Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

Optogenetics and the Mechanism of False Memory

Abstract

Constructivists about memory argue that our memories of past events are produced by building representations of those events from a generalized information store (e.g., De Brigard 2014a; Michaelian 2012). The view is motivated by the memory errors discovered in cognitive psychology. Little has been known about the neural mechanisms by which false memories are produced. Recently, using a method I call the *Optogenetic False Memory Technique* (O-FaMe), neuroscientists have created false memories in mice (e.g., Ramirez et al. 2013). In this paper, I examine how Constructivism fares in light of O-FaMe results. My aims are two-fold. First, I argue that errors found in O-FaMe and cognitive psychology are similar behaviorally. Second, Constructivists should be able to explain the former since they purport to explain the latter, but they cannot. I conclude that O-FaMe studies reveal details about the mechanism by which false memories are produced that are incompatible with the explanatory approach to false memories favored by Constructivism.

§1 Introduction

Many memory theorists, in both philosophy and cognitive science, now endorse a view about the nature of memory, which I call *Constructivism*. Constructivists argue that memory is a capacity for building (i.e., constructing) plausible representations of past events from a generalized network of information. The view is understood as an alternative to the traditional 'warehouse conception' of memory, according to which discrete, well-preserved representations are retrieved from a memory store. Contemporary philosophical accounts of Constructivism derive motivation from the nature and extent of memory errors uncovered in cognitive psychology (e.g., De Brigard 2014a; Michaelian 2012; 2013). Decades of research into false memory reveal that memories of past experiences can be easily and systematically distorted, and further, that these distortions often have little to no influence on the felt sense of remembering. These Constructivists urge a rethinking of the capacity to remember past events in light of this evidence of persistent and pervasive errors.

Constructivists' rethinking of memory has focused on capturing results from behavioral studies and functional neuroimaging. It has not been constrained by molecular and circuit level neuroscience, and for good reason: very little has been known about the

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

neural mechanism(s) by which memory errors are produced, owing in large part to the lack of an animal model of false memory. The experimental paradigms by which false memories are produced in humans have no straightforward translation to non-human cases. This is because false memories are misrepresentations, distorted representations of past experiences or previously acquired information. The memories involved are declarative—in the most intriguing cases they are memories of particular events from the rememberer's past.[1] It is a matter of debate whether non-human animals possess the forms of memory and mental representation required for producing such errors, and even supposing some of them do, it has been difficult to imagine how to design experiments that could detect these memory errors.

This is beginning to change. Using a method I call the *Optogenetic False Memory Technique* (O-FaMe), neuroscientists have demonstrated recently the ability to create false memories in mice (e.g., Ramirez et al. 2013; Redondo et al. 2014).[2] O-FaMe makes use of optogenetics, a new method for manipulating neurons through induced sensitivity to light. The method has been credited with "spurring a revolution in neuroscience research" because it allows precise temporal control of cellular activity in living, behaving organisms (Häusser 2014: 1012). Optogenetics is of particular interest for investigating the mechanisms of false memory because it provides a way to reactivate memories in a non-human animal without returning the animal to the learning context. As O-FaMe demonstrates, this technique makes it possible to not only reactivate these memories, but to distort them as well, creating a false memory. In this paper, I explore the following question: how does Constructivism fare in light of the discoveries about the mechanism of false memory provided by O-FaMe?

[1] I resist labeling these memories of particular past events as *episodic* so as to avoid debates over whether episodic memory must involve a rich phenomenal character and whether non-human animals are capable of episodic remembering. I prefer the more neutral category *event memory*, for reasons elaborated on in §4.1.

[2] The genealogy of non-human animal models of false memory extends back further, of course. I discuss the recent history of this research in §3.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

My aims are two-fold. First, I argue that the memory errors produced in O-FaMe studies are sufficiently similar to those produced by studies in cognitive psychology. In both cases, the errors should be understood as cases of misremembering. Constructivism should be able to explain the former since it purports to explain the latter. My second aim is to argue that philosophical accounts of Constructivism are ill-equipped to do so. O-FaMe studies reveal details about the mechanism by which memory errors are produced that are incompatible with the Constructivist's explanatory approach. Specifically, O-FaMe studies suggest that false memories are the result of an interaction between the memory trace (or engram) and additional, misleading information and that there are mechanistic differences in the production of successful memories and various kinds of memory error. Constructivist theorizing is in tension with both of these claims. Constructivism's difficulty explaining the false memories produced in O-FaMe exposes an even more fundamental tension, between cognitive and neurobiological approaches to memory regarding the need to appeal to engrams, or memory traces, in the study of remembering. I conclude with a brief discussion of this issue.

§2 Memory Constructivism

Memory *Constructivism* is a view with many ancestors. Versions of the view can be found in various historical studies of memory (e.g., Sutton 1998 and Draaisma 2000). Another, possibly distinct strand runs through experimental cognitive psychology (e.g., Bartlett 1932; Neisser 1967; Loftus 2003; Klein 2013). Most proponents of *Constructivism* claim that memory is a process for building plausible representations of past events to suit one's current interests and future plans. My focus in this paper is on Constructivism as articulated by contemporary philosophers of memory (e.g., De Brigard 2014a; Michaelian 2012; Sutton and Windhorst 2009). More specifically, I am interested in versions of philosophical Constructivism that are motivated by a desire to account for the preponderance of memory errors revealed by decades of research in cognitive psychology, as well as more recent evidence from cognitive neuroscience. Understanding the details of

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

this view requires, first, an introduction to the empirical evidence motivating it. I offer a review of the motivating evidence in §2.1, focusing on the two well-established experimental paradigms that have guided this strand of Constructivism. Then, in §2.2, I present the view's two central commitments, concentrating my discussion on the most fully developed version of philosophical Constructivism available to date: De Brigard's (2014a) *Episodic Simulation Theory*.

2.1 Motivations for Constructivism

Studies of memory's malleability abound. I focus on two of the best-established experimental paradigms for eliciting memory errors: the *DRM Paradigm* and *Loftus' Misinformation Paradigm*, which are the focus of philosophical Constructivists as well (e.g., De Brigard 2014a; Michaelian 2012). Results obtained by use of these paradigms illustrate the two key features of false memory echoed throughout the empirical investigation of remembering. First, attempts to recall a particular past experience often contain information from multiple sources, which results in inaccurate, distorted memories. These sources include not only the event in question, but other similar events, as well as the rememberer's background knowledge, cultural assumptions and expectations, and her aims and desires in a given context. Second, confidence and accuracy in memory are orthogonal. The feeling of remembering pulls apart from successful retrieval. When memory is distorted by other sources, this often goes unnoticed by the rememberer. People remain confident in their memories even as the details vary substantially across time.

The DRM Paradigm. The Deese-Roediger-McDermott (DRM) paradigm is one of the best-established techniques used to elicit false memories. In DRM studies, participants are presented with a set of similar items—e.g., a set of semantically related words like *nurse*, *sick*, *medicine*, *ill*, *clinic*, *patient*, *health*, etc. Later, they are asked whether they recognize certain items as members of the set presented previously. Participants do well at recognizing items from the original set and at rejecting items that are dissimilar from those

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

in the set (items like *clinic* and *judge*, respectively). Participants struggle, however, with items that are related to those in the original set but were not presented. For the set above, these would be items like *doctor*. Participants report recognizing related but not presented items at rates comparable to items that were in the original set. That is, they claim to recognize doctor as often as they claim to recognize *clinic* or *medicine* (Roediger and McDermott 1995). What's more, participants often insist that they remember hearing or seeing the non-presented item and are in some cases willing to provide additional details about what they were thinking when it was presented. The effect persists across variations in participant age and background, as well as type of stimuli, retention interval, and recall format. The error persists even when participants are warned to be vigilant against making such errors.[3] The tendency to "recognize" these non-presented items is so well-established that the DRM effect is now often used as a baseline measure against which the efficacy of other experimental manipulations can be tested.

Loftus' Misinformation Paradigm. Loftus has developed a similar set of misinformation studies, which show how susceptible the act of retrieval is to misleading information. In this paradigm, participants witness an event (often a video or re-enactment of a crime) and are then asked a series of questions about what they saw. In some cases, use of this paradigm reveals that eyewitness accounts can be manipulated easily by the language used to prompt recall. When participants watched a video of a car accident, for example, they reported different rates of speed for the car involved depending upon whether they were asked if it *hit*, *bumped*, or *smashed* the other car (Loftus and Palmer 1974). Further exploration of eyewitness reports reveals that participants fail to notice changes in the details of remembered events, claiming for instance to recognize a roadway scene where the stop sign had been replaced by a yield sign (Loftus, Miller, and Burns 1978). These memory distortions are not restricted to subtle or incidental features of the

[3] Gallo (2006) offers a thorough review of the DRM and its various permutations.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

events witnessed or observed. Participants often make substantial errors, such as misidentifying central actors, confusing the order of events, and in some cases even come to "remember" events that never happened, producing elaborate accounts of spilling punch at a family wedding or being lost in a shopping mall as a small child (Loftus and Pickrell 1995).

The memory errors produced in *DRM* and *Misinformation* studies are not easily dismissed as contrived, laboratory tasks.[4] The results are similar for studies of memory for significant personal and cultural events, like one's first job or the explosion of the *Space Shuttle Challenger* (Neisser and Harsch 1992). Memories of such events are often vivid and emotion-laden, and it is easy to assume that these features indicate veridicality. And yet, when people are asked to recall these events periodically over a number of years, studies show that the details of these retellings change over time in ways that are unrelated to the rememberer's confidence in the accuracy of their recall. Participants often add in details that were not part of the original experience, and in some cases remain confident in the veracity of these details even in light of contravening evidence (Paradis, Solomon, Florer, and Thompson 2004).

The accumulated evidence of these memory errors places pressure on the traditional, preservative account of memory.[5] It is not that the traditional view cannot explain memory errors. Even if memory's aim is preservation, as such views suggest, it may occasionally malfunction. The difficulty comes from the frequency and kinds of errors that are made. Constructivists believe that memory errors occur so regularly that any attempt to explain them away as occasional glitches in a preservative process will fail. Evidence

[4] It is worth noting that these results reflect performance *tendencies* across large groups of participants. Not everyone who engages in such tasks produces false memories, nor does everyone produce the same rates of error or fall prone to the same manipulations. Thanks to an anonymous reviewer for pressing this point.

[5] This does not mean that all experimentalists who conduct studies of these memory errors endorse Constructivism, although of course some do (e.g., Loftus, 2003). My focus is on contemporary philosophical versions of Constructivism, which take such evidence as motivation for their views.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

from the DRM and Misinformation Paradigms indicates that memory errors are pervasive, forcing proponents of the traditional account to say that memory malfunctions more often than it functions. As De Brigard explains, "saying that false and distorted memories are a failure of memory may force us to accept that we have a memory system that regularly and systematically malfunctions" (2014a: 159). Constructivists resist the conclusion that memory is inherently faulty, choosing instead to look for an alternative account of the nature and function of memory. The favored alternatives are inspired by the kinds of errors observed. The DRM and Misinformation paradigms show evidence of false memory for previous events—the rememberer's representations of particular past experiences. The "memories" produced feel, to the rememberer, like representations of particular events, but the content reported often includes information from multiple distinct events, as well general background knowledge, expectations and assumptions, etc. These blended representations provide insight into memory's underlying architecture. Information must be stored in a way that favors blended, malleable representations. If memory's structure is refashioned in this way—as a system designed to produce plausible representations of what could have happened during a past event, rather than a system designed to faithfully reproduce individual events—then the DRM and Loftus results no longer compel the view that memory is faulty. What once appeared to be errors are now recast as instances of the memory system functioning as it should. These Constructivist commitments are fleshed out further in the next section.

2.2 Constructivism: Central Commitments

In response to the accumulated evidence of memory errors described above Constructivists advocate a rethinking of both the architecture and process of memory. Here I present De Brigard's (2014a) *Episodic Simulation Theory* and use it as a guide to

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

identifying the two central commitments of Constructivism.[6] First, Constructivists advocate for a change in our understanding of memory's cognitive architecture. Memory does not store discrete representations of particular past events, as traditional views have supposed. Instead, memory relies on a more generalized network of information—one possibly shared with other capacities—that privileges patterns that emerge across a range of similar experiences. Second, Constructivists recommend a corresponding change in how the process of remembering is understood. Remembering is not an act of retrieval, but rather one of reconstruction. Memories are built, as needed, at the moment of recall. During construction, the rememberer makes use of any and all available sources of information to build a plausible account of what could have happened. Importantly, this reconstructive process is the same across all attempts at remembering, whether they result in accurate recall or error.

Memory's Cognitive Architecture. Constructivists argue that the evidence from the empirical study of memory errors illustrates memory's preference for patterns over particulars and gist over detail. As results from the DRM and Misinformation Paradigms show, attempts at remembering a particular past event combine information from multiple sources. These amalgamated recollections are understood as a reflection of memory's underlying architecture. Information from past events must be stored in a way that makes such blending and generalization possible. To this end, Constructivists reject the traditional assumption that memory comprises discrete representations of particular past events. That is, Constructivists reject the traditional understanding of memory traces as stored mental representations of particular past events.

In its place, Constructivists propose a less restrictive account of the storage that is required for memory traces, appealing to distributed networks, gist-based representations,

[6] There are many variants of Constructivism; where individual accounts disagree, my exposition below follows the commitments of De Brigard.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

dispositional states, and the like to explain how information is retained from past events.[7] De Brigard (2014a) favors a view centered around *schemas*, which correspond to the kinds of events and ideas that the person encounters frequently. He characterizes these schematic networks as an individual's "expertise," defined as the person's "relative frequency of exposure to a set of items" (p. 170). Many people, for example, have a restaurant schema—a generalized set of representations about what to expect when dining out. The features of this schema will differ along with an individual's dining habits and preferences. For some, celebratory dinners may involve white tablecloths and dim lighting. For others, it might mean eating with large groups in loud, brightly decorated spaces. Schemas provide the framework into which information from particular experiences is absorbed.

The network's structure shapes both the encoding and retrieval of information from the experience. Encoded events are used to update the schema, strengthening and fine-tuning its associations. The details of any particular event are of use only to the extent that they aid this process. A recent restaurant dinner may have involved servers wearing bowties rather than neckties, but the difference may go undetected by the network, which is far more accustomed to seeing the latter. De Brigard explains the confusion over stop signs and yield signs in Loftus' misinformation paradigm in this way (2014a: 172). Retrieval is, similarly, a process of using this network to construct a plausible representation of what could have happened during a particular event. Discussion of this point leads us to the second Constructivist commitment: the constructive process of remembering.

Process of Remembering. Since Constructivists deny the existence of discrete representations of particular past events in the memory store, there are no longer

[7] Versions of Constructivism can be distinguished by the specific distributed architecture endorsed. For a discussion of these variations, see Robins (forthcoming).

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

individual memory traces on hand waiting to be retrieved. And so the act of remembering can no longer be characterized as a process of retrieving such traces. Instead, it is a constructive process, whereby the information retained is used to build plausible representations of the past event. The act of remembering is "an inferential process, constructive not reproductive" (Sutton 1998: 219).

De Brigard's account of remembering is the most detailed but also the most radical, in that he denies the existence of memory as a distinct cognitive capacity. Memory is one of many cognitive abilities subsumed under a general cognitive system geared toward reasoning hypothetically about personal experiences. This earns his account of Constructivism its name: *Episodic Hypothetical Thinking*. The function of this general cognitive system is to create "self-referential mental simulations about what happened, may happen, and could have happened to oneself" (2014a p. 174–175). This system guides a person's consideration of his or her experiences, both past and future, real and imagined. The outputs of this hypothetical thinking system are governed by the patterns of expertise its schematic organization provides.

Remembering is thus only one way of consulting this general network. This network is used to generate inferences about the plausibility or likelihood of various scenarios, based on the patterns of expertise in the schematic network. Remembering is one way of consulting this network, taking the probability-based outcomes it generates as evidence of what *could have* or was *most likely to have* happened during the event one wants to recall. This explains the kinds of memory errors found using the DRM and misinformation paradigm, De Brigard claims. The patterns in the schematic network can lead to memories whose details have been altered, swapping more common features for the less. And similarly, by making use of a set of common features in a given schema, one can produce memories that are allegedly of one event, but actually combine details from several distinct events.

The generation of these false memories is no cause for alarm, however. Nothing in

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

the memory system, or larger episodic hypothetical reasoning system, has malfunctioned in the process. All attempts at remembering make use of the same inferential process. There is no deep, functional difference between successful remembering and misremembering. The two can be distinguished by a check against the facts, if evidence of what happened is available for the case in question. As De Brigard explains:

> Most of the time what you recall accurately depicts the witnessed event. Sometimes it does not. In both cases, however, the system is doing what it is supposed to do (2014a: 172).[8]

Appeal to a shared function for both instances of remembering and memory error is taken to explain another feature of the evidence from the DRM and Loftus' paradigms, namely, that confidence and accuracy in remembering are orthogonal. People should not be expected to be able to detect a difference between veridical and distorted memories, the Constructivist argues, because both are produced in the same way.

Not all Constructivists support the idea that memory is part of a larger reasoning system. But despite differences amongst Constructivists over how the details of this constructive process are understood, all philosophical Constructivists stress the similarity of the process by which accurate and false memories are produced. Sutton and Windhorst, for instance, claim that "veridical memories...are no less constructed than false memories" (2009: 87). In other words, it is constructive in all cases of remembering, both those that result in success and those that result in error.

[8] The claim that memory responds accurately most of the time may strike the reader as difficult to reconcile with De Brigard's other claim, quoted in §2.1, that memory "regularly and systematically malfunctions" (2014a: 159). Proponents of Constructivism must maintain a fine balance here, between claiming that memory follows certain patterns that are generally reliable and that the possibilities for error are pervasive enough to motivate this alternative account of memory. This is an interesting tension in the view, and I am grateful to an anonymous reviewer for highlighting it, but I do not explore it further in this paper.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

To summarize, the Constructivist claims that the process of remembering is one of building a representation of a past event to suit one's interests, needs, and expectations at the moment of recall. This general characterization of remembering exposes two central commitments of Constructivism, regarding memory's architecture and processing. First, memory does not rely on discrete representations of particular past events, instead uses some broader mechanism of information retention (in De Brigard's case, event schemas) to produce representations. Second, the process of remembering is one of constructing plausible inferences about what could have happened during past events. Even though the veracity of the content differs across the production of true and false memories, the process is constructive in both cases.

With this review of Constructivism's central commitments complete, I turn now to a review of how optogenetic manipulation has been used to produce false memories in non-human animals.

§3 The Optogenetic False Memory Technique (O-FaMe)

O-FaMe pairs established methods for identifying engrams in laboratory animals with the optogenetic technique for manipulating neurons via induced sensitivity to light.[9] In this section, I offer a brief summary of each of these methods (in 3.1 and 3.2, respectively) and then, in 3.3, I present the findings of two emblematic O-FaMe studies.

3.1 Engram Detection

The neurobiological study of memory is the study of engrams. An engram is the neurobiological mechanism by which information from previous experiences is encoded in the brain. Semon (1921) coined the term as part of his proposal that memory had a biological basis. Cellular and molecular neuroscience is now governed by *engram theory*,

[9] For the purposes of this paper, the terms "engram" and "memory trace" are being used interchangeably.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

the view that "when a memory is formed, a subpopulation of neurons will be excited and stay excited latently for the storage of the memory information (engram)" (Liu, Ramirez, and Tonegawa 2014: p. 1).[10] The transition from Semon's initial speculation to the theoretical foundation of cellular neuroscience is the result of a steady trajectory of experimental discovery across many areas and levels of neuroscience (Silva, Landreth, and Bickle 2014). The mechanistic details of memory formation are relatively well understood and often serve as a paradigmatic example of mechanistic explanation for philosophers of science.[11] Here I explain briefly the detection of engrams in non-human animals.

The search for the engram begins with the identification of a model system in a model organism—typically, a form of conditioning in a rodent species. The search for the engram of classical conditioning, for example, focused largely on eye-blink conditioning in rabbits (Thompson 2005). The O-FaMe technique of interest here uses contextual conditioning to fear and reward in mice (Ramirez et al. 2014). Contextual conditioning instills an associative memory for a novel environment. The mouse is first placed in a conditioning chamber. After initial exploration, the mouse is given either a positive or negative stimulus. Positive stimuli include food rewards and the opportunity to engage with a mouse of the opposite sex. The negative stimulus is often a foot shock, applied through the chamber's floor. When the mouse is later returned to the chamber, its behavior indicates memory of the previous (pleasant or unpleasant) experience in this context. Mice that received a positive stimulus will now actively explore the chamber, whereas mice that received a negative stimulus will now freeze (i.e., refrain from all voluntary movement).[12]

---

[10] It is an interesting to ask whether, in the neuroscience of memory, commitment to the existence of discrete memory traces is a pretheoretical commitment or empirical discovery. For a discussion of this issue, see De Brigard (2014b).

[11] Although, of course, philosophers of neuroscience disagree about the explanatory lessons to be drawn from consideration of this example. Bickle (2003) uses memory formation as an example of "ruthless" reduction, whereas Craver (2007) advocates multi-level mechanisms. For concerns about the explananda, see Sullivan (2010).

[12] Freezing is an adaptive response to fear, as predators are often sensitive to motion, and is found in most rodents.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

With the model system and organism identified, the search for the engram is then steered toward the area(s) of the brain known to be involved with the activity in question. For learning and memory, the primary region of interest is the *hippocampus*—a bilateral structure found in the medial temporal lobe. There is a well-established link between hippocampal damage and impairment of declarative memory in both humans and non-human animals (e.g., Squire and Zola 1996). This lesion data has guided decades of extensive research into the anatomical and physiological features of the hippocampus that support this memory mechanism. In a recent survey of the Library of Medicine, Silva and colleagues report the discovery of more than 110,000 scientific articles regarding the hippocampus (2014: 29). The hippocampus exhibits intricate patterns of connectivity, allowing for this brain region to be distinguished from others and for distinct subregions of the hippocampus to be identified as well. The engrams associated with the kind of memories produced in O-FaMe studies are found in the granule cells of the hippocampal dentate gyrus (Ramirez et al. 2014).

The final step of engram detection is the identification of the specific engram corresponding to a particular learning event. This requires knowing what to look for. In order for a neuron to encode information from an event, it must undergo significant modification. The modification will involve genetic activity on the part of the neuron. And so, encoding can be detected by identifying which neurons are engaged in transcription and translation processes immediately following the learning event. The neurons that initiate genetic activity just after the stimulus application are the engram for the contextual memory. Because engram encoding requires genetic modification, it is an ideal target for optogenetic manipulation, which I turn to in the next section.

3.2 Optogenetic Manipulation

Optogenetics is a new and exciting intervention technique in neuroscience. Its development makes good on a speculation offered by Francis Crick, namely, that the most promising, surgical interventions into neural circuits would be achieved by the use of light

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

(Crick 1999). Such intervention requires, first, finding a way to make neurons light-sensitive. This is done by use of light-sensitive proteins, or *opsins*, which respond to particular wavelengths of light. The protein first used for optogenetic manipulation is Channelrhodopsin-2 (ChR2), a membrane protein found in algae (Boyden 2011). ChR2 is a light-sensitive ion channel—when exposed to blue light, the channel opens. Opsins like ChR2 are genetically encoded, making it possible for this protein code to be spliced with the regulatory portions of genes from another organism—mice, in the O-FaMe studies of interest here, but also flies and other mammals. This hybrid gene can then be introduced into a particular type of neuron so that the opsin is expressed, thereby rendering that subset of brain cells light-sensitive. When exposed to blue light, these neurons will now generate an action potential.[13] Other identified opsins, like halorhodopsin (NpHR), can be used to inhibit cell activity, suppressing rather than promoting action potentials (by making the targeted cells sensitive to yellow light). Together, these excitatory and inhibitory interventions allow for precise manipulation and control of neural activity in intact systems and living organisms (Häusser, 2014).

In the decade since the publication of the first paper employing optogenetics (Boyden, Zhang, Bamberg, Nagel, and Deisseroth 2005) the method has received numerous accolades. *Science* declared it one of the Breakthroughs of the Decade and it was awarded Method of the Year in in 2010 by *Nature Methods*. Optogenetics recently became a central tool for the U.S. *National Institutes of Health's* BRAIN Initiative and the method's six developers were awarded the 2013 Brain Prize. Optogenetic manipulation has captured the interest of neuroscientists because of the ways that it allows for real-time control of the behavior of highly particular sets of neurons in living, behaving organisms (Craver forthcoming).[14] The method has been used to explore a range of research questions in neuroscience, from perception to Parkinson's (Fenno, Yizhar, and Deisseroth 2011).

---

[13] For a recent and thorough review of optogenetic techniques, see Deisseroth (2011).
[14] This is not to say that the method is without its limitations (Häusser 2014). Some even recommend more focus on alternative molecular interventions, such as Designer Receptors Exclusively Activated by Designer Drugs (DREADDs), and the relative experimental advantages and disadvantages of each (Bickle in prep).

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

For our purposes here, what is particularly exciting about optogenetic manipulation is that it offers a way to reactivate memories while circumventing the standard route to retrieval in non-human animals. That is, optogenetics allows researchers to sidestep what had previously appeared to be an intractable difference between studies of human and non-human animal memory. In human memory studies, asking participants to retrieve a particular memory is relatively straightforward—the experimenter asks the participant to call the desired event to mind. There is no analogous way to make this request of non-human animals. In order to reactivate a contextual memory, the animal is returned to the original, remembered context. Optogenetics provides a way around this hurdle. If the engram is encoded by neurons with engineered opsins, then the engram—and its concomitant memory—can be reactivated with the application of light. As we will see, the ability to reactivate the engram outside of the original context is what makes possible the creation of false memories in the mice. I explore the details of this technique in the next section, through a discussion of two recent studies.

3.3 O-FaMe Studies

Below I discuss a set of findings from an extensive research project based in the Tonegawa Laboratory at the Massachusetts Institute of Technology (http://tonegawalab.org/), which applies optogenetic techniques to the study of learning and memory. My focus below is on two studies that use a contextual conditioning technique to create false memories in mice.

As with other uses of optogenetics, this research program relies on the creation of a well-specified transgenic population—here, a set of genetically engineered mice. These mice have three especially important features. [15] First, they possess the ChR2 transgene,

[15] The transgenic population used by the Tonegawa laboratory for the O-FaMe studies discussed below are c-fos/tTA/Dox-off mice. In addition to the features discussed in the text, these mice are also engineered to have: 1) A Tetracycline-Responsive Element (TRE), which provides a binding site for the protein that allows expression of the engineered opsin gene), and 2) Monomeric fluorescent protein gene (mCherry), which expresses a protein that appears red under standard light microscopy, allowing the engram cells to be

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

discussed in §3.2, so that the light-responsive ChR2 protein will be expressed when neurons with this transgene are active. Second, these mice are a *dox-off* variety: the ChR2 transgene will only be expressed when the animal is *not* exposed to doxycycline (Dox), an antibiotic applied through its water supply. This confers control over when the light-sensitive proteins are expressed and subsequently translated (i.e., when the mice form light-sensitive engrams). The mice are given Dox until the experimental condition begins, and then returned to Dox immediately after, so that the only light-responsive engrams the mouse has will be those formed during the experimental condition. Third, the mice are given an optical fiber implant, so that the light-sensitive proteins can later be activated by turning on this (blue) light.

In an initial study, the Tonegawa group demonstrated the ability to contextually condition these mice, identify the resultant engrams, and then—most importantly—use optogenetic intervention to reactivate the engram and produce a behavioral expression of the memory (Liu et al. 2012).[16] The two O-FaMe studies discussed below build on this result, pairing the activated engram with additional, misleading information before testing behavioral expression. The first, Ramirez et al. (2013), adds valence to a previously neutral memory and the second, Redondo et al. (2014), reverses the valence of a memory. I discuss these experiments in turn. Throughout this section, I refer to the memories involved as *contextual*—the mice form and retain a memory of an encounter with a particular context or environment. Discussion of the similarity between these memories and human false memories is withheld until §4.1.

Ramirez et al (2013). In this experiment, mice learn to fear a context that they have encountered before, but which contained no fearful stimuli during the initial encounter. The false memory is produced via a two-step process. First, the transgenic mice are taken

---

detected by researchers. Many thanks to [name withheld for purposes of blind review] for helping me understand the details of this mechanism.

[16] The lineage of O-FaMe, and inquiries into the possibility of animal models of false memory, can be traced back further, to studies that manipulate which neurons are involved in the engram by interventions into the CREB transcription factor (Han, Kushner, Yiu, Cole, Matynia, Brown, et al. 2007).

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

off Dox and each is introduced to a novel conditioning chamber—*Context A*. The mouse is allowed to explore its environment and, since Dox is no longer inhibiting transgene expression, the resultant engram for this contextual memory comprises neurons with the light-sensitive ChR2 protein. The mouse is then removed from Context A and Dox is reintroduced to the diet. Second, the mouse is introduced to another novel conditioning chamber—*Context B*—and, while here, the optical implant is turned on. This light reactivates the engram from Context A via the light-sensitive proteins of its neurons. While the Context A engram is active in Context B, the mouse is given a set of foot shocks sufficient to instill a fear memory for Context B.

As a result of this two-step process, each mouse now has a light-sensitive engram that has been activated twice, but paired with a negative stimulus only once. The mice are then tested in three conditions:

1) Returned to Context A
2) Returned to Context B
3) Introduced to Context C, a novel conditioning chamber

When mice are returned to Context B, where they previously received foot shocks (Condition 2), they display typical fear behavior—freezing in place. When returned to Context A (Condition 1), mice display the same freezing, fear behavior, even though they were not exposed to any fearful stimuli in this environment. It is tempting to infer that the previous foot shocks have made the mice generally fearful so that they will exhibit fear behavior in any context. Condition 3 was included to test for this possibility. The results tell against the 'generally fearful' interpretation: when placed in Context C the mice explore the chamber in ways characteristic of exposure to a new environment.

Together, the results of Conditions 2 and 3 indicate the mice form a fear memory for Context B. The mice freeze in this context, but do not freeze in all contexts. The results of Condition 1 indicate the formation of a false memory for Context A: mice respond to this environment as familiar or remembered, but behave in a way that does not reflect their previous experience in this context.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

Redondo et al (2014). This study expands on the Ramirez et al. (2013) results, using the O-FaMe to create false memories by switching the engram's valence—from positive to negative and vice versa.

Redondo and colleagues created false memories in mice using the same two-step process described above, with slight modifications to each step. When the mice were introduced to the first conditioning chamber (Context A), the mouse's initial exploration was paired with either a positive or negative stimulus. The mouse was either given exposure to a female mouse (a positive stimulus for the experimental mice, which were male) or foot shocks (a negative stimulus). The mice were taken off dox during their time in Context A, so the ChR2 transgene was expressed and the protein was synthesized, creating a light-sensitive engram. Next, when the mice were transferred to Context B and the Context A engram was reactivated, the mice were given an addition stimulus that was either consistent or inconsistent with the stimulus received in Context A. That is, for the mice that were fear-conditioned in Context A, half received additional foot shocks in Context B (consistent) and half received exposure to a female mouse in Context B (inconsistent).[17] And similarly for the mice who were reward-conditioned in Context A. The result is four groups of mice: Consistent-Fear, Consistent-Reward, Inconsistent-Fear, and Inconsistent-Reward. For present purposes, our interest is in the behavior of mice who received different stimuli across Contexts A and B—mice from the Inconsistent-Fear and Inconsistent-Reward Groups. These mice have a light-responsive engram that was activated twice, but that was paired with distinct stimuli each time. What happens when mice from these groups are tested in the three retrieval conditions? When returned to Context B, the mice display behavior consistent with the stimulus they received in this context. If they previously received foot shocks in Context B, they freeze in Context B; if

[17] Another portion of the Redondo et al. study involved a comparison of encoding-related changes in the amygdala versus the dentate gyrus. This portion of the results confirms that it is the neural changes in the dentate gyrus, not the amygdala, that are responsible for engram formation. For this reason, I do not include further discussion of this portion of the study here.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

they previously encountered a female mouse in Context B, they actively explore Context B. When returned to Context A, the mice display the same behavior, which is inconsistent with the stimulus received in this context. That is, mice that were initially conditioned to fear Context A now display exploratory behavior in Context A, and, conversely, mice that were initially reward-conditioned now display fear behavior when returned to this context. These results indicate that the valence of the original engram has been changed, resulting in behavior that contradicts their previous experience in this environment.

The O-FaMe studies presented in this section offer examples of distorted, false memories in mice. The mice treat a previously experienced environment as familiar, but as a result of optogenetic manipulation of the involved memory, behave in ways that misrepresent their previous experience. I turn now to the question of whether Constructivist theories of memory can accommodate these results.

§ 4 Constructivism after O-FaMe

The O-FaMe studies discussed in §3 produce what appears to be an animal model of false memory. Because of the detailed interventions optogenetic manipulation makes possible, discovery of these false memories is accompanied by insight into the mechanisms by which they are produced. We are now ready to explore the question at the heart of this paper: How does Constructivism fare in light of the results of O-Fame? I begin in §4.1 with a defense of O-FaMe as an animal model of false memory, arguing that the resultant behaviors are best understood as misremembering errors. In §4.2, I go on to identify two tensions between De Brigard's Constructivist commitments and O-FaMe and then use these observations to draw attention to a more general tension between cognitive and neurobiological approaches to memory.

4.1 O-FaMe as Misremembering

The first point to address is whether the results of the O-FaMe studies are aptly characterized as a non-human animal model of false memory. Importantly, the researchers who conduct these studies understand their results in this way:

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

> Although our design for the formation and expression of a false memory was for a laboratory setting, and the retrieval of the contextual memory during conditioning occurred by artificial means (light), we speculate that the formation of at least some false memories in humans may occur in natural settings through the internally driven retrieval of a previously formed memory and its association with concurrent external stimuli of high valence (Ramirez et al. 2013: 390).

What's more, the Tonegawa research group references the DRM and Loftus paradigms, drawing explicit comparison between their results and the experimental paradigms that have been the centerpiece of Constructivist theorizing (*Ibid*.). Of course these claims should not be taken at face value—and aside from the remark above, the researchers do not further elaborate on or defend the comparison. Our exploration of the similarity between these results must go further.

There is room for concern over the depths of similarity between the memory errors displayed in human memory experiments and those obtained from use of the O-FaMe technique with mice. One might, for example, have concerns about the very possibility that mice and other non-human animals possess the kinds of memories that could be false or distorted. The cases of false memory discussed in §2, which have captured both popular and academic attention, involve memories standardly labeled as *episodic*—memories for particular past events that, in human cases, are often richly detailed with elaborate phenomenology. If episodic memory is a uniquely human capacity, then this would obviate questions of its potential distortion in non-human animals.

Concerns about the possibility of episodic memory in non-human animals should not forestall consideration of O-FaMe studies as an animal model of false memory. First, whether non-human animals are capable of episodic memory is a matter of ongoing controversy. While some continue to claim that animals other than us lack the kind of self-knowledge and autonoetic consciousness required for episodic remembering (e.g., Tulving 2005), there are many others who are happy to grant episodic—or at least episodic-like—memory to various non-human animals, from scrub jays to chimpanzees (Templer and Hampton 2013). Hasselmo (2012) has proposed recently an account of episodic memory

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

meant to explain the capacity in both human and non-human cases, based in large part on studies of maze-running in mice. Second, even if one denies mice and other non-human animals possession of the richest forms of episodic memory, this need not preclude consideration of O-FaMe results as instances of false memories of a more mundane sort. Many cases of human false memory fail to exhibit the rich phenomenal character often taken as definitive of episodic remembering. Consider the examples of false memory from the DRM Paradigm introduced in §2. In these studies, participants report having heard or seen a word that was not on a previously presented list. The memory involved is of a particular past episode, but recalling a list of words may not otherwise contain any richer sense of mental time travel into one's past that is often associated with episodic remembering. Such false memories may be better characterized as distortions of *event memory*, which involves scene reconstructions of past experiences (Rubin and Umanath 2015). The contextual memories displayed by mice in the O-FaMe studies are plausibly construed as event memories, even if one wants to resist describing these or any other non-human animal memories as episodic. Mice in the O-FaMe studies remember particular encounters with particular environments. If at least some false memories are event memories, and some non-human animals are capable of event memory, then deep skepticism about the possibility of discovering an animal model of false memory can be set aside.

I want to go further, however, and argue that O-FaMe and the experimental paradigms that have been used to elicit false memories in humans are sufficiently similar, warranting consideration of O-FaMe as an animal model of false memory. Pressing on is critical for the coming argument, namely, that philosophical Constructivists who focus on results from cognitive psychology and cognitive neuroscience lack the resources to explain O-FaMe results. Exposing the similarity between O-FaMe and other false memory paradigms is critical for identifying the tension between the approaches to memory from neurobiology and higher level cognitive neuroscience. The similarities, I will argue, are both methodological and behavioral.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

First, the methodology. Each paradigm makes use of a shared general technique for eliciting false memories. A memory for a particular past event is created and then paired with information that is similar enough to induce distortion either before or during recall. This is clearest in the Loftus' Misinformation paradigm. In the Loftus et al. (1978) study, participants were shown a series of images depicting a car accident. Then, participants filled out a questionnaire about the accident they had just witnessed. For some, the questionnaire involved misleading information. Specifically, they were asked, "Did another car pass the red Datsun while it was stopped at the stop sign?" (1978: 22) when the previously viewed scene had depicted a yield sign. This misinformation was sufficient to produce a false memory in these participants: when shown a set of images some time later, participants selected photos including the stop sign as having been in the initial set. In other cases, the misinformation is not presented directly to the participant. Instead, the prompt used to solicit recall is selected because its similarity to the original event is likely to produce distortion. In the same study just described, Loftus and colleagues found that some participants would claim to recognize the accident photos including a stop sign even when they had not received misleading information in the questionnaire.[18] And similarly, the DRM paradigm asks participants whether they recognize words that are highly similar to, but were not part of, the previously presented list of words. The items are selected as 'critical lures,' whose similarity encourages distortion of the participant's memory for the prior learning event.

O-FaMe studies make use of the same method. The experiments begin by instilling a contextual memory for a particular past experience: the mouse's encounter with a novel environment. Then, the mouse is made to recall that past experience by activating the corresponding light-sensitive engram while the mouse is given additional, misleading contextual information. Specifically, the memory for the original context is paired with an experience that either adds or changes the valence of what was previously experienced.

[18] The rates of false recognition were lower, however. Only 25% of participants "recognized" the photo without the misleading question on the questionnaire, whereas 60% "recognized" it when the misleading question was included.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

The similarity between O-FaMe and the Loftus and DRM paradigms is unsurprising, given that the O-FaMe researchers were explicit about their aim to produce a similar phenomenon (Ramirez et al. 2013: 390).

Consideration of O-FaMe as an animal model of false memory is strengthened by the fact that use of a similar method to Loftus and DRM produces results that are behaviorally similar as well. Although these errors are often referred to as *false* memories in the literature, a practice I too have adopted up to this point, the terminology is misleading. The memories produced are not *entirely* false. Importantly, the distortions exhibited rely on the participant remembering some information about the event in question. For this reason, Robins (in press) has argued that the Loftus and DRM results should be understood as cases of *misremembering*: errors that rely on successful retention of the targeted event. When a person misremembers, her report is inaccurate and yet the error is explicable only on the assumption that she *has* retained information from the event that her representation mischaracterizes.

In the Loftus misinformation paradigm, the experiments are designed to distort the participant's memory of the car accident, which can only happen if information from the event has been retained. A participant cannot be led to misremember the speed with which the car was traveling before the accident unless she remembers seeing the car. And similarly, when a participant claims that an accident involved a stop sign rather than a yield sign, she is making an error that relies upon her remembering that the accident occurred at an intersection with a road sign. And similarly for DRM experiments: here participants falsely "recognize" items that are similar to those from a previous set, an error they can only make if they remember the types of items that were in that set. Consider the version of this experiment discussed in §2. The participant sees a list of words: *nurse, sick*, *medicine*, *ill*, *clinic*, *patient*, *health,* etc. To claim that *doctor* was on the list while denying the same for *judge* indicates the participant's recollection that the list items were *doctor*-related, even if she has erred in remembering the specific items listed.

O-FaMe studies produce comparable results. The participants are mice; it does not make sense to characterize the non-verbal memories elicited in terms of

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

accuracy/inaccuracy. Nonetheless they can be understood as appropriate/inappropriate given past conditioning. In the Ramirez et al. (2013) study, mice respond inappropriately when returned to Context A. Mice fear this context, even though they never experienced fearful stimuli in this context. This inappropriate response relies on retained information about that past event. The mice can only be misled into expressing fear in Context A if they have retained information about this context—namely, that they have been there before. Similarly, in the Redondo et al. (2014) study, it is inappropriate for the mouse to avoid the context where it previously received a reward or to explore a context where it previously received shocks. In both cases, the mouse's inappropriate response is dependent upon its retention of information about the context that it now misrepresents. Crucially, the environment has to be treated as familiar in order to be feared or explored, as the behavior does not extend to other, novel contexts. The behavioral similarity between the results of O-FaMe and those of the Loftus and DRM paradigms suggests that O-FaMe errors should be understood as cases of misremembering, too.

One might question my characterization of the mouse's behavior as *inappropriate* in the O-FaMe studies. The mouse's response is, in one sense, entirely appropriate given its full conditioning background. The associative conditioning has worked; the mouse has learned to associate its memory of Context A with painful foot shocks and behaves in kind.[19] The point is worth noting. Doing so, however, only serves to strengthen the similarity between the human cases and O-FaMe. The experimental manipulation is successful in each of these paradigms because it exploits the similarity between what is learned and what is distorted. And further, human cases of misremembering are also understandable and thus difficult to characterize as fully inappropriate or even inaccurate. Stop and yield signs look similar, and it may be beneficial to take note of the fact that all of the words on a given list are doctor related. This behavior may only be seen as an error from the perspective of the experimenter, whose interests are focused exclusively on the learning event and subsequent manipulation. From the participant's perspective, the similarities between this event and other events and background information may be

[19] I am grateful to [name withheld for purposes of blind review] for raising this issue.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

equally or more salient. Such observations serve as motivation for the Constructivists' rethinking of memory's function.

O-FaMe differs from human studies of misremembering in one important respect: the manipulation is carried out at the cellular level. Such low-level intervention not only confers more precision and control over the memory's formation, reactivation, and distortion than is possible in human cases, it also provides a first glimpse into the mechanism by which these memories are produced. I turn to the lessons that can be learned from this glimpse in the next section.

4.2 Tensions Between Constructivism and O-FaMe

I have just argued that, as with the Loftus and DRM paradigms, the results of O-FaMe should be understood as cases of misremembering. Given that philosophical Constructivists like De Brigard purport to explain the errors that occur in use of the human paradigms, we should expect the view to explain the errors that occur in O-FaMe as well. And yet, O-FaMe studies reveal features of the mechanism by which memory errors are produced that are in tension with much of Constructivist theorizing. In what follows, I identify two tensions between the Constructivist commitments outlined in §2.2 and O-FaMe and then use these observations to draw attention to a more general tension between cognitive and neurobiological approaches to memory.

Recall that Constructivists characterize memory as relying on retained information that is distributed, blended, or schematized, from which plausible representations of what could have happened during a past event are built at the time of recall. This account of memory involves two important claims. First, memory does not store discrete representations of particular events. Second, all attempts at remembering make use of a constructive process, whether they result in accurate recall or error.

O-FaMe results challenge both of these claims. The misremembering errors in O-FaMe are produced by reactivating an engram from a particular past event and then pairing it with additional, misleading information. The Liu et al. (2012) study provides the initial

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

proof of concept: an engram can be tracked from its formation and then reactivated (via optogenetic intervention) to induce remembering. The Ramirez et al. (2013) and Redondo et al. (2014) studies build on this result, adding distortion to the reactivation so as to induce misremembering. The mechanistic details of O-FaMe fit well with the characterization of misremembering as a memory error that relies on successful retention of a particular past event. If misremembering is an interaction between the retention of an engram or memory trace from a particular event and information available at retrieval, then it makes sense that the mechanism by which they are created involves reactivating and then altering a retained engram.

The first point of tension with Constructivism should be readily apparent: Constructivists deny that engrams, or memory traces of particular past events, are retained. Constructivists do not deny that memory involves the retention of information. They do reject, however, the proposed structure of retention upon which this account of misremembering relies. There are differences amongst Constructivists in the alternative characterization of retention preferred. In De Brigard's case, memory is output from a general, distributed store of information, one that is shared with other capacities like counterfactual reasoning and imagining the future. There is no place for discrete representations of particular past events. The assumption is that there is no need for such discrete traces, as memory errors can be accommodated more efficiently without them. Robins (forthcoming) challenges this claim in the case of human memory errors. Philosophical Constructivists are right to note that the empirical evidence shows that memory for particular events can be influenced by information from other sources. But these other sources are not the only—nor the most significant—influence on the errors produced. The primary source is *information retained from the particular past event*, stored discretely as an engram or trace. This information may be distorted by factors added during a future event or by cues used at retrieval, but the signal from this discrete event is necessary to explain the particular error produced. The addition of O-FaMe results provides further support for this account of misremembering. Misrememberings are

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

distortions or manipulations of what is retained, not cases where a new representation is constructed from the current patterns, schemas, and trends in a generalized network.

The second point of tension comes from the explanation of *how* distorting information influences the process of remembering. Constructivists construe the effects of misinformation generally; they play the same role in all constructed memories, whether the resultant memories are accurate or not. For De Brigard, the shared process is use of one's expertise to reason about what was likely to have happened during the event in question. This guides his explanation of Loftus' misinformation studies. If participants have encountered stop signs more than they have encountered yield signs, then stop signs will feature in their reconstructions of events that take place at intersections, both when this actually occurred and when it did not. In both cases, the system makes the same prediction (2014a: 172). The misinformation is poorly characterized as *mis*information; it is simply general information about what is most likely or most frequent, applied equivalently in both cases.

This is not how misinformation gets a grip in O-FaMe studies. One can hope that, life as a laboratory animal notwithstanding, receiving foot shocks is not the most likely occurrence for a mouse entering a novel context. But even if it is, the mouse is not making use of this general likelihood in its behavioral response to the contexts it is introduced to in the O-FaMe experiments. Consider Ramirez et al. (2013). The foot shocks applied in Context B, while the engram for Context A was activated, influence the mouse's response to re-encountering both of these contexts. The aversive conditioning does not influence the mouse's response to the novel context C, as would be expected if a general expectation about likelihoods was driving the behavior.[20] The misinformation is being paired with a particular engram, neatly circumscribing its effects. This may not always be the route

[20] To make this point more forcefully, O-FaMe studies should include a fourth condition, where the mouse is taken to a familiar, neutral context. If the mouse treats the environment as familiar, but does not freeze, then it would be clear that the misinformation has not spread to all remembered contexts. To my knowledge, no O-FaMe study has yet included such a condition.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

through which misinformation has its influence. The point is that it can sometimes occur in this way, a possibility Constructivism is not well situated to accommodate.

Identifying these points as *tensions* between Constructivism and O-FaMe is important. Constructivism has focused primarily on explaining human memory error. It would be unfair to argue for the view's rejection on the grounds of its inability to accommodate the unanticipated arrival of O-FaMe. What's more, it is still early days for O-FaMe. Optogenetics in general and O-FaMe in particular have produced striking discoveries quickly, giving good reason to expect the success to continue. But it would be a mistake to assume that the current batch of O-FaMe results provide complete or definitive understanding of the mechanisms of false memory, even in mice. Still, the disparities between Constructivism's explanatory approach and O-FaMe results are serious enough to motivate a re-consideration of the central tenets around which philosophical Constructivism has been built up to this point.

Exploration of these difficulties also serves to direct our attention to a larger tension that has emerged between cognitive and neurobiological approaches to memory. The tension concerns the need for engrams, or memory traces, in the study of remembering. Commitment to the existence of memory traces was once central to scientific theorizing about this capacity. Even Karl Lashley, famous for conducting experiments that led him to deny the existence of the engram, felt compelled to acknowledge their essential role in remembering, writing: "I sometimes feel, in reviewing the evidence on the localization of the memory trace, that the necessary conclusion is that learning just is not possible" (Lashley 1950). At the cognitive level, explanations of memory grow ever more distant from any commitment to memory traces, at least from any traditional view of traces as discrete, well-preserved entities. Memory is a radically reconstructive endeavor, more akin to imagination than preservation. It is a dynamic process, with little need for static traces (e.g., Looren de Jong and Schouten 2005, but see also Bickle 2005 in reply). Philosophical Constructivists find support for these claims from systems level neuroscience (e.g., De Brigard et al. 2013). These cries for revolution have had little influence on low-level neuroscience, where the search for engrams continues in earnest. As noted in §3, engram

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

theory still guides the study of memory at the cellular and molecular level. Low-level neuroscientists explore the process of reconsolidation and various ways that memory traces might be updated over time, and as quoted in §3.1, these traces may be described as "excitedly latently" (Liu et al., 2014: p. 1) rather than static, but the commitment to their existence remains firm.

The growing disconnect between cognitive and neurobiological approaches to memory may have gone unnoticed because memory has been, historically, one of the best examples of inter-level integration in cognitive science. Or, if it has been noticed, the apparent tension has been explained away because the false memories that have provoked the disconnect are presumed to be exclusively human. O-FaMe results are thus exciting, if for no other reason than the trouble they make for such quick dismissals. Evidence of the role of engrams in the production of memory errors may require an account of memory that incorporates more from the traditional warehouse model of memory than current versions of Constructivism now suppose. The preponderance of memory errors surely makes accounts of memory focused entirely on preservation difficult to defend. But this need not be seen as warrant for jettisoning the commitment to discrete traces altogether. O-FaMe results encourage a search for more intermediary options.

§ 5 Conclusion

O-FaMe provides the first non-human animal model of false memory, challenging the assumption that such memories are exclusively human and providing insight into the mechanism by which such errors are produced. As I have argued here, O-FaMe studies are similar to the standard paradigms used to elicit false memories in cognitive psychology, both in method and results, and thus should be incorporated into our best theories of how and why such errors are produced. Constructivism has served as the best account of false memory to date, reconfiguring our understanding of the capacity to remember so as to make the propensity to memory errors understandable, and even beneficial, for

rememberers. Adding O-FaMe results to the data that must be explained exposes potential weaknesses in the Constructivist approach to explaining false memory. Specifically, O-FaMe studies suggest that false memories are the result of an interaction between the memory trace (or engram) and additional, misleading information and that there are mechanistic differences in the production of successful memories and various kinds of memory error. Constructivist theorizing, up until this point, has been in tension with both of these claims. As research into animal models of false memory grows, using O-FaMe and other methods, rethinking of Constructivist commitments may be required.

References

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Drodrecht: Kluwer Academic Publishing.

Bickle, J. (2005). Molecular neuroscience to my rescue (again): Reply to Looren de Jong and Schouten*. Philosophical Psychology, 18*, 487–494.

Bickle, J. (in prep). Laser lights and designer drugs: The new faces of ruthlessly reductionistic neuroscience.

Boyden, E.S. (2011). A history of optogenetics: the development of tools for controlling brain circuits with light. *F1000 Biological Reports*, 3, 1–12.

Boyden, E.S., Zhang, F., Bamberg, E. Nagel, G. & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience, 8*, 1263–1268.

Craver, C.F. (2007). *Explaining the Brain*. Oxford: Oxford University Press.

Craver, C.F. (forthcoming). Thinking about interventions: optogenetics and makers knowledge of the brain. In K. Waters (Ed.) *Causation in Biology and Philosophy*. Minnesota Studies in Philosophy of Science, University of Minnesota Press.

Crick, F. (1999). The impact of molecular biology on neuroscience. *Philosophical Transactions of the Royal Society of London, Series B*, *354*, 2021–2025.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

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

De Brigard, F. (2014b). The nature of memory traces. *Philosophy Compass, 9*, 402–414.

De Brigard, F., Addis, D., Ford, J.H., Schacter, D.L., & Giovanello, K.S. (2013). Remembering what could have happened: neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*, 2401–2414.

Deisseroth, K. (2011). Optogenetics. *Nature Methods*, 8, 26–29.

Draaisma, D. (2000). *Metaphors of Memory: a history of ideas about the mind*. Cambridge, UK: Cambridge University Press.

Fenno, L., Yizhar, O., & Deisseroth, K. (2011). The development and application of optogenetics. *Annual Review of Neuroscience, 32,* 389–412.

Gallo, D. A. (2006). *Associative Illusions of Memory: False Memory Research in DRM and Related Tasks*. New York, NY: Taylor & Francis.

Han, J., Kushner, S.A., Yiu, A.P., Cole, C.J., Matynia, A., Brown, R.A., Neve, R.L., Guzowski, J.F., Silva, A.J., & Josselyn, S.A. (2007). Neuronal competition and selection during memory formation. *Science, 316*, 457–460.

Hasselmo, M.E. (2012) How We Remember: Brain Mechanisms of Episodic Memory. MIT Press: Cambridge, MA.

Häusser, M. (2014). Optogenetics: the age of light. *Nature Methods, 11*, 1012–1014.

Klein, S. (2013). The temporal orientation of memory: It's time for a change of direction. *Journal of Research in Applied Memory and Cognition, 2*, 222–234.

Lashley, K. (1950). In search of the engram. Symposium on Explorations of Biology, No. 4. (pp. 454–482). Cambridge, UK: Cambridge Univ. Press.

Liu X., Ramirez, S., Pang, P., Puryear, C., Govindarajan, A., Deisseroth, K., & Tonegawa S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature,* 484, 381–385.

Liu, X., Ramirez, S., & Tonegawa, S. (2014). Inception of a false memory by optogenetic manipulation of a hippocampal memory engram. *Philosophical Transactions of the Royal Society B*, 369, 2013–2042.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

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

Lofuts, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology, 4*, 19–31.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of auto-mobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585–589.

Loftus, E. F. & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720–725.

Looren de Jong, H., & Schouten, M.K.D. (2005). Ruthless reductionism: a review essay of John Bickle's "Philosophy and Neuroscience." *Philosophical Psychology, 18,* 473–486.

Michaelian, K. (2012). Generative memory. *Philosophical Psychology, 24*, 323–342.

Neisser, U. (1967). *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts.

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd & U. Neisser (Eds.), *Affect and Accuracy in Recall: Studies of Flashbulb Memories* (pp. 9–31). Cambridge: Cambridge University Press.

Paradis, C. M., Solomon, L. Z., Florer, F. & Thompson, T. (2004). Flashbulb memories of personal events of 9/11 and the day after for a sample of New York City residents. *Psychological Reports, 95*, 304–310.

Ramirez, S., Liu, X, Lin, P., Suh, J., Pignatelli, M., Redondo, R.L., Ryan, T.J., Tonegawa, S. (2013). Creating a false memory in the hippocampus. *Science*, *341*, 388–391.

Ramirez, S., Tonegawa, S., Liu, X. (2014). Identification and optogenetic manipulation of memory engrams in the hippocampus. *Frontiers of Behavioral Neuroscience, 7*, *226*.

Redondo, R.L., Kim, J., Arons, A.L., Ramirez, S., Liu, X., & Tonegawa, S. (2014). Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature, 513*, 426–430.

Robins, S.K. (forthcoming). Misremembering. *Philosophical Psychology.*

Roediger, H. L., and McDermott, K. B. (1995). Creating false memories: Remembering words that were not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Robins, S.K. (2016). Optogenetics and the Mechanism of False Memory. *Synthese, 193,* 1561–1583. doi:10.1007/s11229-016-1045-9

Rubin, D.C., & Umanath, S. (2015). Event memory: a theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review, 122*, 1–23.

Schacter, D. L. & Addis, D. R. (2007). On the constructive episodic simulation of past and future events. *Behavioral & Brain Sciences, 30*, 299–351.

Semon, R. (1921). *The Mneme*. London: George Allen & Unwin.

Silva, A.J., Landreth, A., & Bickle, J. (2014). *Engineering the Next Revolution in Neuroscience*. New York: Oxford University Press.

Squire, L.R., & Zola, S.M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences, 93,* 13515-13522.

Sullivan, J. (2010). Reconsidering Spatial Memory and the Morris Water Maze. *Synthese, 177*, 261–283.

Sutton, J. (2007). *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge, UK: Cambridge University Press.

Sutton, J., and Windhorst, C. (2009). Extended and Constructive Remembering: Two Notes on Martin and Deutscher. *Crossroads: An Interdisciplinary Journal for the Study of History, Philosophy, Religion, and Classics, 4*, 79–91.

Templer, V.L., and Hampton, R.R. (2013). Episodic memory in nonhuman animals. *Current Biology, 23*, R801-R806.

Thompson, R. F. (2005). In search of memory traces. *Annual Review of Psychology, 56*, 1–23.

Tulving, E. (2005). Episodic memory and autonoesis: uniquely human? In *The Missing Link in Cognition* (H.S. Terrace and J. Metcalfe, Eds.). Oxford: Oxford University Press (pp. 3–56).