010; Mulligan & Scherer, 2012). Therefore, to answer the question in a manageable way, I pick phenomena of interest, which form the foundation of a model of emotions that covers intuitive, scientific and philosophical aspects. By combining an empirical basis with a philosophical framework, I try to answer the question whether and how emotions *per se* can be remembered. Bringing these different approaches together, will enable this investigation to serve as a bridge between scientific research and philosophical debate on the interplay between memory and emotions.

Outline and Objectives

After a short introduction that points out the main kind of memory that is under consideration here, I first lay out a differentiation between two components often attributed

to emotions, which I describe as implicit emotions and explicit emotions. Implicit emotions are identified as certain behavioral and physiological responses, while explicit emotions are defined here as certain conscious experiences. Having set the stage, I answer the question of what it might mean to remember each of these components of emotions. Finally, I summarize how exactly these components of emotions and their respective memories can relate to an event, which includes taking a look at cognitive, evaluative and intentional aspects often attributed to emotions.

3.1) The issue at hand – remembering an emotion

Intuitively, it does not seem clear what exactly is meant when someone says *remember an emotion*. Different interactions between memory and emotions are imaginable, and it seems easiest to analyze this by using an analogy to a more familiar concept such as remembering a visual experience.

When I was a child, I moved from a rural mountain area to a busy metropolitan city. The thing that horrified me the most in my new life were the neighborhood dogs. All dogs I had encountered in the mountains were big livestock guardian dogs, which did a great job of scaring every child that would dare cross their field of view. Since a few neighbors in this new city had dogs, I was terrified of going outside in fear of an unwanted encounter with the dogs. One time, one of them got loose from its leash and stormed towards me while I was on my way back home from school. Suffice it to say, I ran for my life in a state of horror.

When I think back to those dogs there is a lot I can remember, but only a few of those things will be of immediate concern here. I can remember facts about them such as that one of them was ash-gray. Memory of concepts and facts like these is often termed *semantic* or *propositional memory* (Debus, 2017; Squire, 2004; Bernecker, 2011), and does not require that one makes reference to something they have experienced. After having read this, you could also remember that one of the dogs was ash-gray, although you have never seen it. But intuitively speaking, you could not remember anything *experiential* about the dogs in the same way I do, as I have experienced events involving those dogs while you probably have not. You could remember seeing other dogs, but it seems that you could not remember seeing *those* dogs I mentioned earlier, unless you actually saw them. Memories of experiences like these are often subsumed under the terms *episodic* or *experiential memory* (Teroni, 2017; Debus, 2017; Squire, 2004; Bernecker, 2011). If I think back, I can in a way 'see' the color of that dog's fur again.[20]

Similarly to remembering these visual impressions, one might ask if it is possible to experientially remember an emotional experience. If I think back to that dog breaking loose, barking and chasing me, I can feel a slight shiver down my spine, and am overcome with a certain uneasiness. But is this enough to say that I remember the emotion of fear I had experienced back then?

[20] In what follows, semantic or propositional memory will not be playing a role. Instead, experiential remembering is the kind of memory that will be the focus here.

3.2) Parts of Emotions

The debate on what exactly constitutes an emotion is long, arduous and with no agreed upon answer (Dixon, 2012; Izard, 2010; Mulligan & Scherer, 2012). Likewise, I will not try and resolve it here. Instead, I will stick to a simplified commonsense notion that serves as a working definition and assumes only so much as to sufficiently reflect the phenomena under consideration here.

Intuitively, two aspects that seem prototypical of emotions are often mentioned in commonsense definitions of emotions (Merriam-Webster Online Dictionary, 2020). One is that emotions involve some kind of bodily change or physiological response, such as fear involving an increase in heart rate or faster breathing. At the same time, emotions are often described as certain conscious experiences or feelings with a specific phenomenology, which can be difficult to describe intersubjectively. Fear often seems to involve a feeling of danger, for example. These two parts are also found in various accounts of emotions and will form the basis of this analysis (Prinz, 2004; Deonna & Teroni, 2012). Possible interactions between them and their necessity will be discussed later (section 3.4). I will call the involved physiological response an *implicit emotion* E_i and the respective conscious experience an *explicit emotion* E_e in the following.[21]

3.2.1) Implicit emotions (Ei)

Much of contemporary neuroscientific research dealing with emotions is concerned with behavioral change or physiological responses, both in theorizing as well as in experimentation (Damasio, 1994; Prinz, 2005). Often, emotions are (implicitly) reduced to or identified with these responses, either for pragmatic reasons to specify the research topic, or to further experimentation (Nader, Schafe, & Le Doux, 2000; Ramirez, et al., 2013; Arun, Kandel, & Rayman, 2019). Usually it is not denied that such behavioral or physiological responses are accompanied by, or can lead to, a conscious emotional experience, but such an experience is often not considered a necessary aspect of an emotion in these frameworks. This is reminiscent of the James-Lange theory of emotions, where something (the stimulus) is perceived which causes physiological responses which are then felt. This feeling of physiological responses is, according to the James-Lange theory, identical with the emotion (James, 1983a). However, as large parts of current neuroscientific research on emotions is carried out on non-human animals such as mice, and since we lack a reliable source informing us of the contents of their conscious on-goings, emotions are often not seen as the feeling of such physiological responses, but as the physiological responses themselves. This notion is further supported by cases where people do not have any kind of particular conscious emotional experience but still show physiological signs usually attributed to certain emotions (Winkielman, Berridge, & Wilbarger, 2005; Zajonc, 2000; Prinz, 2005).

An emotion defined as certain behavioral or physiological responses is what I term an implicit emotion (E_i). An E_i therefore does not need any kind of conscious experience or thought. What exactly constitutes a certain E_i as compared to another or compared to

[21] This is done in loose analogy to the distinction between *implicit memory* (defined as memory absent of conscious recollection) and *explicit memory* (defined as memory that requires conscious recollection) (Graf & Schacter, 1985), but similar distinctions can be found in the emotion literature as well, e.g. (Rüsch, et al., 2007).

another set of physiological responses not attributed to any emotion is a difficult question, but one that (neuro)scientists are trying to discern and answer.

3.2.2) Explicit emotions (E_e)

The philosophical landscape provides many different views to the question of what an emotion is (Deonna & Teroni, 2012; Goldie, 2009; Scarantino, 2016). For explicit emotions, I will mostly follow a feelings account, since it seems to account quite well for this part of the commonsense notion of what emotions are (but aspects of other accounts will be considered later on). Under this view, emotions are a kind of subjective conscious experience, sometimes also referred to as the phenomenological part of emotions (Deonna & Teroni, 2012). In contrast to implicit emotions as described above, explicit emotions are always a conscious experience. Therefore, there cannot be unconscious (or 'unfelt') explicit emotions under this definition.

In this sense, explicit emotions are similar to perceptual experiences. For example, having the visual experience of an ash-gray dog means that I consciously experience something which has the visual appearance of an ash-gray dog.[22] Generally speaking, I could have such a visual experience of an ash-gray dog by for example seeing a dog or visually hallucinating one. Similarly, since I define that such an experience has to be conscious, it is not enough to 'see' an ash-gray dog, for example by it being presented only for a fraction of a second (Zajonc, 1980), if I do not consciously experience it. There most likely is a gray area in which it is difficult to decide whether or not something was consciously perceived, but it is not my intention to sort out such atypical examples here. For the purpose of this investigation, it is enough to look at prototypical examples where one clearly does, or does not, consciously experience something.

I take explicit emotions to be much like these perceptual experiences just described (with some differences such as perceptions having dedicated sensory organs, while this does not seem to be the case for emotions). An emotion defined as certain subjective phenomenological experiences is what I term an explicit emotion (E_e). As with implicit emotions, I do not intend to sort out if a particularly atypical conscious experience should count as an emotion or not. Instead I stick to those subjective conscious experiences which are usually quite clearly recognized as emotions, such as fear (Fehr & Russel, 1984). In what way explicit and implicit are related to, and may depend on, one another will be elaborated later on (section 3.4).

3.2.3) Beyond implicit and explicit emotions

A variety of contemporary theories define emotions not simply through physiological responses or subjective conscious experiences, but also include evaluations and objects which emotions are about (Deonna & Teroni, 2012; Goldie, 2009; Scarantino, 2016). Under such views, emotions are directed at something, that is, they have an intentional object (i.e. an object which they are about) (Searle, 1983; de Sousa, 1987). In the case of fear for example, someone is usually fearful *of* something, such as being fearful of a dog, because it is evaluated as being dangerous. Even though these aspects are quite popular in the current literature on emotions, I will not make intentionality and evaluation a necessary condition of

[22] I do not take a stance here on the debate whether a concept of an ash-gray dog is needed to have the experience of an ash-gray dog, or any closely-related debates such as in (Siegel, 2016).

emotions here, primarily because it is often thought to exclude states such as moods, which do not seem to be about anything, and because such 'objectless emotions' seem to play an important role when it comes to anxiety or in psychiatric disorders that involve an interplay between memories and emotions (Arun, Kandel, & Rayman, 2019; LaBar, 2016). However, I do not negate such aspects, and intentionality and evaluation will play a role later on.

Downsides of excluding intentionality and evaluations from emotions include that motivational aspects and differentiation between some emotions will be lacking. Having or not having an intentional object can lead to different consequences concerning the actions one takes. Being afraid of a dog can lead me to thinking of ways to flee from that dog, but simply feeling anxious without knowing what is eliciting it puts me in a position that is a lot different. Similarly, some emotions we typically differentiate on cognitive grounds will ultimately be regarded as the same. Errol Bedford for example mentions that indignation and annoyance are not distinguished because they feel different (1956). Consequently, if only physiological responses and feelings are taken into account for emotions, some of the emotions we usually refer to by different names will be the regarded as the same. However, my primary concern here is not to try and distinguish closely related emotions or give a detailed taxonomy of emotions, but to analyze if emotions can be remembered. Thus, I take prototypical cases of emotions, such as fear, without trying to distinguish closely related emotions from another, such as fear, panic and angst. I leave it open to further distinguish closely-related emotions on other grounds such as their intentional objects or appraisal for future accounts (Arnold, 1960; de Sousa, 1987; Lazarus, 1991; Scherer & Moors, 2019).

Thus, for now I will only look at implicit emotions and explicit emotions as characterized above and show how each of them can be remembered, and later consider other aspects of emotions, such as evaluations, as well.

3.3) What does it mean to remember an emotion?

Implicit emotions and explicit emotions as described above differ in many aspects and are also used differently with regards to how essential each of them is to the concept of emotion depending on the field of study. Likewise, it seems advisable to consider different definitions of memory, not only regarding the different uses of the term between disciplines, but also with the different types of memory which are usually distinguished. Behavioral or physiological responses, in the popular memory taxonomies, come closest to phenomena which belong to *nondeclarative* memory, which is taken to be an umbrella term for those memories expressed through performance (Squire, 2004; Bernecker, 2011). On the other hand, subjective conscious experiences come closest to phenomena in the context of *episodic* or *experiential* memories (Teroni, 2017; Debus, 2017; Squire, 2004; Bernecker, 2011). Since these notions of emotions and the corresponding memory types are quite different, it seems reasonable to propose that they have different conditions for when someone remembers an implicit emotion compared to when someone remembers an explicit emotion.

3.3.1) Remembering implicit emotions E_i

Different scientific disciplines usually define memory or remembering differently. The traditional definition in biology (and related scientific fields) usually held that "memory is an imprinting of past experience" (Dudai, 2007a, p. 11). For experimental purposes, the

traditional definition often was (and still is) that (especially nondeclarative) memory is the change of (the potential for) behavior due to previous experience (Dudai, 1992). This imprinting approach has been questioned considerably in the last decades, leading to more and more scientists shifting to views that see memories as retention or reconstruction of representations which are formed through experience (Dudai, 1992; Dudai, 2007b). Not wanting to deal with defining the concept of memory further, contemporary experimental neuroscientific research often (implicitly) takes memories to be at least inferable from behavioral or physiological data. If a mouse solves the same maze faster and faster (compared to a control), it is usually concluded that this increase in task completion speed is due to memories the mouse has formed concerning the maze or task (Olton, 1979; Morris, 1981). Whether or not the actual memories amount to more than this change of behavior or are based on some kind of representation is then (implicitly) taken to not be of primary interest.

Since I am concerned with implicit emotions defined as certain behavioral or physiological responses, which are studied primarily in neuroscientific research, the standard definition of certain changes of behavior or physiology due to experience will suffice, although the kinds of experience will have to be limited to intuitive and prototypical cases.[23] To do this I will go through all conditions one at a time, which are judged to be individually necessary and jointly sufficient for the subject x to remember the implicit emotion E_i. For illustration I will take the example of the implicit emotion E_i of fear.

P1: x responds to stimulus S at time t(2) in a manner M that constitutes the implicit emotion E_i, given the appropriate circumstances

Given that memory is taken to be changed behavioral or physiological response due to experience, and implicit emotions are taken to be specific responses, it is necessary that there actually is a certain kind of response. This response can either be a behavioral response such as fidgeting or a physiological response such as heart rate increase, but can also be the lack thereof such as no movement at all or no increase in heart rate (which is especially important in extinction (Quirk & Mueller, 2008)).

For example, in the standard conditioning experiments a mouse (or typically a group of mice for statistical purposes) first hears a neutral tone which causes no particularly distinctive response. Then the mouse hears the tone again, but at the same time receives an aversive electrical shock. It is usually concluded that the mouse formed a memory, which is recalled, if the mouse responds in a specific manner which constitutes a fear response when hearing the tone again (such as showing almost complete immobility called *freezing* (Blanchard & Blanchard, 1969), among other things,) compared to a control group of mice which heard the tone, but did not receive an electrical shock and does not show such a fear response. The mouse (x) responds to hearing the tone (S) after having heard it (t(2)) by responding in a manner (M) that constitutes the implicit emotion of fear (E_i) (given the appropriate circumstances). A more relatable example would be if I (x) now (t(2)) respond in a way (M, for example by, among other things, breaking out in sweat) that constitutes the

[23] This is done to avoid common objections, such as that we would usually not call being shot in the foot and because of that limping a kind of memory (Dudai, 2007b; Hacker, 2013). Naturally, such an approach runs the risk of being circular by already assuming a definition of memory which enables us to restrict the kinds of experiences that count as relevant for memory and those that do not.

implicit emotion of fear (E_i) when seeing the dog that chased me (S) years ago (given the appropriate circumstances).

The 'appropriate circumstances' clause is introduced to counteract objections such as whether or not I would still be said to remember when I see the dog but it is too cold outside to sweat. What exactly falls under appropriate circumstances will be difficult to define and depends on the context. In many cases there probably is not a clear-cut distinction, but it seems reasonable to make these conditions not too wide nor too narrow, so that only relevant phenomena are captured. It does not seem to be reasonable to say that if I am deaf, blind and crippled I would still have to be expected to respond in a specific manner to the stimulus S (and interpret my lack of response as some kind of forgetting or extinction). At the same time, it seems plausible that the circumstances should be allowed to vary at least in some degree from the time stimulus S was experienced before.

P2: x has learned to respond in the manner M that constitutes the implicit emotion E_i to stimulus S at time t(2) given the appropriate circumstances at a time(span) t(1) which preceded t(2)

Both, the intuitive understanding of the word *remembering*, and the scientific idea of memory as stated above, seem to presuppose that there must have been some kind of experience at an earlier time.[24] There might be some responses to certain stimuli which are classified as implicit emotions, but which are not responses to previously encountered stimuli and therefore not memories. For example, hearing a lion roar for the first time in my life (as an infant) might lead to my showing physiological responses that are clearly classified as the implicit emotion of fear. But it does not seem to make sense to say that a one-year old who has never encountered a lion or anything similar before in his life remembers something specific about it. Remembering seems to presuppose that the changed behavioral or physiological response due to experience has been acquired in the past through learning. In the case of the lab moue (x) this would mean that it has heard the tone (S) at a previous time (t(1)) and because of the pairing of the tone with the electrical shock has learned to respond in a manner (M) that constitutes the implicit emotion of fear (E_i) when hearing the tone (S) (given the appropriate circumstances).

P3: P2 establishes a cause of P1

It is conceivable that a subject responds to a stimulus S in a manner M that constitutes an implicit emotion E_i without ever having encountered that stimulus, even though such a response to such a stimulus is usually acquired and not innate. However, to speak of remembering in the intuitive as well as traditional biological definition of memory described above, the change of behavior or physiology needs to causally depend on the learning experience. For experimental purposes this is often expressed in the way of a counterfactual condition such as *if not P2, then not P1*, i.e. had the subject not had the learning experience at t(1), it would not respond in the corresponding manner at t(2). This counterfactual condition often seems to be what is the motivation behind including control groups, i.e. in

[24] Some recent accounts of memory explicitly deny the necessity of the dependence on (particular) earlier experiences for mental states to count as memories however (Michaelian, 2016b; Michaelian & Robins, 2018).

the example above mice that might have heard the tone but did not receive the foot shock and therefore did not learn to respond to the tone in a specific manner.

Caveats

While such an analysis might seem rather straightforward and is often used in neuroscientific studies, it is not without its fair share of problems. For one, scientists are moving away from a simple definition of memory as change in (the potential of) response due to previous experience to other notions of memory. More and more findings are suggesting that there is no simple connection between stimulus and acquired response as often portrayed, since other factors such as context or body position seem to play an important role (Mendes, 2016; Dickinson, 2007). Furthermore, serious doubts have been voiced whether there actually are any stable patterns in physiological or behavioral response that could be ascribed to any specific emotion (Clark-Polner, Wager, Satpute, & Barrett, 2016). Lastly, as the following section on remembering explicit emotions shows, a simple counterfactual dependence as a realization of condition P3 mentioned above often seems insufficient to exclude cases outside the lab, where circumstances and experiences cannot be controlled for, and are much more diverse and interdependent.

3.3.2) Remembering explicit emotions E_e

As mentioned before, I take explicit emotions to be certain subjective experiences, usually called feelings. As will be elaborated later, such experiences can be triggered by different events. Hearing a lion's roar for the first time might lead to an intense feeling of fear. Likewise, remembering something can lead to feeling fearful. However, just because a current explicit emotion is triggered by the remembering of something does not establish that that explicit emotion was remembered. Dorothea Debus (2007), for example, argued that it is impossible to remember emotions, and that every emotion we have is a new emotion. I agree that not *every* emotion we have necessarily is a remembered one, but I think that it is nevertheless possible to remember emotions. Like with implicit emotions, I will take the example of feeling fear for illustration in the following, and go through the conditions one at a time which are judged to be individually necessary and jointly sufficient for the subject x to remember the explicit emotion E_e.

P1: x felt the explicit emotion E_e1 at event(1) at time t(1) which precedes t(2)

Similar to remembering perceptual experiences and implicit emotions, it seems to be the case that to remember explicit emotions, I must have experienced an explicit emotion at a previous point in time. If I want to remember the experience of seeing ash-gray for example, it seems necessary to have had the experience of seeing ash-gray at some previous point in time. To say that I remember what ash-gray actually looks like to me without ever having had any experience of the color ash-gray seems to be contradictory from an intuitive viewpoint. Likewise, if I try to remember the explicit fear emotion I had when confronted with the ash-gray dog that haunted my childhood, I need to have felt that fear in the past, otherwise it seems doubtful if 'remember' would be the right word.

P2: x feels the explicit emotion E_e2 at time t(2)

To say that I experientially remember the fur color of my neighbor's dog, presupposes that I now have some kind of (visual) experience. If I do not have any kind of (visual) experience, it does not seem to make much sense to say that I am *experientially* remembering something. I could be remembering something propositionally and remember *that* the dog was ash-gray. However, this would not be an experiential memory, but a propositional one, which anyone who knew that proposition in the past could remember. Similarly, to say that I experientially remember an explicit emotion, presupposes that I am now having some kind of emotional experience. Saying that I remember the fear I had as a child when that dog chased me but feeling nothing seems to be contradictory, or to refer to something else, such as propositional memory.

P3: E_e2 is sufficiently similar to E_e1

The concept of memory seems to presuppose at least some kind of identity, similarity or entailment relation. It strikes me as counterintuitive to say that I remember the visual experience of my neighbor's ash-gray dog, when I am now having the auditory experience of a frog's croak. These two experiences seem too dissimilar to warrant that one is a genuine memory of the other. Thus, for two experiences to be in an experiential memory relation they need to be at least type-identical (among other things). In the case of emotions, it seems false to say that I remember the fear I had of that dog, when I am now experiencing only joy. How similar two type-identical experiences need to be for them to count as memory-related, seems to be rather dependent on the actual context and speech community. When I am now having the visual experience of charcoal instead of ash-gray, some who do not distinguish those colors might say that they are sufficiently similar to count as the same, while in a suit tailoring class it might be considered as too dissimilar to be the same. Consequently, the two experiences, in this case explicit emotions, have to be at least sufficiently similar for one to count as a memory of the other, even if they might not be completely identical.

P4: If E_e1 were substantially different, E_e2 would be different in a way that P3 would still be true

Thinking back at the incident where my neighbor's dog broke loose and chased me, I could try and remember it as vividly as possible. In case I have not overcome my anxiety of dogs, it could also very well be that I would experience an explicit emotion E_e2^* that is sufficiently similar to the explicit emotion E_e1 I had felt when I was actually chased by the dog. However, this would not be enough to say that I actually remember the explicit emotion E_e1, because they could be sufficiently similar by coincidence. Even if I had completely forgotten the experience, including all experiential aspects, I could be told this story by a friend who witnessed it and develop the completely new emotion E_e2^* that just happens to be sufficiently similar to the one I had felt at the time of the event (cf. (Martin & Deutscher, 1966) for a similar example concerning memory in general).

There are other ways of coming up with experiences of explicit emotions that coincidentally are sufficiently similar, but the important thing is that it seems that the explicit emotion E_e2 should be determined by the explicit emotion E_e1 to count as memory-

related (cf. (Bernecker, 2010, pp. 128–154; Bernecker, 2017b) in the case of memory and more generally (Nozick, 1981) for so-called 'tracking' conditions). Had I felt joy back then, I should feel something sufficiently similar to that joy now. This also seems to be true for memories concerning perceptual experiences. If in my childhood I have had the visual experience of orange instead of ash-gray when seeing that dog, I should now 'see' orange when I remember the dog's fur color, and not ash-gray or something completely different from orange or ash-gray.

P5: P1 produces the structuring cause of P2

It is conceivable that P1 through P4 would be met, but we would be hesitant to call something an experiential memory. To rule out that P3 and P4 are not true simply by coincidence, it seems to require that P1 is instrumental in, or is causally related to, P2. Traditionally such connections have been established by appeal to so-called *memory traces*, which have been described with varying degrees of abstraction but generally seem to express 'that which makes past information/representations available or carries the causal connection' (Bernecker, 2010; Bernecker, 2017; Debus, 2017; Martin & Deutscher, 1966; Robins, 2017). Instead of an appeal to memory traces, I use the notion of structuring and triggering causes described by Fred Dretske (2010; 1988).[25] These two views might not be mutually exclusive and describe overlapping notions, but it seems that viewing the dependence between the past experience and the current state of remembering in terms of structuring and triggering causes presupposes fewer things which seem to be contingent (but reasonable), such as the reliance on neurons or neural networks, and at the same time allow for variability of memories depending on retrieval cues.

According to Dretske, a structuring cause is that which produces conditions under which certain triggering causes can produce certain events (2010; 1988). Applied to my cases of remembering, this can be described as follows. Something, the triggering cause, causes P2. But, according to P5, in the case of remembering, the triggering cause causes P2 only if the structuring cause is instantiated by P1. More concretely, thinking back (triggering cause) to the experience where the dog chased me causes me to feel an explicit emotion of fear (E_e2). However, intuitively we would only say that this is a case of remembering if my former emotional experience with the dog is in a certain way instrumental in this triggering of E_e2. The emotional experience with the dog at t(1), which established E_e1, is therefore a structuring cause of the later emotional experience of E_e2. Without the initial emotional experience, the triggering cause should not lead to E_e2, if the initial emotional experience is a structuring cause for my thinking back causing me to experience E_e2.

An advantage of this view is that it allows to not only give an answer to what happens how (thinking back to P1 leading to experience E_e2), but also why it happens. P2 could happen for a number of reasons, which are not cases of remembering. But viewing P1 as a structuring cause of P2, can explain why in the case of remembering P2 is happening. Yet, at the same time such a view is not restricted to cases which contingently depend on certain physiological aspects, such as specific neural architectures, and thus captures the broad concept of what remembering often is taken to be.

[25] I take it that an analogous account using memory traces instead of structuring and triggering causes could establish the causal connection just as well, or less abstractly, but would presuppose different things.

Caveats

The account proposed here for remembering explicit emotions is in many respects similar to more traditional accounts of memory, which are often termed *causal theories of memory* (Martin & Deutscher, 1966; Debus, 2017; Bernecker, 2017b). Yet, causal theories of memories have been somewhat questioned in recent years, which seems in large part to be due to issues concerned with the nature of memory traces, and constructivists account are on the rise, most notably in the form of reliabilist or simulationist alternatives, which take a reliably working memory system, and not the dependence on past representations, as the distinguishing feature of memories (Michaelian, 2016; Michaelian & Robins, 2018; Robins, 2017). Since the presented account above does not explicitly rely on memory traces, I suggest that it could be adapted to either a traditional causal theory of memory with memory traces, or a reliabilist variety which combines reliability and causality if one wanted to. Yet, such an endeavor would be far beyond the scope of this paper.

3.4) What is the relation between an event, an emotion and memory of that emotion?

So far I have considered implicit emotions and explicit emotions separately. However, usually both of these aspects are mentioned to occur or even be necessary for someone to have an emotion. Not showing any significant behavioral or physiological response but still claiming that someone is afraid, strikes many as missing a vital point about emotions. At the same time, it has been questioned whether or not certain physiological changes really correspond to only one emotion and it has been proposed that the distinguishing factor might be outside behavioral or physiological changes (Clark-Polner, Wager, Satpute, & Barrett, 2016; Bedford, 1956; Deonna & Teroni, 2012; Schachter & Singer, 1962).

3.4.1) Events can cause implicit and explicit emotions which can be remembered

As mentioned before, it is possible to have no conscious subjective emotional experience but still show certain physiological responses usually attributed to emotions. At the same time there is evidence that the same physiological changes can lead to conscious subjective experiences that are described as different emotions (Clark-Polner, Wager, Satpute, & Barrett, 2016; Schachter & Singer, 1962). To include such phenomena, I propose that an event can cause an implicit emotion as well as an explicit emotion, each of which can be remembered later on.

The view that a stimulus causes both certain physiological responses attributed to emotions and a conscious feeling is reminiscent of the so-called Cannon-Bard theory (Cannon, 1927; Bard, 1928). In contrast to it, however, I do not presuppose that physiological responses and conscious feelings are independent (and also do not imply the role of a specific brain region). While implicit emotions without explicit emotions do seem to occur, it seems questionable if one can only have a conscious emotional feeling without any kind of behavioral or physiological response usually attributed to emotions. Nevertheless, I leave this question open here as it does not seem to have an established answer yet.

3.4.2) Perception and cognition can be emotion-eliciting events

According to the James-Lange theory emotions are perceptions of bodily changes. Such bodily changes alone are described as implicit emotions here, and not taken to be the only thing that should be regarded as emotions. I take the perception of bodily or physiological changes as a separate event, that can again cause implicit or explicit emotions. If I see a huge dog, my heart rate might increase, and I might experience fear. However, it might also be the case that I do not really experience fear at first, but my heart rate increases, which I perceive leading me to worry and feel fear as a result. But the perception of my heart rate increase is a different event than the perception of the dog.

Similarly, I take that cognition can work in much the same way. Some emotions are caused by events that are complex and which presuppose some degree of cognition or judgment. Take complex emotions such as *schadenfreude* for example, which is the pleasure derived from becoming aware of someone else's misfortune. During my undergraduate studies I once saw a classmate who carefully peeled a mandarin for over ten minutes during class to remove all the peel and pith, only to accidentally drop it on the floor when he was finished, causing me to burst into laughter. Such an outburst on my side would generally not be given if I see someone drop a mandarin. What made the whole situation so amusing to me was the (in my opinion) excess amount of work he had put into peeling that mandarin which ultimately was wasted. But taking all this into account presupposes that I can think about the relatively complex situation that transpired.

Concerning cognitive involvement, this seems a lot different from hearing a lion's roar and as a result feeling intense fear, before even realizing what exactly occurred. This mirrors in some way the popular distinction between top-down and bottom-up processes often employed in psychology. However, both are events (perceiving and cognizing) that can elicit implicit and explicit emotions. I take remembering to be a form of cognizing, which can of course also elicit emotions, such as me thinking back to that mandarin incident and still having to smirk.

3.4.3) Evaluations can distinguish closely-related emotions and influence behavior

Evaluative theories of emotions usually pose that judgments or evaluations stand in a relation of necessity to emotions, such as being a part of emotions or being a cause of emotions (Deonna & Teroni, 2012; Nussbaum, 2001; Arnold, 1960; Lazarus, 1991; Scarantino, 2016). However, I think that such evaluations are either one form of emotion-eliciting events as described above, or one way of how we can distinguish closely-related types of emotions which might neither be physiologically different nor feel much different. To feel fear or show physiological responses usually attributed to fear can in some cases presupposes that some kind of judgment or evaluation has taken place, but it need not be so. As mentioned above, hearing a lion's roar and immediately afterwards having an increased heart rate as well as an intense feeling of fear, does not seem to necessitate that I think about what actually happened, it just happens.

As mentioned before, it sometimes is difficult to pinpoint whether and how closely-related emotions such as fear, panic and angst differ in physiological responses or regarding how they feel. Evaluations and intentional objects, on the other hand, can help distinguish these. Fear could be seen as a more general word for the emotion to be described. Panic

usually describes a more sudden outburst of fear, and angst a diffuse kind of fear not directed to anything in particular. Evaluations about factors such as the eliciting circumstances can lead to differentiation of emotions into types which are physiologically or phenomenologically indistinguishable.

Consequently, evaluations and intentionality can play an important role when it comes to emotions but are not necessary for the implicit and explicit emotions I set out to analyze here.

Conclusion

Here I have tried to sketch an initial account of what it could mean to remember emotions *per se* by combining insights from both science as well as philosophy. By first distinguishing between implicit and explicit emotions, a first step was taken to disentangle misunderstandings that sometimes arise in the debates between scientists and philosophers. Furthermore, such a distinction allows to relate implicit emotions to nondeclarative memories and explicit emotions to experiential memories, which helps incorporate both phenomena into the framework of memory research.

However, as was noted throughout the text, both the definition of emotion as well as memory in scientific research are subject of lively debates, which will have to be clarified in the near future to enable fruitful collaborations between the disciplines. Similarly, while I have opted for a largely causal account of memory, a reliabilist alternative with its focus on empirical findings might make for an interesting addition.

Since I have based this analysis largely on a natural language understanding of emotions, and restricted it to implicit and explicit emotions, aspects such as intentionality, motivation or appraisal have not factored in the direct account of remembering emotions, but did play a role in the how emotions or memories could be initiated. Yet, I hope that this analysis can be the starting point for future accounts of what it means to remember emotions that can cover such aspects as well.

3.5) Falsely Remembering Emotions

False memories have rarely been considered when it comes to emotions in contemporary philosophy. Using the characterizations given above of what it means to remember implicit and explicit emotions, I will now offer a brief suggestion of what, if anything, it might mean to falsely remember emotions.

3.5.1) Falsely Remembering Implicit Emotions

There are a number of cases where people seem to show dysfunctional emotional responses. Many of these are thought to be connected to previous experiences the affected individual had, but here I will focus only on post-traumatic stress disorder (PTSD). PTSD is a stress disorder which can develop after a traumatic incident such as exposure to threat to life or serious injury (American Psychiatric Association, 2013, p. 271). Among the symptoms typically found in patients suffering from PTSD are unwanted and upsetting memories, uncontrollable flashbacks and emotional distress when exposed to reminders of the traumatic experience (American Psychiatric Association, 2013). Concretely, PTSD can

manifest itself in such examples as war veterans showing panic attacks in non-dangerous and seemingly random contexts.

I have characterized the remembering of an implicit emotion by first acquiring the showing of a certain behavioral or physiological response (the implicit emotion) towards a stimulus due to the experience with that stimulus. Importantly, in the case of genuine remembering, this acquisition needs to be instrumental in the display of that behavioral or physiological response after re-exposure to that (or a sufficiently similar) stimulus. In a pathological case such as PTSD, however, affected individuals often respond to a stimulus by showing an implicit emotion, while any possible previous experience with that stimulus is not instrumental in such a response, even though such a response to that stimulus is usually acquired and not innate. Therefore, the causal condition (P3) does not hold (or at least not in the same way) in many pathological cases. In the case of PTSD, this seems to be the case in people who in neutral settings show behavioral or physiological response patterns (implicit emotions), which are due to a traumatic experience. A war veteran who walks down a peaceful street might start to show a response pattern such as being overcome by a rush of adrenaline, increased heart rate and heavy breathing, which is reminiscent of a response they had acquired during a harrowing war experience. It seems likely that this is a result of maladaptive fear generalization, where response to a conditioned stimulus (such as seeing a helicopter fly over one's head) is generalized to other normally neutral stimuli (such as the rotating spokes in a bike wheel) in a detrimental way (Arun, Kandel, & Rayman, 2019). In this manner, it could be argued that the 'false' part in falsely remembering emotions should express maladaptation, rather than simply generalization of responses to other stimuli.

Following the more general definition, falsely remembering an implicit emotion would occur when an individual responds to a neutral stimulus in a specific way, while any possible previous experience with that stimulus is not instrumental in such a response. Crudely put, it is the showing of a certain response towards the 'wrong' stimulus, although that stimulus does not innately trigger such a response. Consequently, this kind of 'false memory' could be seen as an incorrect association, but similar cases meeting the criteria characterized above have been called false memories as well (Ramirez, et al., 2013).

3.5.2) Falsely Remembering Explicit Emotions

Usually, the term false memory is used to refer to cases where someone seems to remember something which never occurred (in the way seemingly remembered). In the case of remembering explicit emotions as described above, there seems to be no room for such a kind of falsity. Either a previously felt explicit emotion is the structuring cause for a currently felt, sufficiently similar, explicit emotion, in which case it would be a case of genuine remembering, or it is not, in which case it is not a case of remembering. Trying to characterize false memory as not fulfilling the sufficient similarity condition seems questionable as well, since it seems highly counterintuitive to claim that a substantially different emotion is a memory, whether genuine or false, of a past emotion (which is not to say that the two could not be causally related).

Some definitions of false memory add another aspect, namely that the subject who seemingly remembers takes the false memory to be true. While this view has been shown to be problematic (Bernecker, 2017b; Michaelian, 2020), one might argue that it might fit the context of remembering explicit emotions. In case of remembering explicit emotions, I have

noted that a currently felt emotion could be sufficiently similar to a past explicit emotion by coincidence. Feeling fear now when imaging that ash-gray dog is not a guarantee that I am remembering the explicit emotion of fear I felt when that dog chased me. However, I could believe that I feel fear now because I felt fear back then, even though there is no such dependence. That is, it could be argued that someone falsely remembers an explicit emotion if they take their current emotion to be due to the past emotion (i.e. being a memory of the past emotion) even though the current emotion does not in fact depend on the past emotion in a relevant way. However, as Debus (2007) has noted, generally it seems doubtful whether people take emotions they currently feel to be due to, or informative of, emotions felt in the past. If I think back to the situation where I was chased by the dog, I usually do not take it that the fear I feel now is due to the fear I felt back then (even if it might be).

It seems unclear if a simulation-theoretic version would fare better. If instead of being causally dependent on the past, you were to take use of reliable processes as the distinguishing criterion between memories and false memories, it still seems counterintuitive to speak of false memory in the case of emotions. The main issue seems to be epistemic rather than ontological in nature concerning the relation to the past. In the case of episodic memories, one might argue that the employed mental processes aim at creating a simulation of a personal past event. Likewise, it could be argued that in the case of remembering emotions, what the processes aim at is the simulation of the past subjective experience, i.e. the past explicit emotion. However, as mentioned before, we usually do not make judgments about whether currently felt emotions are memories of past felt emotions. If there are processes which simulate past explicit emotions, you might argue that you would genuinely remember if the processes used were reliable (and the simulation were accurate), and falsely remember if the processes used were unreliable. However, it seems questionable how you would know that your currently felt emotion is a simulation of a past emotion constructed by processes of your episodic construction system (and not just a completely new emotion), without making reference to any type of causality to the particular past event. Thus, calling such an instance a false memory seems rather counterintuitive.

Currently, it seems unclear how exactly current theories of false memories would deal with what I have described as remembering emotions. The main point seems to be that in the case of remembering emotions, we usually do not take ourselves to remember emotions, even if what I call remembering emotions fulfill the same hallmarks as, for example, remembering visual experiences. An interesting approach might be the further development and differentiation of the notion of *seeming to remember* which according to Robins (2020, p. 122) “occurs when a person has an occurrent mental representation, the content of which targets a representation in her personal past”. However, it seems advisable to leave such an application for a future endeavor when a full account of remembering all components of emotions has been developed and it becomes more clear what kind of phenomenon or phenomena emotion actually is.

4) Are You Morally Responsible for the Veracity of Your Memories?

Memory research in the past few decades has shown that memories can change because of a range of different reasons. This can lead to memories no longer standing in the alleged relation to their purported content, that is, memories not being veracious anymore. At the same time, it seems quite intuitive that memory and moral responsibility can interact, and that you can in some cases ensure the veracity of your memories. Thus, in this paper I will answer the question if and under what conditions you are morally responsible for the veracity of your memories. I argue that you do not have a moral responsibility to ensure the veracity of your memories *per se*, that is, if you are remembering only for the sake of remembering. But it can be the case that due to your moral responsibility for something other than the veracity of your memory, you become morally responsible for ensuring the veracity of certain memories. Put differently, you are morally responsible to ensure the veracity of your memories only if remembering veraciously is a potential way to act morally responsible. I distinguish between two different ways in which the veracity of your memories can stand towards moral responsibility, a necessary way and a non-necessary way, which differ in whether acting morally responsible necessarily presupposes remembering veraciously or not. I conclude the paper by pointing out future endeavors which can arise from this account, such as moral responsibility in relation to other forms of memory.

Introduction

When it comes to the ethics of remembering, questions in contemporary philosophy usually revolve around such issues as whether and in what circumstances you have a moral duty to remember (Blustein, 2017), a right to actively alter or delete your memories (Kolber, 2006; Liao, 2017), or the interplay between memory and other issues such as consent (Craver & Rosenbaum, 2018) or forgetting (Bernecker, 2018). At the same time, philosophical theories of memory are starting to incorporate so-called *false memories* or *mnemonic confabulations* into their overall frameworks.[26] Often, a false memory is understood as the memory of something that did not happen (Pezdek & Lam, 2007) or that did not happen in the way remembered (Dalla Barba, 2002).[27] While the notion of false memory is making its way into current philosophical theories of memory, the intersection between the ethics of remembering and false memories has scarcely been investigated in contemporary philosophical analysis. Yet, as I argue here, since there is more and more research suggesting that many of your memories are susceptible to substantial change over time, concerning moral responsibility, it is not only important whether or not you remember or forget, but

[26] To answer this question in a manageable way, in this paper I will mainly look at declarative individual memory (Squire, 2004) and not sub-divide memory types further, leaving questions concerning the intriguing interactions between moral responsibility, veracity and collective memory (specifically, whether or not *we* are morally responsible for the veracity of *our* memories) for a future endeavor.

[27] Recently, objections have been voiced against such definitions of false memory, which (among other things) state that confabulations can also be (entirely) veridical (Bernecker, 2017b; Michaelian, 2016b; Robins, 2020).

also whether what you seem to remember stands in the alleged relation to its purported content, that is, whether your memories are veracious.[28]

Outline and Objectives

In this paper I analyze if and under what conditions you are morally responsible for the veracity of your memories. To answer this question, I first delineate in what way false memories are different from simply not having memories. Having set the stage, I take a look at memories 'in isolation' and argue that you are not morally responsible for ensuring the veracity of your memories *per se*. However, it can be the case that, due to your moral responsibility towards something other than ensuring the veracity of the memory in question, you become morally responsible for ensuring the veracity of certain memories. I conclude the paper by pointing out future endeavors which can arise from this account.

4.1) No Memory, False Memory and Moral Responsibility

In this section, I will narrow down and specify the question I am trying to answer by distinguishing the issue of veraciously remembering from other closely related topics such as simply not remembering, and by clarifying the terms *veracious memory* and *morally responsible*.

4.1.1) Why the distinction between 'no memory' and 'false memory' can be morally relevant

Imagine you are in a town you have never been before only to find that you got lost. You happen to see an elderly man and ask for directions to the nearest train station. He kindly tells you to simply go straight and take the first road on the left. "Only takes about 15 minutes", he says. Hurriedly you follow his directions. You walk for 15, 20, 25, 30, 35 minutes, but there is no train station in sight. At some point you realize that the elderly man must have given you wrong directions. Thus, you make your way back and ask an elderly woman this time who tells you to go straight and take the next road on the right. This time you successfully (and finally) make your way to the train station by following her directions.

There are several possible reasons as to why the elderly man gave you wrong directions. Suppose that he has been to the train station a couple of times and in fact believes that he remembers the way. In this case, he might simply find it enjoyable to prank foreigners, and intentionally gave you what he believed to be the wrong way. Or, he might have actually believed that this was the way but was mistaken. As a third option, it is possible that he might have never been to the train station, had never even heard of a nearby train station and as a consequence believed that he did not remember, but too stubborn to admit, he just guessed, incorrectly in this case.

While the result of what the elderly man says in each case just mentioned would be the same for my not finding the train station, morally speaking we might intuitively make a distinction between someone who believes themselves to remember something (even though they do not) and claiming to do so, and someone who believes themselves to not remember something but still claiming to remember. Concretely in the case above, we might

[28] Since a more thorough explanation of what it means for memories to be veracious will take a bit of space, such an explanation will follow at the end of the next section (see section 4.1.3).

say that if the elderly man genuinely believed that he was actually remembering but was mistaken (disregarding for now possible circumstances such as whether or not he knew himself to easily mix things up), it seems like an honest mistake for which he might be exculpated. But, believing that he does not remember and simply guessing, thereby pretending to remember, seems at least morally questionable, because, on the one hand, it is deceptive to pretend to remember in this case, and, on the other, reckless to make such an unjustified claim. This distinction, as I argue, is important for the ethics of remembering but seems to have been somewhat disregarded in contemporary philosophy of memory.

4.1.2) Falsely remembering vs. not remembering

Increasingly more research in the last few decades has been suggesting that our memories are malleable to substantial change, up to the point where entire events are fabricated and allegedly remembered (Loftus, 2005; Loftus & Pickrell, 1995). The idea that memories are not immutable has made its way into philosophy as well, where different strands of memory theories have picked up on it (Bernecker, 2017b; Michaelian, 2020; Robins, 2020). False memories are often not considered to be memories at all and therefore 'recalling' them is not seen as remembering by some philosophers. Yet, as I show here, even if one accepts that 'recalling' false memories is not genuine remembering, including false memories in an analysis of the ethics of remembering is still crucial. Since false memories are often not seen as memories, current memory theorists might view the question of whether we are morally responsible for the veracity of memories as more or less identical with the question of whether we are morally responsible for remembering or not. However, the difference between simply not remembering and falsely remembering can be important concerning moral responsibility.

Not remembering can be realized in different ways depending on the attitude you might have towards your mental states.[29] I) One type of not remembering is the absence of remembering or even seeming to remember the thing in question ('no memory') which might be accompanied by the attitude of believing that you do not remember the thing in question. For example, you could say that a distant relative of yours does not remember what I had for breakfast ten years ago. This can express two things. First, it might express that the relative does not have any relevant kind of attitude such as belief or knowledge towards it, given that they do not know me and, for example, have never even considered the question. Second, it might express that the relative has considered the question but concluded that they do not remember what it was that I had for breakfast then. Thus, in the 'no memory' sense of not remembering I either lack any relevant attitude such as belief or knowledge towards the respective content, or I believe that I do not remember the content in question.

II) Apart from this 'no memory' sense of not remembering, I might seem to remember something but in fact what I am seemingly remembering does not reflect that claimed by the alleged memory ('false memory'). Thus, I might (tacitly) believe that I am veraciously remembering when in fact I am not. For example, I might seem to remember

[29] It might be helpful to highlight a distinction here. If you believe that you are remembering that p, it is possible to distinguish two attitudes: the attitude of remembering that p; and the attitude of believing that you are remembering that p. For a more comprehensive investigation of memories and attitudes cf. (Bernecker, 2010, pp. 231-239).

that my wife's birthday is in May, when in fact it is in April. In contrast to the 'no memory' sense of not remembering, in this second sense ('false memory') the (tacit) belief that I am remembering (even though I am not) is added.[30] Another way of expressing this idea is to say that in the 'false memory' sense of not remembering I (tacitly) take myself to have the attitude of remembering towards a certain purported content, when in fact I really have a different attitude such as imagining towards the purported content. Therefore, while false memories are neatly defined by different contemporary philosophical theories without reference to any (tacit) belief, I will mainly use the term in the intuitive sense of a 'memory' of something that did not happen (in the way seemingly remembered) or that is false, but which is usually (tacitly) taken to actually be a memory.

4.1.3) When are memories veracious?

Generally, I propose that a memory is veracious if it stands in the alleged relation to its purported content. This characterization of veracious memory is (kept) abstract to allow for different views of what it means to remember. In most cases this will simply reduce to a memory being factual (corresponding to objective reality) or authentic (corresponding to the initial representation) or both (cf. (Bernecker, 2010; Bernecker, 2017a) for use of these term in context of the memory debate), depending on what is being alleged. Since this use of *veracious* in the contemporary philosophy of memory is rather novel, I will offer two ways of looking at it in the following to help explain what I mean.

Concerning the idea of *purported content*, consider the example that I might seem to remember my divorce, in which case something like my divorce, the experience of my divorce, or a representation of one or both of those two is the purported content of my memory. If it turns out that I (luckily) never got divorced, my memory would not be veracious since it cannot stand in the alleged relation to its content, because there is no content to stand in relation to. Another way of looking at it would be to employ a recently made distinction by Sarah Robins (2020) between *target* and *content* in the philosophy of memory. Crudely put, if we take memories to represent propositions or events, then the target of the memory is that which you take yourself to represent, while the content is that which it in fact represents. What I call *purported content* is what Robins (2020) calls *target*. Thus, another way to express what it means for a memory to be veracious would be to say that it stands in the alleged relation to its target. If I take myself to remember my divorce, then the target is my divorce. However, in the case that I never got married in the first place, the content of my seeming to remember cannot be identical with the target, but might instead be a divorce I have heard or read about and am falsely attributing to myself.

Concerning the idea of *alleged relation*, assume, for example, that memories are taken to represent certain factual states. In that case an alleged memory is veracious only if it actually represents the corresponding factual states. For example, I veraciously remember that my wife was born in April, if (among other things) she indeed was born in April and my memory represents this. If, instead, memories are taken to involve a reexperiencing of the past, then your alleged memory is veracious only if it involves such a reexperiencing of the past.

[30] NB: a similar distinction, which might be of interest to the reader, is drawn when it comes to types of ignorance in the literature on moral responsibility, cf. for example (Peels, 2014).

What this implies is that, when it comes to veracity, the primary concern is not necessarily (but it can be) whether a memory is strictly speaking a mnemonic confabulation as defined by current memory theories. If it is taken that memory implies, for example, that your current representation counterfactually depends on a particular past representation (Bernecker, 2017b), or that your current representation is the result of a reliable process for constructing past representations (Michaelian, 2020), then these are important concerns because they constitute, at least in part, what memories are taken to imply in each of these cases. Given the definition of veracious memory above, I leave it open whether one expresses any of the two views mentioned or a completely different view, but whether an alleged memory is veracious will depend on the view taken. In most contexts and cases, however, it is usually clear what is meant when someone takes themselves or others to remember something, and the usual definition of a false memory will suffice.

For the purpose of this paper, i.e. answering whether you are morally responsible for the veracity of your memories, details such as whether my memory must correspond to both the facts in the world as well as my initial perception when I acquired the memory (Bernecker, 2017b), will in most cases not play a substantial role. Matters such as these can indeed be crucial when it comes to moral responsibility and memory, but, to avoid straying too far from the main topic, they will only be brought up explicitly if they make a difference concerning moral responsibility.

4.1.4) Moral responsibility

The debate of what moral responsibility is and when you are morally responsible in general is long and it seems like it will not be resolved soon. Likewise, I will not try and resolve it here. Instead, I will choose sides and pick the traditional account of moral responsibility, since it has been considered to be a standard for quite a while, and since it overall seems rather intuitive (Talbert, 2019). However, I will briefly describe another approach (attributionism) afterwards which might be particularly interesting in the context of veracity of memories.

The traditional view of moral responsibility holds that for you to be morally responsible you have to be a moral agent (Uniacke, 2010). Moral agency in turn is thought to presuppose two aspects about the moral agent, control and awareness (Talbert, 2019; Rudy-Hiller, 2018). The control condition often relied heavily on debates about free will, and is usually taken to be fulfilled if you could have done otherwise or if what you did was in accordance with your reasons, desires etc., but details are more complex (Talbert, 2019; Fischer & Ravizza, 1998). Since the control condition is not crucially important here, I will not elaborate on it further.

The awareness condition demands that you are in possession of adequate cognitive capabilities and epistemic states concerning your actions for you to be morally responsible (Rudy-Hiller, 2018). Put oversimply, this means that you need to be aware of what you are doing in order for you to be morally responsible for what you are doing. What exactly you need to be aware of is a controversial question, but here I will focus on only two aspect, because they seem to be the ones that are important later on, namely awareness of action and awareness of consequences of actions (Rudy-Hiller, 2018). It is often argued (leaving details and atypical cases aside) that awareness of consequence requires, at least to some degree, reasonable foreseeability (Uniacke, 2010; Rudy-Hiller, 2018). If you hand someone a

knife and they proceed to kill others with it, whether or not you are morally responsible for the ensued deaths depends on what you thought you were doing and whether you could have reasonably foreseen the consequences of your action. If you (through no fault of your own) thought you were handing them a spoon or if you knew the person you handed the knife to to be a balance-minded individual, you might not be morally responsible, since it does not seem reasonably foreseeable that they would go on to kill others with the knife. If, however, they in rage proclaimed they wanted a knife to kill others, it seems likely that you were aware of your action and the likely consequences of your action were reasonably foreseeable. Thus, you might justifiably be judged to be morally responsible for the deaths that ensued (Uniacke, 2010).

In contrast to the traditional view are accounts belonging to *attributionism*, which, crudely put, claims that you are morally responsible for your actions if they are the expression of your underlying attitude (Levy, 2005; Sher, 2009; Talbert, 2013; Watson, 1996). Control and awareness conditions as such are not necessary for moral responsibility according to attributionists, since what is important is that your actions are an expression or reflection of your attitudes. For example, if my wife forgets my birthday, I might hold her responsible for forgetting my birthday, if this is a reflection of her underlying disregard for me. And this could be true even if it might be the case that she was neither in control of, nor aware of, the fact that she was forgetting my birthday (Bernecker, 2018). Thus, attributionists could argue that neither the control, nor the awareness condition are necessary for moral responsibility. I will not dwell on attributionism further, since I assume the traditional account of moral responsibility in this paper, but I will mention it when discussing the necessity of reasons to doubt (see section 4.2.3) because it might provide an alternative take on the matter.

4.2) Veracity of Memories and Moral Responsibility

In this section I argue that you do not have a moral responsibility to ensure the veracity of your memories if you are, or anticipate that you will be, remembering veraciously only for the sake of remembering veraciously. However, if you have a moral responsibility towards something else, it can be the case that you have a derived moral responsibility to ensure the veracity of your memories.

4.2.1) No Moral Responsibility: Remembering Veraciously for the Sake of Remembering Veraciously

Before delving into situations where remembering is considered in interaction with other acts, it is worth looking at whether you might have a moral responsibility to ensure the veracity of your memories *per se*. By this I mean whether you have a moral responsibility to ensure the veracity of your memories if you are, or anticipate will be, only remembering veraciously for the sake of remembering veraciously. For example, imagine in my old age, many years after my wife passed on, I for no other reason than wanting to remember my wedding veraciously, try to remember the color of my wife's wedding dress, and later compare it with photos we had taken. Let's assume for a moment that you do have such a moral responsibility. It seems to me that this would imply (excluding possible exculpating factors) that you would have a moral responsibility to ensure the veracity of *every* memory you have, since there would not be any circumstantial factor that could make possible a

differentiation between moral responsibility for different memories. However, the claim that you have a moral responsibility to ensure the veracity of *every* of your memories seems wrong for at least two reasons.

First, if we assume that if you ought to do something it implies that you can do it, it seems contradictory to say that you have a moral responsibility to ensure the veracity of every one of your memories, because this is something beyond your ability. Contemporary cognitive science paints a clear picture when it comes to the question of whether memories change over time. In contrast to the previously held belief that memory serves as a device to exactly reproduce the past, current scientific evidence suggests that memories are frequently changed when remembering (Laney & Loftus, 2013; Spear, 2007). If this is the case, then you cannot ensure the veracity of every single one of your countless memories, and consequently you cannot have the moral responsibility to do so.

Second, even if we assumed that you could overcome the natural boundaries of your mental faculties and through enough effort ensure a kind of 'super memory' which is always veracious, in cases in which the veracity of a memory is completely irrelevant to anything else, it is difficult to see why it would enter the realm of morality in the first place, i.e. where the moral responsibility lies. An example might illustrate this point. When I was 14, I went mountain climbing and after two days of hiking and climbing finally made it to the top of the mountain. For reasons unbeknownst to me, I noticed three small rocks there which happened to lay apart in a way so as to form a more or less equilateral triangle (which has neither then, nor now, struck me as relevant for anything in particular). For an even more obscure reason, I remember this to this day. However, apart from writing it here, I cannot think of any way this memory has had any influence on anything in my life or for anyone I interacted with, and had I not written it here, it is reasonable to assume that it would not have had any influence in the future. It seems to make no difference whatsoever if, instead of three rocks aligned in a triangle, there were actually four rocks aligned in a rectangle. Likewise, concerning moral responsibility, it seems to make no difference whatsoever if my memory is veracious or not. If there is no reason why the veracity of my memory of the three rocks is, or will be, of even the slightest significance for any other action, it is quite natural to say that the matter is amoral, that is, it does not even enter the moral sphere.

Objection: Potential future use of memories

One could argue that it might be, because of some unlikely or unforeseeable reason, that the veracity of the memory becomes morally relevant in the future. It might be that the way the rocks lay there was a clue left by some former mountain climber which in the end leads to solving a murder case (and let's assume that I have a moral responsibility to help solve such a case if I can). It could be claimed that since theoretically every memory has the potential to be of such use in a morally relevant situation, you have a moral responsibility to ensure the veracity of your memories. While there might always be an infinitesimal small chance that any memory might be of use in a morally relevant situation, you would then not be remembering veraciously for the sake of remembering veraciously anymore, but for something else that is morally relevant, namely in the example above solving a murder case. Remembering veraciously for the sake of remembering veraciously would have to be evaluated independently of any potential or actual use it had, has or will have for anything other than the veraciously remembering itself.

Objection: Veracity of memories per se only for some memories

It could also be objected that a moral responsibility to ensure the veracity of your memories *per se* does not imply that you have a moral responsibility to uphold the veracity of *every* of your memories, but only for certain memories. However, it seems questionable on what ground to distinguish morally relevant memories from others if circumstantial factors are excluded. I am not claiming that you should never ensure the veracity of your memories. In cases where you are remembering for something else, it can be the case that you do have such a moral responsibility (as I elaborate in the next subsection). But, in such cases you are ensuring the veracity of your memories not (only) for remembering veraciously itself but for something else.

4.2.2) Derived Moral Responsibility: Remembering for the Sake of Something Else

While you might not have a moral responsibility to ensure the veracity of your memories when you are remembering veraciously only for the sake of remembering veraciously, from everyday life we certainly are familiar with situations where we ascribe moral responsibility to others or ourselves concerning the veracity of memories. For example, I might mistakenly take myself to remember when my wife's birthday is and surprise her with a gift, only to later find out how far off I was and how hurtful the implications of my falsely remembering are. Since intuitively it seems that you do have a moral responsibility to ensure the veracity of your memories in at least some cases, one might ask where this moral responsibility comes from if it does not come from the remembering veraciously itself.

The remainder of this paper will explore details of why and when exactly you have a moral responsibility to ensure the veracity of some of your memories. For this, I distinguish between two different ways in which the veracity of your memories can stand towards moral responsibility, a necessary one and a non-necessary one.

Derived moral responsibility: relation of necessity

One way in which you can be morally responsible for the veracity of your memories is if you explicitly consent to remembering veraciously. The prototypical cases are constituted by promises in the form of promising to remember veraciously. If I promise my dying friend to remember the kind of man he was, then I explicitly consent to remembering veraciously.[31] Keeping such a promise necessarily presupposes that I remember veraciously, for example by remembering that he was a computer scientist, not a biologist. If we accept that I have a moral responsibility to keep this promise, and keeping it necessarily presupposes that I remember veraciously, then I also have a derived moral responsibility to remember veraciously.

If, additionally, I am or should be aware of reason to doubt the veracity of my memories, then I also have a derived moral responsibility to ensure the veracity of those memories. Reason to doubt is necessary in the account presented here, but it can be realized in a rather general manner. Details about reason to doubt will be elaborated later on (see section 4.2.3).

[31] I use this example for illustrative purposes, and simply assume that 'graveside' or 'deathbed' promises are binding. Cf. (Albrecht, 2018; Dressel, 2014) for some discussion on the matter.

This case of derived moral responsibility is an example of one way in which you can become morally responsible for ensuring the veracity of your memories where the only way to act morally responsible is by remembering veraciously. In the case above, the only way to act morally responsible (to fulfill the promise) is by remembering veraciously (if we grant that 'remembering the man he was' implies remembering veraciously). If I have reason to doubt the veracity of my memories, for me to act morally responsible in such cases necessarily presupposes that I ensure the veracity of the corresponding memories. Put in a more formal sense, one might say that

Derived moral responsibility – Necessary cases
subject S is morally responsible for ensuring the veracity of a memory M, if
i) S is morally responsible for ϕ-ing
ii) the only way for S to ϕ in a morally responsible way is by remembering the contents of M veraciously
iii) S is, or should be, aware of reason to doubt the veracity of M
iv) there are no exculpating factors present.

Derived moral responsibility: relation of non-necessity
Apart from cases where the only way to act morally responsible is to ensure the veracity of your memories, there are other ways in which you can be morally responsible for the veracity of your memories. These other ways differ in that ensuring the veracity of your memories is generally not necessary for acting morally responsible but can be necessary under certain circumstances.

Why ensuring the veracity of memories can be non-necessary
Consider another form of promise, where one promises something other than remembering veraciously. For example, after failing to show up at a previous rendezvous, I promise my wife this time I will meet her at 9:00 on Sunday at a certain restaurant. If I show up there at 10:00 on Sunday convinced that I remembered time and place veraciously, I can expect to justifiably be blamed for failing to keep my promise (excluding possible exculpating circumstances). However, just showing up at 9:00 on Sunday exactly at that restaurant would not be enough to keep that promise in a morally relevant way.[32] It might be the case that I thought we agreed to meet on Monday, not Sunday, but since I was nearby and felt a slight rumble in my stomach, I decided to go eat there. As luck would have it, it is around 9:00 on Sunday and I meet my wife there. While I did show up at the right time and place, I did not keep my promise in a morally relevant way. It was only through luck that I did not stand her up (but see the debate surrounding moral luck (Nelkin, 2019)).

One quite natural way to keep such promises is by veraciously remembering what was promised and acting on it. If this were the only way to keep such a promise, the case would be of the same type as in the previous section where I promised to remember veraciously. However, there are other ways to keep that promise. For example, being aware that I failed to show up last time, I might, immediately after I promised to meet my wife,

[32] As with the deathbed promises mentioned above, I do not have the space to argue why more is needed, and doing so seems to stray too far from the main topic. I use this example mainly for illustrative purposes. If it is found unconvincing it can be exchanged for any other fitting example of a case in which one is morally responsible. Thus, I simply assume that in such a case the promise would not be kept in a morally relevant way.

write down the exact date, time and place on a to-do-list. Having established a habit of simply going through the to-do-list, I would no longer have to remember veraciously what I promised. I would simply have to go through each point, one of which would be the rendezvous with my wife. While keeping the promise in a morally relevant way presupposes that I do not just out of luck do what I promised, it does not necessarily presuppose that I have to remember veraciously what I promised. Since there are other ways of keeping that promise, ensuring the veracity of my corresponding memory is not necessary.[33]

Moral responsibility and alternatives

For the sake of simplicity, let's assume that in the case of the promise for the rendezvous, there are only two ways of keeping the promise, one characterized by remembering veraciously, and the other by writing things down on a to-do list. This case can be seen as analogous to the necessary case described in the previous section. If I am morally responsible for keeping the promise, and the only way to keep the promise is by remembering veraciously or writing things down, then I am morally responsible for remembering veraciously or writing things down (or doing both). Therefore, I am not necessarily morally responsible for ensuring the veracity of the corresponding memories. However, I can be morally responsible for remembering veraciously and in turn morally responsible for ensuring the veracity of my memories under certain circumstances.

If in principle there are multiple ways of acting morally responsible, but there are good reasons why remembering veraciously should be preferred, then I can be morally responsible for remembering veraciously. For example, while writing what I promised my wife down on a list and then just following one thing after the other on the list might be enough to keep my promise, it seems doubtful whether this is generally adequate. If I do this for everything concerning my wife, she might take it as an expression of disregard that I do not bother to remember at least some of the things concerning her. In some cases, you are expected that you achieve some things by remembering veraciously, even if there are alternative ways imaginable.

Even if all possible ways of acting morally responsible in a given situation are of 'equal value', we can still be morally responsible for a particular one. This can be the case when one intends to use that particular way and no other way to act morally responsible. Imagine that I am aware that I could either write down the date and place and then follow the to-do-list, or remember when and where I should meet my wife in order to keep the promise. Yet, I choose not to write it down and instead rely on my memory. While there is a possible alternative way to act morally responsible, I opted for the way that presupposes that I remember veraciously. Thus, for as long as I do not intend to employ any other way of equal value, I am morally responsible for remembering veraciously.

Derived moral responsibility in non-necessary cases

Consequently, if there are multiple ways of keeping the promise and one of the ways is characterized by remembering veraciously, I can be morally responsible for remembering veraciously if no other way of acting responsibly is intended to be realized or all other

[33] Cf. John Mackie's idea of an insufficient but necessary part of an unnecessary but sufficient (INUS) condition (1965). Ensuring the veracity of the corresponding memory, while in itself insufficient for keeping the promise, is a necessary part of one possible way which is sufficient for keeping the promise.

possible ways are disfavored. Additionally, if I have reason to doubt the veracity of my memories, I can be morally responsible to ensure the veracity of the corresponding memories. Put in a more general way, one might say that

> **Derived necessary moral responsibility – Non-necessary cases**
> subject S is morally responsible for ensuring the veracity of a memory M, if
> i) S is morally responsible for ϕ-ing
> ii) there are multiple ways W(x) for S to ϕ in a morally responsible way
> iii) one way W(M) for S to ϕ in a morally responsible way presupposes remembering the contents of M veraciously
> iv) S intends to ϕ only by way W(M), or W(M) is favorable to any other way of W(x)
> v) S is, or should be, aware of reason to doubt the veracity of M
> vi) there are no exculpating factors present.

4.2.3) Reason to Doubt and Exculpating Circumstances

In the previous sections it was simply assumed that there is reason to doubt the veracity of your memories, and that there are no exculpating factors present. But it might seem like this is often not the case. Therefore, this subsection elaborates on these two points.

Reason to Doubt

In a nutshell, reason to doubt as described here aims at the awareness condition of moral responsibility described before (section 4.1.4). More specifically, reason to doubt in the case of veracity of memories seems to imply awareness of action and especially of the consequences of those actions. If you (tacitly) believe that your memories are veracious and have no reason to doubt their veracity, you are unaware when you are not remembering veraciously and as a result might be unaware of the likely consequences of your action (i.e. failing to be morally responsible). However, in order not to diverge too far from main question and given spatial limitations, important aspects such as when ignorance is culpable and exactly what kind of awareness is necessary, which are highly debated, cannot be resolved here, but see (Rudy-Hiller, 2018) for an overview and (Peels, 2017) for a more comprehensive debate of such and similar matters.

According to my account, a necessary condition for being morally responsible to ensure the veracity of memories is that one is, or should be, aware of reason to doubt the veracity of a certain memory. Yet, intuitively you might be expected to ensure the veracity of your memories in cases of utmost importance, even if you do not seem to have any reason to doubt. Consequently, it does not seem obvious that a reason to doubt is necessary. Imagine a doctor who has to decide between two different drugs to administer to a patient they know well. The doctor is aware that the patient is highly allergic to one but not the other drug. Furthermore, the doctor is quite certain that they remember which drug the patient was highly allergic to, but a look at the patient file to double-check would only take a second. Intuitively it might be claimed that the doctor has a moral responsibility to double-check, independently of whether they have a reason to doubt their memory or not.[34]

[34] See (Rosen, 2004, pp. 303-304) for a similar example and accompanying discussion viewed from a different angle.

Thus, at first it seems that at least in some cases reason to doubt is not necessary. To show why a reason to doubt is necessary, I think it is advisable to differentiate between two different kinds of reasons to doubt the veracity of your memories.

Specific reason to doubt veracity of a particular memory

It seems obvious that in some cases a reason to doubt the veracity of a particular memory is present. In case of my dying friend whom I promised to remember, I would have a strong reason to doubt whether my memory of his occupation as a computer scientist is veracious, if I also remember that he did not know how to use a computer (and could not otherwise explain this inconsistency). It seems implausible that he was a computer scientist and at the same time knew nothing about computers. Thus, I would conclude that there is something wrong with at least one of these memories, and I would have reason to doubt that they are veracious, making me aware that I might not be remembering veraciously and of possible consequences potentially resulting from it (such as that I might not be holding my promise). In such a case I would have a specific reason to doubt the veracity of a particular memory.

Yet, in case of the doctor, no such specific reason is available. Quite contrary, it might be that their memory seems to be quite coherent with the rest of the memories they have about their patient. Why then might it still seem right to say that the doctor has a moral responsibility to double-check if what they are remembering is veracious? Like I have proposed in the previous subsection, I do think that reason to doubt is necessary. Even more, it must be the case that there is reason to doubt the veracity of the particular memory in question. However, even if we do not have a *specific* reason to doubt a particular memory, there is another way.

General reason to doubt veracity of a particular memory

If I know myself to easily mix up occupations of others, I would not have an entirely specific reason to doubt that my alleged memory of my dying friend is veracious. Instead, I have a more generalized reason to doubt the veracity of my alleged memories of occupations others have. But since the occupation of my dying friend is an instantiation of this general class, my reason to doubt applies to it as well. Thus, I would have a reason to doubt the veracity of this particular memory.

Taken further, generally speaking, because of personal experience or the common and readily available knowledge about the malleability of memory provided by memory science among others, you are, or at least should be, aware that your alleged memories are not always veracious. Therefore, one could speak of a general reason to doubt the veracity of every of your memories, or of your memory as a mental faculty, and with that in most cases a reason to doubt the veracity of your memories will be given. This general reason to doubt instantiates that you are aware of your actions, i.e. potentially not remembering veraciously, or, in the case that you are not but at least should be aware, makes potential consequences of your action, to a degree, foreseeable.

Intuitively however, it does not seem to follow from this that it is *reasonable* to always try and ensure the veracity of *every* memory. In cases where you are quite certain of the veracity of a particular memory of yours, and the repercussions of your falsely remembering would be of little consequence, one might say that it would not be reasonable to double-check, especially if the cost of ensuring the veracity of your memories would be high. All-things-considered it could also be the case that the costs of ensuring the veracity of

memories of such almost negligible importance would be in competition with and outweighed by the moral responsibility you have towards something else (such as using your time and effort more reasonably). For instance, I might be morally responsible to remember veraciously what juice my wife wanted me to buy for her, but unless for example she is highly allergic to a certain kind, it seems unreasonable to expect me to drive home and check what kind of juice it was, if I am quite certain that it was mango juice, even if I am fully aware that we do falsely remember sometimes, and every memory could in principle be false. While it is true that I could have ensured the veracity of that memory and generally speaking might be responsible for ensuring it because of the awareness of a reason to doubt, the consequences of not ensuring the veracity in this case would be so marginal that it seems negligible to actually be held responsible for it.

Coming back to the doctor however, it seems different. Here the repercussions of falsely remembering could be immense, and ensuring the veracity by double-checking the patient file seems of negligible effort. But, precisely because we know that even if you are quite certain that your memories are veracious it can in principle always be the case that they are false, it seems unreasonable to take the risk and not double-check. The doctor does have a reason to doubt the veracity of their particular memory in so far as they are (or should be) aware that every one of their memories could in principle be false, and they would thus be morally responsible to ensure the veracity of their memory. Such a reason however is so generalized that it has lost much of its strength, and we would only find it reasonable to be accounted for in cases where the consequences of falsely remembering are immense, but the costs are minimal.

Alternative accounts

Summing up, reason to doubt is a necessary condition for it to make sense why, under certain circumstances, we are held responsible for not ensuring the veracity of our memories even if we feel certain that we are remembering veraciously. In cases where we have a specific reason to doubt the veracity of a particular memory, the reason to doubt usually plays a substantial role. As a rule of thumb, it seems that the more generalized the reason gets, the less weight is attributed to it, up to the point that, while you might still be morally responsible to ensure the veracity of your memories, all-things-considered it would seem unreasonable to ensure the veracity of it in highly generalized cases which are of little consequence and thus we would not bother holding someone else morally responsible.

Application of other accounts of moral responsibility to veracity of memories, could be construed in such a way as to not see reason to doubt as a necessary condition to ascribe moral responsibility. In the case of attributionism as described above (section 4.1.4), one might argue that neither reason to doubt nor awareness of action or foreseeability of consequences is necessary. Simply put, if I buy the wrong kind of juice for my wife due to my falsely remembering, I could be held responsible for not ensuring the veracity of my memories even if I neither am nor should be aware of reason to doubt the veracity of that memory, if not ensuring the veracity of the memory would be expressive of my disregard towards my wife's interests. It should be kept in mind, however, that this is just an oversimplification, and that there are caveats to attributionism as well (Rudy-Hiller, 2018) which, at least in some cases, might be viewed in such a way that reason to doubt is

necessary as well (for example, in the case just mentioned if not ensuring the veracity would only be an expression of disregard if it was made contrary to existing reason to doubt).

Thus, while in the standard view of moral responsibility reason to doubt seems to be necessary for you to be morally responsible to ensure the veracity of your memories, other accounts of moral responsibility might possibly do without it, but further investigation of the interplay between reason to doubt and veracity of memories in these accounts is needed.

Exculpating Circumstances

In the investigation so far, exculpating factors have been excluded. However, as was hinted at in the last section, it sometimes can be the case, that while you might generally speaking be morally responsible to ensure the veracity of certain memories, it can also be the case that all-things-considered this moral responsibility is outweighed or suspended by something else. In the example of my dying friend, it might be argued that while it is true that I have a moral responsibility to ensure the veracity of the memories of my dying friend if I promised to remember the man he was, it could be that ensuring the veracity would put undue burden on me such that I would have to neglect the moral responsibility I have towards other things. If, for example, the only way I could ensure the veracity of those memories would entail that I have to quit my job and as a result could not provide adequate food for my children (and there was no way to circumvent this constellation), intuitively it could be judged that all-things-considered I could not be held morally responsible to ensure the veracity of those memories under these circumstances. Not letting my children become malnourished could reasonably be judged to take precedence over ensuring the veracity of those memories. I cannot give a full account of when exactly factors exculpate concerning veracity of memories here since the details of when factors exculpate are more complex, but for the purpose of this paper it suffices to acknowledge that such factors can and do exist in at least some cases.

Conclusion

Here I have tried to show that you can be morally responsible for the veracity of your memories. This seems especially important since the more we are finding out about how memory works, the more we are getting aware of how susceptible to change memories are. Research on memory and false memory is currently flourishing and it seems like a promising endeavor to investigate how insights from memory science can be useful for the debate on the interaction between moral responsibility and veracity of memories. I have argued that reason to doubt is necessary given the standard view of moral responsibility, but it would be interesting to see if this might also be the case in more non-traditional accounts of moral responsibility such as attributionism. Furthermore, here I have restricted myself to individual, declarative memory, but it would be intriguing to apply this analysis to other forms of memory, and to collective memory in particular, both for moral responsibility directed at the past, as well as the future.

Future Prospects

Memory is fundamental to life as a human being. This importance is what motivated this thesis. My aim in writing it was to add a few interesting new ideas to the interdisciplinary study of memory and false memory. As I outlined at the beginning of the thesis, the scientific as well as philosophical study of memory and false memory has been flourishing in recent decades, and new bridges are being built between the different disciplines. These advances bring up intriguing new questions in the realm of memory research. In the process of trying to answer a few of these questions, many others are coming up.

Kinds of false procedural memory and false nondeclarative memory

The first topic I looked at was whether procedural memory could be false and what it would mean to falsely remember-how to do something. This question came up simply because current philosophical theories of memory and false memory were only concerned with declarative forms of memory, but at the same time nondeclarative memory seemed to require fulfillment of different conditions to be genuine. Thus, the question arose what it would mean for procedural memory to be false. While it is obvious that you might make a mistake in remembering-how to do something, it is far less obvious what exactly it could mean to falsely remember-how to do something. Answering this question required first to give an account of what it means to remember-how to do something. Philosophers have been studying procedural memory in some form, but usually conflated the notion of remembering-how with the notion of knowing-how. After setting the stage by defining what it means to remember-how to do something, I addressed the question of what it means to falsely remember-how to do something. I suggested that, roughly put, you falsely remember-how to do something if you try to perform an act which you have learned, but instead of performing the intended act, you correctly perform another act which you have also learned but ultimately it was not what you tried to perform. This notion turned out to possibly be quite helpful in refining well-established psychological research paradigms and their application in the area of interference research.

However, I have only looked at one form of procedural memory, namely motor memory. It does not seem obvious if the account presented here can easily be extended to all forms of procedural memory, i.e. to perceptual and cognitive procedural memory. This question could, however, be of great importance since most complex skills require the interaction of all three forms of procedural memory. Furthermore, a more generalized notion of false procedural memory might turn out to be applicable to or at least of some merit to other forms of memory as well. Thus, the question arises whether other forms of nondeclarative memory could be false, or at least what memory errors in those types of memory would look like. Additionally, it would be interesting to see if distinctions made by current false memory theories, such as misremembering vs. confabulation, could be applied to false procedural or nondeclarative memory. I suggest that answering these questions would not only be interesting in its own right out of intellectual or philosophical curiosity, but could, like in the case of false procedural memory presented here, be useful for applications such as the refinement of research paradigms of nondeclarative memory.

Remembering emotions fully and falsely

While researching what it would mean for other forms of nondeclarative memory to be false, I came across scientific publications which classified certain emotional responses as a form of nondeclarative memory. Intuitively this seemed wrong, since I took emotions to primarily be a kind of experience, much like visually experiencing an event. Thus, the question came up, what it might mean to remember an emotion, if it is possible at all. Answering this question turned out to be quite difficult, since scientists as well as philosophers disagree strongly amongst each other on what emotions are in the first place. I have opted to look at two components which most accounts of emotions identify as essential parts of an emotion, physiological responses and a subjective experience. However, these two components are markedly different in many ways, and it did not seem reasonable to assume that remembering them would be covered by only one type of memory. Thus, I proposed to study physiological responses in the framework of nondeclarative memory, and subjective experiences in the framework of episodic or experiential memory. Since there were barely any accounts that tried to do something similar, it took quite a bit of work to first establish what it might mean to just remember those two components of emotions.

A couple of things necessarily fell short in answering this question. First, obviously other components often attributed to emotions, such as cognitive, intentional or motivational aspects were not given the proper extensive examination they would deserve. However, it seems questionable if it is possible to give a general account of remembering emotions, given the quite diverse accounts of what emotions are. One way might be to simply pick an account of emotions and see what remembering emotions as defined by that account could be. Another, similar to the approach I have chosen, would be to look at individual components attributed to emotions and investigate what remembering each of these components could mean. Both approaches come with advantages and drawbacks. Simply picking one account of emotion would obviously disregard some components others deem essential to emotions. On the other hand, giving definitions of what it means to remember each single component of an emotion might miss the main point of what it means to remember emotions as a whole. A second point which fell short, but which might naturally follow an account of what it means to remember emotions, is the question of what it might mean to falsely remember an emotion. While I have offered some initial thoughts, this question will likely depend strongly on what remembering emotions is taken to be. This endeavor is even more complex at this point since there is no agreement on what an emotion is. However, I think ultimately answering that question will pay off, since the interactions between emotions and memories seem crucial, both in everyday life as well as in pathological interactions often found in mental disorders such as PTSD or generalized anxiety disorder (GAD).

When should you, or we, try to remember veraciously?

Intuitively *false* memory might have a negative connotation to it. You might associate it with such notions as ‘not genuine’ which intuitively on the level of value judgments seem negative. However, while researching treatment of PTSD and GAD for the remembering emotions topic, it seemed to me that some treatment options aimed at changing the emotional response caused by remembering experiences in order to ameliorate suffering. Thus, the question came up if these treatments intentionally try to change how you

remember, be that declaratively or nondeclaratively, your past and under what circumstances that is something that should or should not be done. Out of this the more general question arose if and when you might be morally responsible to ensure the veracity of your memories. Here I have argued that you can be and are morally responsible to ensure the veracity of your memories under certain conditions, one of them being that remembering veraciously is important concerning the moral responsibility towards something else. Thus, by looking at whether you are morally responsible not only to remember or not, but to ensure the veracity of your memories, I have tried to look at the topic of false memories from a normative perspective.

The answer I developed was intentionally kept both quite general in some ways and quite specific in others. It was general in the sense that it could be applied to different strands of normative ethics, and be useful in topics of applied ethics, such as the ethics surrounding the role of false memories in treatment of mental disorders. This approach left open questions of what concretely might be expected of you for you to ensure the veracity of your memories, which undoubtedly will necessitate taking a close look at what scientific research can tell us about memory and how memories change. The answer I developed was specific in the way that I only looked at a specific type of memory, namely declarative individual memory. Yet, as I have shown here, other types of memory can go wrong, which can have immense repercussions. Thus, it seems reasonable to ask whether you could be morally responsible to ensure that those memories are veracious as well, and what this might mean. At the same time, I have only looked at whether or not *you* are morally responsible to ensure the veracity of *your* memories, but memories might also be understood in a collective or societal sense. An interesting question for the future would thus be whether *we* are morally responsible for the veracity of *our* memories.

Conclusion – where do we go from here?

I had the chance to look at only some interesting questions in the neurophilosophy and the ethics of false memories and false beliefs. As it often is the case, in trying to answer these questions many more came up. It is encouraging to see how diverse the views about the nature of memory and false memory have become in recent years. Undoubtedly some of this diversity is, at least in part, due to our growing knowledge provided by scientific research about the processes of remembering, and the openness of philosophers to try and include this new knowledge into their theories. Since memory is such an immensely complex topic, interdisciplinary interplay will be crucial in trying to understand more about it. Here I have tried to include perspectives from philosophy, biology and psychology. Yet, this seems hardly enough given how intertwined memory is with almost every aspect of human life. The delineation of different kinds of errors in procedural memory has shown that the design of interfaces can be decisive in what and why different kinds of errors are made. Trying to answer the question of what it might mean to remember emotions made it clear that it is necessary to not only think across different types of memory but also across different theories and disciplines to understand what intuitively might appear as a single complex phenomenon. Lastly, I was able to get only a glance at the ethics of false memories and false beliefs, which, however, was limited to only one type of memory on an individual level. To understand the ethics of false memories and false beliefs, it will be necessary to take an interdisciplinary dive into other research areas such as sociology or anthropology.

While my thesis ends here, I hope it will be just the beginning of a future full of intriguing inquiries about the neurophilosophy and ethics of false memories.

References

Adams, J. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101*(1), 41–74.

Albrecht, I. V. (2018). Graveside and Other Asymmetrical Promises. *Social Theory and Practice, 44*(4), 469–483.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Arcangeli, M., & Dokic, J. (2018). Afective memory: A little help from our imagination. In K. Michaelian, D. Debus, & D. Perrin, *New directions in the philosophy of memory* (pp. 139–157). New York: Routledge.

Arnold, M. B. (1960). *Emotion and personality. Vol. I: Psychological aspects; Vol. II:Neurological and physiological aspects.* New York: Columbia University Press.

Arun, A., Kandel, E. R., & Rayman, J. B. (2019). The Neurobiology of Fear Generalization. *Frontiers in Behavioral Neuroscience, 12*(329).

Bard, P. (1928). A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology-Legacy Content, 84*(3), 490–515.

Beaunieux, H., Hubert, V., Witkowski, T., Pitel, A.-L., Rossi, S., Danion, J.-M., . . . Eustache, F. (2006). Which processes are involved in cognitive procedural learning? *Memory, 14*(5), 521–539.

Bedford, E. (1956). Emotions. *Proceedings of the Aristotelian Society, 57*(January), 281–304.

Bengson, J., & Moffett, M. (2011). *Knowing How: Essays on Knowledge, Mind, and Action.* New York: Oxford University Press.

Bergson, H. (1991). *Matter and Memory (Matière et Mémoire).* (N. Paul, & S. Palmer, Trans.) New York: Zone Books. Original work published 1908.

Bernecker, S. (2001). Russell on Mnemic Causation. *Principia, 5*(1–2), 149–185.

Bernecker, S. (2008). *The Metaphysics of Memory.* New York: Springer.

Bernecker, S. (2010). *Memory: A Philosophical Study.* Oxford: Oxford University Press.

Bernecker, S. (2011). Memory knowledge. In S. Bernecker, & D. Pritchard, *The Routledge Companion to Epistemology* (S. 326–334). New York: Routledge.

Bernecker, S. (2017a). Memory and Truth. In S. Bernecker, & K. Michaelian, *The Routledge Handbook of Philosophy of Memory* (pp. 51–62). London: Routledge.

Bernecker, S. (2017b). A Causal Theory of Mnemonic Confabulation. *8*(1207), 1–14.

Bernecker, S. (2018). On the blameworthiness of forgetting. In K. Michaelian, D. Debus, & D. Perrin, *New Directions in the Philosophy of Memory* (S. 241–258). New York: Routledge.

Bernecker, S., & Michaelian, K. (2017). Part IX History of the philosophy of memory. In S. Bernecker, & K. Michaelian, *The Routledge handbook of philosophy of memory* (S. 383–571). New York: Routledge.

Bernstein, D. M., & Loftus, E. F. (2009). How to Tell If a Particular Memory Is True or False. *Perspectives on Psychological Science, 4*(4), 370–374.

Betsch, T., Haberstroh, S., Molter, B., & Glöckner, A. (2004). Oops, I did it again—relapse errors in routinized decision making. *Organizational Behavior and Human Decision Processes, 93*(1), 62–74.

Blanchard, R. J., & Blanchard, C. D. (1969). Crouching as an index of fear. *Journal of Comparative and Physiological Psychology, 67*(3), 370–375.

Blustein, J. (2017). A duty to remember. In S. Bernecker, & K. Michaelian, *Routledge handbook of philosophy of memory* (pp. 351–363). London: Routledge.

Bock, O., Schneider, S., & Bloomberg, J. (2001). Conditions for interference versus facilitation during sequential sensorimotor adaptation. *Experimental Brain Research, 138*(3), 359–365.

Bortolotti, L., & Cox, R. E. (2009). 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition, 18*(4), 952–965.

Brainerd, C. J., & Reyna, V. F. (2005). Your Ancients. In C. J. Brainerd, & V. F. Reyna, *The Science of False Memory* (pp. 3–23). Oxford: Oxford University Press.

Brydges, R., Carnahan, H., Backstein, D., & Dubrowski, A. (2007). Application of Motor Learning Principles to Complex Surgical Tasks: Searching for the Optimal Practice Schedule. *Journal of Motor Behavior, 39*(1), 40–48.

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology, 39*(1/4), 106–124.

Cheng, S., & Werning, M. (2016). What is episodic memory if it is a natural kind? *Synthese, 193*(5), 1345–1385.

Cheyne, J., Carriere, J., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition, 15*(3), 578–592.

Christianson, S. Å., & Safer, M. A. (1996). Emotional events and emotions in autobiographical memories. In D. C. Rubin, *Remembering our past: Studies in autobiographical memory* (pp. 218–243). Cambridge: Cambridge University Press.

Clark, A., Parakh, R., Smilek, D., & Roy, E. (2012). The Slip Induction Task: Creating a window into cognitive control failures. *Behavior Research Methods, 44*(2), 558–574.

Clark-Polner, E., Wager, T. D., Satpute, A. B., & Barrett, L. F. (2016). Neural fingerprinting: Meta-analysis, variation, and the search for brain-based essences in the science of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones, *Handbook of Emotions* (4 ed., pp. 146–165). New York: Guilford Publications, Inc.

Cohen, N., & Eichenbaum, H. (1993). The Hippocampal System and the Procedural-Declarative Memory Distinction: A Comprehensive Proposal. In N. Cohen, & H. Eichenbaum, *Memory, Amnesia, and the Hippocampal System* (pp. 55–92). London: MIT Press.

Cohen, N., & Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science, 210*(4466), pp. 207-210.

Collins, A., & Hay, D. (1994). Connectionism and Memory. In P. Morris, & M. Gruneberg, *Theoretical Aspects of Memory* (pp. 195–237). London: Routledge.

Craver, C. F., & Rosenbaum, S. R. (2018). Consent Without Memory. In M. Kourken, D. Debus, & D. Perrin, *New Directions in the Philosophy of Memory* (S. 259–275). New York: Routledge.

Curran, T. (2001). Implicit learning revealed by the method of opposition. *Trends in Cognitive Sciences, 5*(12), 503–504.

Dalla Barba, G. (2002). *Memory, Consciousness and Temporality.* Boston: Kluwer.

Damasio, A. R. (1994). Emotions and Feelings. In A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (pp. 127-164). New York: Avon Books.

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

de Sousa, R. (1987). Emotions and Their Objects. In R. de Sousa, *The Rationality of Emotion* (pp. 107–140). Cambridge, MA: MIT Press.

Debus, D. (2007). Being Emotional about the Past: On the Nature and Role of Past-Directed Emotions. *Noûs, 41*(4), 758-779.

Debus, D. (2010). Accounting for epistemic relevance: A new problem for the causal theory of memory. *American Philosophical Quarterly, 47*(1), 17–29.

Debus, D. (2017). Memory causation. In S. Bernecker, & K. Michaelian, *The Routledge handbook of philosophy of memory* (pp. 63–75). New York: Routledge.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Journal of Experimental, 5*(3), 17–22.

Deonna, J., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction.* London: Routledge.

Dickinson, A. (2007). Learning: The need for a hybrid theory. In H. Roediger, Y. Dudai, & S. Fitzpatrick, *Science of Memory: Concepts* (pp. 41–44). Oxford: Oxford University Press.

Dixon, T. (2012). "Emotion": The History of a Keyword in Crisis. *Emotion Review, 4*(4), 338–344.

Dressel, A. (2014). Directed Obligations and the Trouble with Deathbed Promises. *Ethical Theory and Moral Practice, 18*(2), 323–335.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes.* Cambridge: MIT Press.

Dretske, F. (2010). Triggering and Structuring Causes. In T. O'Connor, & C. Sandis, *A Companion to the Philosophy of Action* (pp. 139-144). West Sussex: Wiley-Blackwell.

Dudai, Y. (1992). Why 'learning' and 'memory' should be redefined (or, an agenda for focused reductionism). *Concepts in Neuroscience*, 9–121.

Dudai, Y. (2007a). Memory. In H. Roediger, Y. Dudai, & S. Fitzpatrick, *Science of Memory: Concepts* (p. 11). Oxford: Oxford University Press.

Dudai, Y. (2007b). Memory: It's all about representations. In H. L. Roediger, D. Yadin, & S. M. Fitzpatrick, *Science of Memory: Concepts* (pp. 13–16). Oxford: Oxford University Press.

Edwards, W. (2010). *Motor Learning and Control: From Theory to Practice.* Wadsworth: Cengage Learning.

Fantl, J. (2017). Knowledge How. In E. Zalta, *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition ed.). Metaphysics Research Lab, Stanford University.

Fehr, B., & Russel, J. A. (1984). Concept of Emotion Viewed From a Prototype Perspective. *Journal of Experimental Psychology: General, 113*(3), 464–486.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge: Cambridge University Press.

Freud, S. (1917). Das Versprechen [The Slip of Tongue]. In S. Freud, *Zur Psychopathologie des Alltagslebens [The Psychopathology of Everyday Life]* (Fünfte Auflage Ausg., S. 44–84). Berlin: S.Karger. Original work published 1904.

Frings, C., Schneider, K., & Fox, E. (2015). The negative priming paradigm: An update and implications for selective attention. *Psychonomic Bulletin & Review, 22*(6), 1577–1597.

Gerrans, P. (2018). Painful Memories. In K. Michaelian, D. Debus, & D. Perrin, *New directions in the philosophy of memory* (pp. 158–178). New York: Routledge.

Ghilardi, F., Moisello, C., Silvestri, G., Ghez, C., & Krakauer, J. (2009). Learning of a sequential motor skill comprises explicit and implicit components that consolidate differently. *Journal of Neurophysiology, 101*(5), 2218–2229.

Glick, E. (2011). Two Methodologies for Evaluating Intellectualism. *Philosophy and Phenomenological Research, 83*(2), 398–434.

Goldie, P. (2009). Part I: What Emotions Are. In P. Goldie, *The Oxford Handbook of Philosophy of Emotion* (pp. 15–117). Oxford: Oxford University Press.

Graf, P., & Schacter, D. L. (1985). Implicit and Explicit Memory for New Associations in Normal and Amnesic Subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(3), 501–518.

Grafton, S., Mazziotta, J., Presty, S., Friston, K., Frackowiak, R., & Phelps, M. (1992). Functional anatomy of human procedural learning determined with regional cerebral blood flow and PET. *Journal of Neuroscience, 12*(7), S. 2542-2548.

Grison, S., Tipper, S., & Hewitt, O. (2005). Long-term negative priming: Support for retrieval of prior attentional processes. *The Quarterly Journal of Experimental Psychology Section, 58A*(7), 1199–1224.

Hacker, P. M. (2013). Memory. In P. M. Hacker, *The Intellectual Powers: a Study of Human Nature* (p. 316352). Oxford: Wiley-Blackwell.

Hawley, K. (2003). Success and Knowledge-How. *American Philosophical Quarterly, 40*(1), 19–31.

Heitz, R. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*(150), 1–19.

Helene, A., & Xavier, G. (2006). Working memory and acquisition of implicit knowledge by imagery training, without actual task performance. *Neuroscience, 139*(1), 401–413.

Izard, C. E. (2010). The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation. *Emotion Review, 2*(4), 363–370.

James, W. (1983a). The Emotions. In W. James, & G. Miller (Ed.), *The Principles of Psychology* (pp. 1058–1097). Cambridge: Harvard University Press. Original work published 1890.

James, W. (1983b). *The Principles of Psychology.* (G. Miller, Ed.) Cambridge: Harvard University Press. Original work published 1890.

Kinsbourne, M. (1981). Single-channel Theory. In D. Holding, *Human Skills* (pp. 65–89). New York: John Wiley & Sons.

Knowlton, B., Siegel, A., & Moody, T. (2017). Procedural Learning in Humans. In J. Byrne, *Learning and memory: a comprehensive reference* (Vol. 3, pp. 295–312). Oxford: Academic Press.

Koedijker, J., Oudejans, R., & Beek, P. (2010). Interference effects in learning similar sequences of discrete movements. *Journal of Motor Behavior, 42*(4), 209–222.

Kolber, A. J. (2006). Therapeutic forgetting: The legal and ethical implications of memory dampening. *Vanderbilt Law Review, 59*(5), 1561–1626.

Kremer, P., Spittle, M., & Malseed, S. (2011). Retroactive interference and mental practice effects on motor performance: a pilot study. *Perceptual and Motor Skills, 113*(3), 805–814.

Künzell, S., & Lukas, S. (2011). Facilitation effects of a preparatory skateboard training on the learning of snowboarding. *Kinesiology, 43*(1), 56–63.

LaBar, K. S. (2016). Fear and Anxiety. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones, *Handbook of Emotions* (pp. 751–773). New York: The Guilford Press.

Laney, C., & Loftus, E. (2013). Recent advances in false memory research. *South African Journal of Psychology, 43*(2), 137–146.

Lazarus, R. S. (1991). *Emotion and Adaptation.* New York: Oxford University Press.

LeDoux, J. E. (1992). Emotion as memory: Anatomical systems underlying indelible neural traces. In S. Å. Christianson, *The handbook of emotion and memory: Research and theory* (pp. 269–288). New York: Lawrence Erlbaum Associates, Inc.

Levine, L. J., Safer, M. A., & Lench, H. C. (2006). Remembering and Misremembering Emotions. In L. J. Sanna, & E. C. Chang, *Judgements over Time: The Interplay of Thoughts, Feelings and Behaviors* (pp. 271–290). New York: Oxford University Press.

Levy, N. (2005). The Good, the Bad and the Blameworthy. *Journal of Ethics and Social Philosophy, 1*(2), 1–16.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs, 13*(4), 455–476.

Liao, M. S. (2017). The ethics of memory modification. In S. Bernecker, & M. Kourken, *Routledge handbook of philosophy of memory* (pp. 373–382). London: Routledge.

Loftus, E. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366.

Loftus, E., & Pickrell, J. (1995). The Formation of False Memories. *Psychiatric Annals, 25*(12), 720–725.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*(4), 245–265.

Malcolm, N. (1963). Three Forms of Memory. In N. Malcolm, *Knowledge and Certainty* (pp. 203–221). Englewood Cliffs: Prentice Hall.

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review, 75*(2), 161–196.

Mayr, S., & Buchner, A. (2006). Evidence for Episodic Retrieval of Inadequate Prime Responses in Auditory Negative Priming. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), 932–943.

Mayr, S., & Buchner, A. (2007). Negative Priming as a Memory Phenomenon: A Review of 20 Years of Negative Priming Research. *Journal of Psychology, 215*(1), 35–51.

McDermott, K. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. *Psychonomic Bulletin & Review, 4*(4), 582–586.

McKone, E., & Murphy, B. (2000). Implicit false memory: Effects of modality and multiple study presentations on long-lived semantic priming. *Journal of Memory and Language, 43*(1), 89–109.

Mendes, W. B. (2016). Emotion and the Autonomic Nervous System. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones, *Handbook of Emotions* (4 ed., pp. 166–181). New York: Guilford Publications, Inc.

Merriam-Webster Online Dictionary. (2020). emotion. Abgerufen am 28. February 2020 von https://www.merriam-webster.com/dictionary/emotion

Michaelian, K. (2016a). The Simulation Theory. In K. Michaelian, *Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past* (pp. 97–121). Cambridge: MIT Press.

Michaelian, K. (2016b). Confabulating, Misremembering, Relearning: The Simulation Theory of Memory and Unsuccessful Remembering. *Frontiers in Psychology, 7*(1857), 1–13.

Michaelian, K. (2020). Confabulating as Unreliable Imagining: In Defence of the Simulationist Account of Unsuccessful Remembering. *Topoi, 39*, 133–148.

Michaelian, K., & Robins, S. K. (2018). Beyond the Causal Theory? Fifty Years After Martin and Deutscher. In K. Michaelian, D. Debus, & D. Perrin, *New Directions in the Philosophy of Memory* (pp. 13–32). London: Routledge.

Michaelian, K., Perrin, D., & Sant'Anna, A. (2020). Continuities and Discontinuities Between Imagination and Memory: The View from Philosophy. In A. Abraham, *The Cambridge Handbook of the Imagination* (pp. 293–310). Cambridge: Cambridge University Press.

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM. *Neuropsychologia, 6*(3), 215–234.

Milner, B., Squire, L., & Kandel, E. (1998). Cognitive Neuroscience and the Study of Memory. *Neuron, 20*(3), pp. 445–468.

Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*(2), 239–260.

Moscovitch, M. (2007). Memory: Why the engram is elusive. In H. Roediger, Y. Dudai, & S. Fitzpatrick, *Science of Memory: Concepts* (pp. 17–22). Oxford: Oxford University Press.

Mulligan, K., & Scherer, K. R. (2012). Toward a Working Definition of Emotion. *Emotion Review, 4*(4), 345–357.
Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature, 406*(6797), 722–726.
Nelkin, D. K. (2019, Summer). *Moral Luck.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2019/entries/moral-luck/
Nikulin, D. (2015). *Memory: A History.* Oxford: Oxford University Press.
Noë, A. (2005). Against Intellectualism. *Analysis, 65*(4), 278–290.
Nozick, R. (1981). Knowledge. In R. Nozick, *Philosophical Explanation* (pp. 172–196). Cambridge, MA: Harvard University Press.
Nussbaum, M. (2001). *Upheavals of Thought: The Intelligence of Emotions.* Cambridge: Cambridge University Press.
Nyberg, L., Eriksson, J., Larsson, A., & Marklund, P. (2006). Learning by doing versus learning by thinking: An fMRI study of motor and mental training. *Neuropsychologia, 44*(5), 711–717.
Olton, D. S. (1979). Mazes, maps, and memory. *American Psychologist, 34*(7), 583–596.
Panzer, S., & Shea, C. (2008). The learning of two similar complex movement sequences: Does practice insulate a sequence from interference? *Human Movement Science, 27*(6), 873–887.
Panzer, S., Wilde, H., & Shea, C. (2006). Learning of similar complex movement sequences: proactive and retroactive effects on learning. *Journal of Motor Behavior, 38*(1), 60–70.
Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin, 116*(2), 220–244.
Peels, R. (2014). What Kind of Ignorance Excuses? Two Neglected Issues. *Philosophical Quarterly, 64*(256), 478–496.
Peels, R. (2017). *Perspectives on Ignorance from Moral and Social Philosophy.* New York: Routledge.
Pezdek, K., & Lam, S. (2007). What research paradigms have cognitive psychologists used to study “False memory,” and what are the implications of these choices? *Consciousness and Cognition, 16*(1), 2–17.
Porter, J., & Magill, R. (2010). Systematically increasing contextual interference is beneficial for learning sport skills. *Journal of Sports Sciences, 28*(12), 1277–1285.
Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion.* Oxford: Oxford University Press.
Prinz, J. (2005). Are Emotions Feelings? *Journal of Consciousness Studies, 12*(8–10), 9–25.
Quirk, G. J., & Mueller, D. (2008). Neural Mechanisms of Extinction Learning and Retrieval. *Neuropsychopharmacology, 33*(1), 56–72.
Ramirez, S., Liu, X., Lin, P.-A., Suh, J., Pignatelli, M., Redondo, R. L., . . . Tonegawa, S. (2013). Creating a False Memory in the Hippocampus. *Science, 341*(6144), 387–391.
Reason, J. (1990). *Human Error.* Cambridge: Cambridge University Press.
Ribot, T. A. (1897). The memory of feelings. In T. A. Ribot, *The psychology of the emotions [La Psychologie des sentiments]* (pp. 140–171). London: Walter Scott, Ltd.
Robins, S. (2017). Memory traces. In S. Bernecker, & K. Michaelian, *The Routledge handbook of philosophy of memory* (pp. 76–87). New York: Routledge.
Robins, S. (2020). Mnemonic Confabulation. *Topoi, 39*(1), 121–132.
Robins, S. K. (2016). Misremembering. *Philosophical Psychology, 29*(3), 432–447.
Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 803–814.

Roediger, H., Balota, D., & Watson, J. (2001). Spreading activation and arousal of false memories. In H. Roediger, J. Nairne, I. Neath, & A. Suprenant, *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association.

Roediger, H., Dudai, Y., & Fitzpatrick, S. (2007). *Science of Memory: Concepts.* Oxford: Oxford University Press.

Rosen, G. (2004). Culpability and Ignorance. *Proceedings of the Aristotelian Society, 103*(1), 61–84.

Rosenbaum, D., Carlson, R., & Gilmore, R. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology, 52*(1), 453–470.

Ross, B. M. (1991). Memory Observed by Introspection. In B. M. Ross, *Remembering the Personal Past* (pp. 12–44). New York: Oxford University Press.

Rudy-Hiller, F. (2018). *The Epistemic Condition for Moral Responsibility*, Fall 2018 Edition. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/

Rüsch, N., Corrigan, P. W., Bohus, M., Kühler , T., Jacob, G. A., & Lieb, K. (2007). The Impact of Posttraumatic Stress Disorder on Dysfunctional Implicit and Explicit Emotions Among Women With Borderline Personality Disorder. *The Journal of Nervous and Mental Disease, 195*(6), 537–539.

Russell, B. (2005). Memory. In B. Russell, *The Analysis of Mind* (pp. 93–111). New York: Dover Publications. Original work published 1921.

Ryle, G. (2009a). Dispositions and occurences. In G. Ryle, *The Concept of Mind* (pp. 100–135). New York: Routledge. Original work published 1949.

Ryle, G. (2009b). Knowing How and Knowing That. In G. Ryle, *The Concept of Mind* (pp. 14–48). New York: Routledge. Original work published 1949.

Scarantino, A. (2016). The philosophy of emotions and its impact on affective science. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones, *Handbook of Emotions* (pp. 3–47). New York: The Guilford Press.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379–399.

Schacter, D., Gallo, D., & Kensinger, E. (2011). The cognitive neuroscience of implicit and false memories: Perspectives on processing specificity. In J. Nairne, *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 353–378). New York: Psychology Press.

Scherer, K. R., & Moors, A. (2019). The Emotion Process: Event Appraisal and Component Differentiation. *Annual Review of Psychology, 70*, 719–745.

Schnider, A. (2018). *The Confabulating Mind: How the Brain Creates Reality* (2nd ed.). Oxford: Oxford University Press.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* Cambridge: Cambridge University Press.

Seli, P., Cheyne, J., & Smilek, D. (2012). Attention failures versus misplaced diligence: Separating attention lapses from speed–accuracy trade-offs. *Consciousness and Cognition, 21*(1), 277–291.

Sharit, J. (2012). Human Error and Human Reliability Analysis. In G. Salvendy, *Handbook of Human Factors and Ergonomics* (4th ed., pp. 734–800). New Jersey: John Wiley & Sons.

Sher, G. (2009). *Who Knew? Responsibility Without Awareness.* New York: Oxford University Press.

Siegel, S. (2016, Winter). *The Contents of Perception.* Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2016/entries/perception-contents/

Spear, N. E. (2007). Retrieval: Properties and effects. In H. L. Roediger, D. Yadin, & S. M. Fitzpatrick, *Science of Memory: Concepts* (pp. 215–219). Oxford: Oxford University Press.

Squire, L. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171–177.

Squire, L. R., & Zola-Morgan, S. (1988). Memory: Brain systems and behavior. *Trends in Neurosciences, 11*(4), 170–175.

Squire, L., Cohen, N., & Zouzounis, J. (1984). Preserved memory in retrograde amnesia: Sparing of a recently acquired skill. *Neuropsychologia, 22*(2), pp. 145-152.

Stanley, J., & Williamson, T. (2001). Knowing How. *The Journal of Philosophy, 98*(8), 411–444.

Talbert, M. (2013). Unwitting Wrongdoers and the Role of Moral Disagreement in Blame. In D. Shoemaker, *Oxford Studies in Agency and Responsibility* (Vol. 1, pp. 225–245). Oxford: Oxford University Press.

Talbert, M. (2019, Winter). *Moral Responsibility*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/

Teroni, F. (2017). The Phenomenology of Memory. In S. Bernecker, & K. Michaelian, *The Routledge handbook of philosophy of memory* (pp. 21–33). New York: Routledge.

Terzis, G. (2001). How crosstalk creates vision-related eureka moments. *Philosophical Psychology, 14*(4), 393–421.

Titchener, E. B. (1895). Affective memory. *The Philosophical Review, 4*(1), 65–76.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson, *Organization of Memory* (pp. 381–402). New York: Academic Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Journal of Psychology, 26*(1), 1–26.

Tulving, E. (2000). Concepts of Memory. In E. Tulving, & F. I. Craik, *The Oxford handbook of memory* (pp. 33–43). New York: Oxford University Press.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace, & J. Metcalfe, *The missing link in cognition: Origins of self-reflective consciousness* (pp. 3–56). New York: Oxford University Press.

Tulving, E. (2007). Are there 256 Kinds of Memory? In J. S. Nairne, *The Foundations of Remembering: Essays in Honor of Henry L. Roediger* (pp. 40–52). New York: Psychology Press.

Tulving, E. (2016). Episodic Memory. In R. J. Sternberg, S. T. Fiske, & D. J. Foss, *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about their Most Important Contributions* (pp. 152–156). New York: Cambridge University Press.

Underwood, B. (1957). Interference and forgetting. *Psychological Review, 64*(1), 49–60.

Uniacke, S. (2010). Responsibility: Intention and consequence. In J. Skorupski, *The Routledge Companion to Ethics* (pp. 596–606). New York: Routledge.

Walter, C., & Swinnen, S. (1994). The formation and dissolution of "bad habits" during the acquisition of coordination skills. In S. Swinnen, J. Massion, H. Heuer, & P. Casaer, *Interlimb Coordination* (pp. 491–513). San Diego: Academic Press.

Wang, P., Zhang, X., Liu, Y., Liu, S., Zhou, B., Zhang, Z., . . . Jiang, T. (2013). Perceptual and response interference in Alzheimer's disease and mild cognitive impairment. *Clinical Neurophysiology, 124*(12), 2389–2396.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics, 24*(2), 260–288.

Werning, M., & Cheng, S. (2017). Taxonomy and Unity of Memory. In S. Bernecker, & K. Michaelian, *The Routledge handbook of philosophy of memory* (pp. 7–20). New York: Routledge.

Wickens, C., Hollands, J., Banbury, S., & Parasuraman, R. (2016). Memory and Training. In C. Wickens, J. Hollands, S. Banbury, & R. Parasuraman, *Engineering Psychology and Human Performance* (4th ed., pp. 197–244). New York: Routledge.

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious Affective Reactions to Masked Happy Versus Angry Faces Influence Consumption Behavior and Judgments of Value. *Personality and Social Psychology Bulletin, 31*(1), 121–135.

Wohldmann, E., Healy, A., & Bourne, L. (2008). A mental practice superiority effect: Less retroactive interference and more transfer than physical practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 823–833.

Wood, C., & Ging, C. (1991). The role of interference and task similarity on the acquisition, retention, and transfer of simple motor skills. *Research Quarterly for Exercise and Sport, 62*(1), 18–26.

Wylie, S., Wildenberg, W., Ridderinkhof, R., Bashore, T., Powell, V., Manning, C., & Wooten, F. (2009). The effect of Parkinson's disease on interference control during action selection. *Neuropsychologia, 47*(1), 145–157.

Wylie, S., Wildenberg, W., Ridderinkhof, R., Bashore, T., Powell, V., Manning, C., & Wooten, F. (2009). The effect of speed-accuracy strategy on response interference control in Parkinson's disease. *Neuropsychologia, 47*(1), 1844–1853.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*(2), 151–175.

Zajonc, R. B. (2000). Feeling and thinking: Closing the debate over the independence of affect. In J. P. Forgas, *Feeling and Thinking: The Role of Affect in Social Cognition* (pp. 31–58). Cambridge: Cambridge University Press.