Philosophy Compass 9/6 (2014): 402–414, 10.1111/phc3.12133

The Nature of Memory Traces

Felipe De Brigard*
Duke University

Abstract

Memory trace was originally a philosophical term used to explain the phenomenon of remembering. Once debated by Plato, Aristotle, and Zeno of Citium, the notion seems more recently to have become the exclusive province of cognitive psychologists and neuroscientists. Nonetheless, this modern appropriation should not deter philosophers from thinking carefully about the nature of memory traces. On the contrary, scientific research on the nature of memory traces can rekindle philosopher's interest on this notion. With that general aim in mind, the present paper has three specific goals. First, it attempts to chart the most relevant philosophical views on the nature of memory traces from both a thematic and historical perspective. Second, it reviews critical findings in the psychology and the neuroscience of memory traces. Finally, it explains how such results lend support to or discredit specific philosophical positions on the nature of memory traces. This paper also touches upon the issues raised by recent empirical research that theories of memory traces need to accommodate in order to succeed.

1. The Need for Memory Traces

In the philosophy of memory, the view according to which remembering is mediated by representations whose contents represent past intentional objects is known as *representationalism*. Although most representationalists accept the existence of memory traces (Sutton 2010), and take them to be identical with said intermediate memory representations, the relationship between a memory representation and a memory trace need not be ~~identified.~~ Identity. As we shall see, memory traces can be conceived as part of, or as causally relevant to, the instantiation of the memory representation carrying the intentional content during recollection.

The *Theatetus* is often cited as the first philosophical text where memory traces are mentioned. Plato compares experiences leaving traces in our memory to seal rings leaving impressions on a wax table. The first *argument* for the existence of memory traces, however, is likely from Aristotle. At the beginning of *De Memoria et Reminiscentia*, Aristotle follows a content-based approach - the idea, roughly, that cognitive faculties are individuated in terms of the contents they operate with (De Brigard 2014a, 2014b) - to distinguish memory from expectation and perception (*DM*, 449b25). Unlike expectation, which is about things that have not happened yet, and unlike perception, which is about things that are happening now, memory deals with things that have already happened. That is, the intentional objects of our memories are things in the past. If so, Aristotle wonders, how is it possible that something that is not present - the remember***ed*** - be recalled by something that is present - the rememb***ering*** (450a25)? Recall that, for Plato, what we presently perceive is an image (*phantasma*) of the past. This answer, however, raises the following concern: if what happens when we remember is that we perceive an image of the past thing, should we then say that what we remember is the image or should we rather say that what we remember is the thing from which the image was produced (450b11)? According to Aristotle, both options are problematic. If we say that what we remember is the current image, then we are forced to say that what we remember is present. But this is clearly false: memory is about the past,

and the past is not present. However, if we say that what we remember is in the past, then how could it be possible that we can perceive that which is not present? Perception is, by definition, of the present. The past is not present. So, Aristotle goes beyond the impression metaphor and offers a more elaborate solution, using of the platonic notion of *eikon* or "copy". "An eikon of X", Sorabji (2006) reminds us, "is both similar to, and derived from, X." Moreover, for Aristotle, this derivation is causal (450a27-b11); the preserved eikon brought to mind when one remembers is both similar to and causally derived from the perceived object. In sum, according to Aristotle, remembering consists of bringing to mind a preserved representation that is both a copy of, and caused by, a previous perception of a past object or event (Annas 1986).[1]

Aristotle's view - i.e., that memory representations are stored and causally derived copies of previous perceptions - is essentially identical to what appears to have been the received philosophical view on memory traces up until the early 1980s. Initially resurrected by Martin and Deutscher (1966), and later articulated - albeit disapprovingly - by Norman Malcolm (1977), this view holds that for a mental representation to count as a memory trace, it needs to meet three conditions: (1) it must play a causal role in the recollection of the event it is a trace of; (2) it must retain the intentional content entertained during the remembered event; and (3) it must be structurally similar or isomorphic to the event that is remembered.[2] In addition to these *causal*, *retention*, and *similarity* conditions, there is another, more subtle parallel between Aristotle's and the more recent view: both take memory traces to be theoretical entities whose presumed existence is the result of an inference to the best explanation (IBE), rather than non-theoretical entities whose existence is determined observationally (Malcolm 1977; Heil 1978; Bernecker 2008; although see Rosen 1975). Aristotle's IBE posits memory traces to explain how we can entertain an intentional content in the present whose intentional object is in the past. He does not come to the conclusion that they exist on the basis of having found them. Similarly, the more recent view posits the existence of memory traces, not on the basis of an empirical observation, but rather as an IBE that avoids acceptance of causation-at-a-temporal distance between the experienced object/event and its subsequent recollection (Martin and Deustcher 1963; Sutton 1998; Bernecker 2008, 2010; whether or not this IBE is tantamount to Aristotle's is an exercise left for the reader).

Consequently, philosophers of memory have come to use the term "memory trace", in its most general form to refer to some sort of mental entity, or mental property instantiated in a particular entity, for which the causal, retention, and similarity conditions obtain, and which exists during a period of time, t_2, between a time, t_1, in which a subject, *S*, experiences a particular event *x*, *Ex*, and a subsequent time, t_3, in which *S* remembers or recollects *x*, *Rx*. As we shall see, although many views accept all of the stipulated constraints - while differing on exactly what each amounts to - some others reject them, arguing that the correct view of memory traces should better accommodate certain facts rather than try to fit some *a priori* agenda.

2. Philosophical Views on Memory Traces: Charting the Logical Space

Representationalists who adhere to the strictest interpretation of the aforementioned definition take memory traces to be mental representations whose contents closely resemble their objects by way of having been causally derived from them during perception. In addition, representationalists hold that such contents are preserved intact until subsequent recollection. Notice, however, that there are at least three possible ways of understanding the relationship between the content of the stored memory trace and the perceptual content it is derived from. *Direct epresentationalism* takes a memory trace to be a mental representation created only after the remembered object has been perceived, for perception is understood as being un-mediated,

i.e., as involving no intermediary perceptual representations. *Semi-direct representationalism* understands a memory trace to be the very same representation created during the initial perception and subsequently stored for later retrieval. Finally, *indirect representationalism*, which views both perception and recollection as representational, takes a memory trace to be numerically different from the perceptual representation entertained during the initial perception of the object, even though both representations carry the same intentional content.

Aristotle's view on memory traces appears to have been a version of direct representationalism. He considered memory traces to be representations of the past things themselves. However, given the prominence of representationalist views of perception, direct representationalism remains underrepresented in the philosophical literature, even though it constitutes an attractive position for theorists who may find direct realism about perception appealing but refuse to commit to a direct realist view about memory, such as Reid's (1785/1849) or Laird's (1920). Semi-direct representationalism has been more popular. Augustine, for instance, thought memory was the "storehouse of images of things experienced" (Augustine, X:12-13). Locke revived this metaphor in the first edition of his Essay (Locke 1690/1975). He thought that memory traces were stored past perceptions that remained unperturbed in our memory, hidden from consciousness, until they were further retrieved during recollection. But Locke quickly recanted this view because it contradicted a main tenet of his philosophy: that there are no unconscious ideas in the mind. If memory stores ideas one is not aware of, then the mind can have unconscious ideas, which - for Locke - was absurd. As a result, Locke revised his view of memory in an oft-quoted passage of the second edition of his Essay:

> But, our ideas being nothing but actual perceptions in the mind, which cease to be anything when there is no perception of them; this laying up of our ideas in the repository of the memory signifies no more but this: that the mind has a power in many cases to revive perceptions which it has once had, with this additional perception annexed to them, that it has had them before. And in this sense it is that our ideas are said to be in our memories, when indeed they are actually nowhere; but only there is an ability in the mind when it will to revive them again, and as it were paint them anew on itself, though some with more, some with less difficulty; some more lively, and others more obscurely. (Essay X:2)

Critics quickly pointed out more difficulties. Reid, for instance, argued that if a mental representation ceases to exist during a certain period of time after which it is revived, then it has two beginnings: when it first appears in consciousness and when it is later recalled. But, as Locke admits, the same thing cannot have two different origins. Therefore, all Locke can say is that memory can create a new mental representation that, at most, resembles the previous one. But resemblance is not enough for a representation to be a memory representation. Suppose - to paraphrase Reid's own example - that you look at an object at time t1 and form a representation. You look away for a moment and then back at it at t2. You then form a second representation of the object that clearly resembles the first one. However, the second representation is not a memory of the object, for you are not remembering it: you are simply seeing it again (Reid 1785/1849, III: 7). How can you tell apart genuine cases of remembering from cases of re-seeing? According to Reid, you can do it if you can recognize that the second representation is, or is very much like, the previous one. But this act of recognition presupposes memory. Therefore - according to Reid - Locke faces the problem of either accepting that one thing can have two origins, or admitting that his definition of memory is circular.

A second problem, crucial for our current purposes, has to do with the cause of the revived idea, i.e. the memory representation. If memories are revived representations formerly

contemplated during perception, and if that which is perceived is caused by the objects affecting our senses, what are the causes of our memories of those objects? Surely the cause cannot be the objects themselves; that would force us to accept some kind of causation-at-a-temporal distance, such as Russell (1921) or Malcolm's (1977) mnemic causation. Alternatively - as some direct realists about memory would posit (e.g., Harvey 1940) - one could claim that past objects do not cease to exist but rather are "prolonged in time" in a way that is no longer accessible to perception, just to memory. This view is not only counterintuitive and metaphysically improbable, but also fails to explain why our grasp of the object is more inaccurate and less vivid during recollection than during perception (Furlong 1948).

The solution is to accept that the intentional content of our memory representation is the same as its corresponding original perception *because* there is a representational vehicle that is not only capable of preserving the initial content but also of bringing it about during recollection. This solution - known as the causal theory of memory (Bernecker 2008, 2010) - reintroduces the necessity of causally linking the initial experience with its subsequent recollection via a memory trace. There are several ways in which a memory trace can guarantee such causal continuity. The most straightforward way is simply to allow that the representational vehicle carrying the intentional content of the encoded experience remains invariant until subsequent recollection. This *content invariantist* version of semi-direct representationalism - close to Locke's view in the first edition of his Essay - has been endorsed by a number of philosophers (e.g., Martin and Deutscher 1966; Zemach 1968), as it easily meets the causal, retention, and similarity conditions stated above. A content invariantist version of semi-direct representationalism can be agnostic as to what the precise nature of the stored representation is, as long as it preserves the structure of the original experience.[3]

A different approach is to deny that what gets stored during encoding is the representational vehicle carrying the intentional content of the initial experience. Rather, encoding could be understood as the storage of a fragment and/or a compressed version of the initial representation which, upon recollection, has the capacity to reconstruct the intentional content entertained during the original experience. This *reconstructivist* (as opposed to *reproductive*) view of indirect representationalism comes in two flavors: either one can be a content *invariantist* and suggest that the intentional content reconstructed during retrieval is identical to the intentional content entertained during encoding, or one can be a content *variantist* and suggest that both intentional contents differ in important respects. Intriguingly, although content invariantism was depicted as the received view at the turn of the 19th century (e.g., Stout 1898/1915), actually most philosophers at the time accepted content variantism, claiming that the contents of our memories were fainter and less vivid than the perceptual contents they were derived from (Hume 1739/1975; Mill 1869). Moreover, many also thought that memories were mental compounds that included, in addition to a (fainter) encoded perceptual content, an added mental state. The nature of this additional state varied. Some philosophers held that, in order to count as a memory, a memory trace must be retrieved along with an appropriate *cognitive* state - such as the belief that it happened in the past (Aristotle, *DM*; Russell 1921) or that it happened in one's own past (James 1890). Some philosophers thought that what must accompany a memory trace is a particular kind of *affective* state, such as a feeling or an emotion - that of duration or "pastness" (Spinoza 1985; Locke 1690/1975; Hume 1739/1975). Finally, some philosophers held that in order to count as a memory, a memory trace needs to be retrieved along with a second-order representation or an *apperception* - e.g., that the currently entertained mental content is the same content one entertained before - so that only organisms capable of self-referential cognitions can have episodic memories (Leibniz 1714/1991; Kant 1787/1998). The main point about content variantist versions of reconstructivist indirect representationalism is that they allow for violations

of the similarity constraint; neither the intentional content stored nor the one retrieved need to be structurally isomorphic with the intentional content entertained during encoding. In addition, content variantism differs from previously discussed versions of semi-direct and indirect representationalism in that it does not take the retrieved memory to be identical with the memory trace, but rather it considers memory traces as constituents – or causal contributors – of the mental representation entertained during recollection.

This leads to the nature of the representational vehicle carrying the intentional content of the memory trace, another dimension across which philosophical views on memory traces tend to vary. As mentioned, the majority of contemporary philosophers take memory traces to be neural entities. They are thereby *restricted* representationalists, as they assume that the intentional contents of retrievable experiences get encoded exclusively in the brain – or, at least, in the central nervous system. There are a number of views as to what sort of neural implementation memory traces can adopt, and some of them will be reviewed in the next section. Some proponents of the extended mind hypothesis, however, have suggested that memory traces are not confined to the brain but rather *extended* to the body and/or external environment and objects. For instance, using a clever thought-experiment, Clark and Chalmers (1998) argued that, if for a certain individual, A, an external component – e.g., a notepad – is functionally equivalent to a neural component of the memory trace of another individual, B, then A's memory trace would include a component outside the brain. A strong reading of Clark and Chalmers' proposal would imply that if A and B's memory traces are functionally equivalent, and memory traces are identical to the memory representations entertained during retrieval, then the intentional contents of A and B's memories would be the same. But this strong reading of extended representationalism is probably false, as the intentional content of A's and B's representations during retrieval can easily differ even if A and B exhibit equally successful recollective behavior (Adams and Aizawa 2001). A weaker interpretation of extended representationalism, according to which memory traces are *not* identical to, but rather causally responsible for and/or constitutive of, the mental representation carrying the memorial content during retrieval, might avoid this objection, as it allows for other contributive factors to bring about the full recollective experience. Further research is needed to clarify whether or not there are critical differences between the ways in which restricted versus extended memory traces contribute to remembering (Sutton 2009; Michaelian and Sutton 2013).

Two other points of departure among different philosophical views on memory traces merit mentioning. First, a tacit assumption among many representationalists is that the causal, retention, and similarity conditions are categorical properties instantiated in a certain entity – presumably neural – from the moment of encoding to the moment of retrieval. Recently, however, some philosophers have suggested that at least some of them can be understood as *dispositional* properties with the propensity to be instantiated when the right conditions obtain. Robins (2012), for instance, suggests that what the memory trace retains is not an actualized intentional content but rather the dispositional property to manifest said content given certain conditions at retrieval (also Sutton 1998; De Brigard 2014a, 2014b). Similarly, De Brigard (2011a, 2011b, 2014) suggests that memory traces do not preserve the structure of the representational vehicle from the moment of encoding to the moment of retrieval, but rather the dispositional property to recreate (a very close approximation of) the representational vehicle carrying the encoded intentional content at the time of retrieval. (I will get back to this issue at the end).

Second, there is an important disagreement among philosophers as to the nature of memory traces based upon the way memories are temporarily experienced during recollection. The received view, at least in the so-called analytic tradition, can be dubbed *discrete*

representationalism. It holds that our memory representations are discrete or discontinuous with each other, and due to some cognitive mechanism (e.g., a habit of the mind, for Hume, or the synthesis of intuition plus the unity of apperception, for Kant), we are aware of them – i.e. they are present to our consciousness – as though they were continuous. There is also an important tradition of non-discrete or *continuous* representationalists, among which one can find philosophers such as James (1890), Bergson (1908), Husserl (1938/1991), and Merlau-Ponty (1962). Despite their differences, these theorists accuse discrete representationalists of failing to take the phenomenology of remembering as primitive and to artificially slice the contents of our mental life. The nature of this disagreement is so fundamental that it goes beyond the nature of memory representation, and thus beyond the scope of this paper. Nonetheless, it is an important avenue for future research in neurophilosophy, as many neuroscientists are starting to move away from static notions of memory representations toward more dynamic approaches in terms of memory processes.

3. The Psychology and Neuroscience of Memory Traces

By the end of the 19th century, when philosophy and psychology parted ways, both sides agreed on the hypothetical status of memory traces. For psychologists, the existence of memory traces was – to use James' term (James 1890: 655) – a psychophysical hypothesis: they were posited as existing in the brain, and as persisting for particular period of time. Take, for instance, Maudsley's (1876) *The Physiology of the Mind*, where he adopts the view that to remember is to reinstate an idea previously held in perception plus the added idea that one had entertained it before. He asks: "What is the modification of the anatomical substrata of fibers and cells, or of the physiological activity, which is the occasion of this *plus* element in the reproduced idea?" (p. 513). Maudsley's suggestion was that, during retrieval, the same "nerve-circuit" affected by the initial perception is "disposed to fall again readily into the same action", and that the modifications said nerve-circuit underwent from the time of the initial perception to the time of it is reactivation during recollection would constitute the neural substratum of the added consciousness. However speculative an answer, Maudsley's was nonetheless consistent with the current understanding of the physiology of perception. For this is the mark of the neurophysiologists of memory: their conjectures are constrained by current knowledge in neurophysiology *and* by one's philosophical stance on the psychological nature of memorial representations and their contents. Disagreements occur when theorists identify a mismatch between these constraints. James, for example, criticized Maudsley's view on account that sameness of nerve-circuitry and an added conscious idea were incompatible claims, so memories must have a different neural substratum than the perceptions they were derived from.

Given the precarious state of the experimental study of memory traces during the first part of the 20th century, these discussions went from the realm of the speculative to that of oblivion. Memory theorists like Richard Semon (1904/1921), who coined the term engram – synonymous with memory trace – and wrote extensively about its nature, only became influential during the second half of the century, and was hardly read by his contemporaries (Schacter 2001). With the advent of methodological behaviorism, the notion of memory trace fell in disrepute (Watson 1930; Skinner 1953). Any mention of "memory" as opposed to "learning" was jettisoned from psychological writings, and those who kept searching for the engram reached rather pessimistic conclusions. In 1950, Karl Lashley – who was trained as a behaviorist by J. B. Watson – published his famous paper *In Search of the Engram*, in which he declared that "it is not possible to demonstrate the isolated localization of a memory trace anywhere within the nervous system" (Lashley 1950). Many interpreted Lashley's view as proof against the prospect of ever finding memory traces.

During the second half of the 20th century, two critical discoveries resurrected the hopes for finding memory traces. The first occurred in 1957, when Brenda Milner described the case of H.M., an individual who had sustained a bilateral resection of the medial temporal lobes, four years earlier, as a result of an intractable epilepsy (Scoville and Milner 1957). The surgery left H.M. unable to store new episodic information, and it impaired his recollection of recent life events, while apparently sparing all other intellectual abilities. This observation opposed Lashley's, for it showed that there was a clear dissociation between brain areas that were required for the creation of new memories and the retention of recent ones, and brain areas that were not. Without a hippocampus you simply cannot memorize new conscious experiences.[4]

The second discovery was equally influential. Researchers in Andersen's neurophysiology lab in Oslo observed that hippocampal cells retained their synaptic potentiation long after the electric stimulus was removed (Craver 2003). The first description of the underling mechanism of this phenomenon, known now as long-term potentiation (LTP), was offered by Bliss and Lomo in 1973. The discovery of a neural mechanism that could preserve the effects of a stimulus once removed, and the fact that such mechanism happened to be found in a region that was demonstratedly necessary for the formation of new memories (i.e., the hippocampus), gave a new life to the research on memory traces.

These discoveries illustrate two complementary approaches to study the neural underpinnings of memory traces. On the one hand, researchers using neuropsychological methods and cognitive neuroscientific techniques (e.g. fMRI, ERP), attempt to unveil the neural substrates of memory traces at the *systems* level, i.e., at the level of gross anatomical structures such as the medial temporal lobe or the sensory cortices. On the other hand, research on the neurophysiology and neurobiology of memory and learning try to uncover the neural substrates of memory traces at the *cellular* level. Results from both the system and the cellular levels have improved our understanding of the nature of memory traces tremendously. However, bridging both levels of description remains a major challenge (Morris 2007).

Both approaches focus on two intertwined questions: (1) *when* do memory traces get formed; and (2) *where* do they get stored. A typical answer is that memory traces get stored as patterns of neural connections once the encoded information is consolidated. Unfortunately, this simple view is fraught with conceptual and empirical difficulties. To begin with, although most researchers would agree that memory comprises three separate processes - *encoding*, *storage*, and *retrieval* - there is little agreement as to when encoding begins and ends, and when a memory is consolidated and stored (Davachi 2007). Consider what it takes to remember an every-day complex multisensory scene such as, say, a car collision. As in most real life situations, we would probably allocate attention to only a subset of the elements of the scene, and for most of them, we would do so merely transiently. Psychological evidence shows that retention of episodic information depends on attentional allocation during encoding, so that information is better retained when it is deeply or attentively encoded than when it is encoded shallowly or under divided-attention conditions (Craik and Tulving 1975; De Brigard 2011b). As a result, not everything that is perceived is encoded, for the conscious perception of an event is a necessary but not a sufficient condition for it to be encoded. However, this does *not* mean that we should conceive of perception and encoding as two successive stages, as a number of attention-dependent neural processes that occur during conscious perception persist during encoding as well - e.g., neuronal depolarization (Jensen and Lisman 2005), sustained spiking (Fransen et al. 2002; Hasselmo 2007) - even after attention has been shifted away from the content being encoded. Perhaps what we see here is a difference of levels of description: whereas perception appears to be the kind of process better described at a personal level, encoding may be best defined sub-personally, in terms

of neural mechanisms performing certain operations. Still, even if the boundaries between the processes of perception and encoding appear fuzzy, the evidence overwhelmingly suggests that not everything that we consciously perceive gets to be encoded.

A related difficulty pertains to the moment in which encoding ends and consolidation occurs. According to the prevalent view during most of the second half of the 20th century, while the encoding of a perceived scene requires the interaction of the hippocampus and the neo-cortex - specifically, modality-specific areas - once memory traces are consolidated, the hippocampus is no longer needed, as the retrieval of the memory trace is supported by the prefrontal cortex and the sensory cortices alone (Squire 1984; McClelland et al. 1995). The advantage of this view is that it accounted for three fundamental pieces of evidence. First, individuals with medial temporal lobe damage (such as H.M.) tend to exhibit a temporally graded retrograde amnesia (i.e., impairment to recover memories for events prior to the lesion), with recent memories being more affected than remote ones (a consistent phenomenon known as Ribot's law). Second, damage in the hippocampus produces severe anterograde amnesia (i.e., incapacity to store new memories for events occurring after the lesion) with preserved short-term memory, suggesting that the hippocampus is required for the successful encoding and consolidation of new declarative memories in long-term memory. Finally, evidence from these case studies also suggests that storage of new semantic and episodic information is equally affected, which indicates commonalities in the brain mechanisms involved in both kinds of declarative memory processes.

Advocates of this "standard model" of consolidation at the systems level also consider that their view dovetails with our current understanding of consolidation at the cellular level. The prevalent view on cellular or synaptic consolidation holds that experiences get encoded as changes of connectivity among the neurons originally involved in processing the perceptual information. From this perspective, learning consists in the activation and reactivation of neural networks whose co-activation strengthens their connection weights until they become highly selective for their proximal stimulus. First articulated by Hebb (1949), this view has found support in molecular and genetic neurobiology (e.g., Kandel 1976; Silva et al. 1998) as well as computational neuroscience (McClelland and Goddard 1996). The precise mechanisms of these cell assemblies are complex, and involve a number of processes such as enzimatic production (Silva et al. 2002), gene regulation (Kida et al. 2002), and the formation of novel dentritic spines (Engert and Bonhoeffer 1999). Nonetheless, the moral of the story is relatively clear: the experience of an event - such as a car collision - involves the activation of a number of regions in the sensory cortex: the auditory cortex processes the sound of the cars crashing, the visual cortex processes the colors and shapes you see through the windshield, the lateral temporal cortices helps to categorize the perceived objects on the street, and so on (Frankland and Bontempi 2005). Next, it is thought that hippocampal activity, presumably modulated by the fronto-parietal attentional network, helps to bind together these cortical areas into a larger hippocampal-neo-cortical network (McClelland et al. 1995). With time, the binding processing of the hippocampus is no longer required, and memory traces become consolidated in the neo-cortex, as neuronal assemblies ready to be reactivated by the pre-frontal cortex given the appropriate cue.

This standard model has been recently challenged by Nadel and Moscovitch (1997). They note, for instance, that a careful examination of individuals with medial temporal lobe damage shows that their retrograde amnesia for detailed autobiographical events extends for decades, sometimes even for their lifetimes, whereas the retrograde amnesia for semantic memory is less extensive, temporally graded, and differentially compromised depending on whether it involves public events, world facts or vocabulary. In fact, Nadel and collaborators (2000) have shown that the degree of retrograde episodic autobiographical amnesia is directly proportional to the amount of hippocampal damage, and that different regions of the medial

temporal lobe differentially contribute to the formation of episodic, spatial, and semantic memories (Moscovitch et al. 2005). Their analyses suggest that the hippocampus is required during retrieval as well as encoding (Eldridge et al. 2000; Ryan et al. 2008), and that the capacity to remember remote autobiographical memories depends on the activation of preserved hippocampal tissue (Mullally et al. 2012).

Nadel and Moscovitch suggest amending the standard model with what they call "multiple trace theory (MTT)". According to MTT, memory traces are initially formed as sparse neural networks whereby a hippocampal index binds disconnected neo-cortical cell assemblies encoding sensory information. Additionally, MTT holds that every time a memory trace gets reactivated, a new neo-cortical network is re-indexed by the hippocampus. New traces, of course, largely overlap with the initial ones, which both help to strengthen the contents of the memory but also renders them liable to distortions and modifications. Finally, the formation and re-formation of memory traces is supposed to help to explain the process by means of which invariant features across experiences become more and more stable through time, and in turn less and less reliant on the hippocampus - a process that is thought to explain why semantic memory traces are relatively more impervious to hippocampal damage relative to episodic memory traces. Although still incomplete and perhaps imprecise (Knowlton and Fanselow 1998), MTT pinpoints difficulties with the standard model while also offering a comprehensive explanation of the way in which the brain appears to record past experiences.

Two further lines of evidence have also put pressure on the standard model. First, a large amount of behavioral evidence demonstrates that the contents of our memories are prone to different kinds of distortions. For example, people are likely to misremember a plausible event as having occurred if they previously imagined it (imagination inflation; Garry et al. 1996), to wrongly recognized as experienced information that was misleadingly introduced at the time of retrieval (post-event misinformation; Loftus and Hoffman 1989), and to false alarm to lures that are conceptually or semantically related to studied items (Roediger and McDermott 1995). Many interpret these results as suggesting that retrieval renders memory traces liable to distortion, as information processed online while one is remembering can infiltrate and/or modify the re-encoded memorial content. Further support for this claim comes from studies in animal neurobiology, where it has been shown that consolidated memory traces which are normally resistant to amnesic manipulations, become highly unstable upon reactivation, which renders them susceptible to the kind of amnesic interventions that are effective during initial consolidation. Thus, these results strongly suggest that, upon reactivation, memory traces undergo a re-consolidation process in which the retrieved memorial contents are updated (Nader and Einarsson 2010; Hardt et al. 2010).

The reviewed empirical evidence has important implications for philosophical theories on memory traces. First, contrary to direct and semi-direct versions of representationalism, it indicates that the retrieved memory trace differs from the encoded one in both number and content, supporting thus indirect and variantist versions of representationalism. Additionally, the evidence supports reconstructivist views, as retrieved contents do not appear to be unchanged replays of perceived contents, nor do they seem to involve the reproduction of the representational vehicle of the initial perception. Instead, these results suggest that remembering involves the online reconstruction of memorial representations from incomplete and disaggregated episodic information stored in patterns of neural activation across disperse cortical areas (Bartlett 1932; Schacter and Addis, 2007). Moreover, recent computational models suggest that accumulated prior experiences may constrain the reconstructive processes during memory retrieval in a hierarchical probabilistic way (Hemmer and Steyvers 2009; Steyvers et al. 2006; De Brigard 2012), which could help to explain why memory distortions tend to be plausible and often beneficial (Schacter et al. 2011).

Finally, this scientific evidence also suggests that it is not accurate to talk in terms of a perception leaving a memory trace in the brain. Encoding does not seem to lead to any kind of brain modification that remains solely dedicated to coding for the stimulus that caused it. On the contrary, the neural networks engaged in the initial encoding get constantly redeployed in a variety of tasks and for a large number of purposes, which likely cause them to undergo further transformations. As such, it may be best to think of a memory trace, not as a neural network constantly carrying a particular memorial content from encoding to retrieval, but rather as instantiating the dispositional property to reinstate, as closely as possible, the complex hippocampal-neo-cortical pattern of neural activation the brain was in during encoding, at the time of retrieval (De Brigard 2011a, 2014a, 2014b). The precise way in which such dispositional properties are manifested, and the retrieval conditions under which their actualization is possible, are among the many questions left for further research. For it is likely that future results in the science of memory will keep challenging our traditional philosophical theories on memory traces, forcing us to revise long-held commonsensical views about the experience of remembering in drastic and perhaps counterintuitive ways.[5]

Short Biography

Felipe De Brigard is an Assistant Professor of Philosophy at Duke University and core faculty at the Center for Cognitive Neuroscience and the Duke Institute for Brain Sciences, where he directs the Imagination and Modal Cognition Lab. He earned a bachelor's degree from the National University of Colombia, a master's degree from Tufts University, and a doctoral degree from the University of North Carolina, Chapel Hill. He then spent two years as post-doctoral fellow in the Department of Psychology at Harvard University. His research centers on the interaction between memory and imagination - in particular, hypothetical and counterfactual thinking - as well as the relationship between attention, consciousness and recollection.

Notes

* Correspondence: Department of Philosophy, Center for Cognitive Neuroscience, Duke Institute for Brain Sciences, 203A West Duke Building, Box 90743, Duke University, Durham, NC 27708-0743. Email: felipe.debrigard@duke.edu.

[1] Aristotle had other reasons to identify a memory trace with an *aikon* rather than a *phantasma*, as Plato did. Phantasmata are simply impressions that resemble their objects. But not all resemblances are due to having been caused by their objects, and Aristotle thought a causal connection to their objects was necessary for memory traces. Aristotle argued, for instance, that in cases of relearning, one re-acquires a phantasma that resembles a past experienced event, even though relearning is not remembering. An eikon, by way of being causally related to the past experienced event, circumvents this objection.

[2] The vehicle/content ambiguity in the literature on memory traces is rampant. Köhler (1938), for instance, talks about memory *experiences* being isomorphic with the events they represent, but in the same text, he talks about their biological correlates to be isomorphic as well. The ambiguity is also pervasive among physiologists, like Penfield (1955) and Gerard (1953), who talk about memory traces being structurally isomorphic to the event they represent, but whether this structure refers to the content or the vehicle of the representation is unclear. Indeed, Malcolm (1977) discusses the ambiguity at length, and - as we shall see - different interpretations of the "similarity" condition have yielded distinct views on memory traces.

[3] A content invariantist can make use of the explanatory machinery of the language of thought (Fodor 1975) to support the view that experiences initially encoded as particular propositions are stored as formulas in mentalese until later retrieved during recollection. Or she might prefer a connectionist view on memory traces, arguing instead that the intentional content of the original experience is stored as connection weights in distributed patterns of activation across neural networks (Sutton 1998). Either way, the main point for content invariantist versions of semi-direct realism is that memory traces are conceived as representational vehicles whose intentional contents are preserved unchanged from encoding to retrieval.

[4] I'm following the traditional interpretation here. H.M. did in fact have cognitive deficits outside the memory realm, suggesting that the hippocampus is critically involved in a number of cognitive processes in addition to episodic memory retrieval (Corkin 2002; MacKay and James 2009; De Brigard 2014a, 2014b).

[5] Thanks to Edouard Machery, Sarah Robins, and an anonymous reviewer for their helpful comments. Also, thanks to Scott Guerin, Nathan Spreng, Peggy St Jacques and Karl Szpunar for many memory-related conversations over lunch at The Kebab, around the time I was starting to write this paper.

Works Cited

Adams, F. and K. Aizawa. 'The Bounds of Cognition.' *Philosophical Psychology* 14 (2001): 43–64.

Annas, J. 'Aristotle on Memory and the Self.' In J. Moravcsik, and M. Woods (Eds.), *The Festschrift for John Ackrill*. Oxford Studies in Ancient Philosophy 4 (1986): pp. 99–118.

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

Bergson, H. *Matter and Memory*. New York: Zone Books, 1908.

Bernecker, S. *The Metaphysics of Memory*. Dordrecht: Springer, 2008.

——. *Memory*. Oxford: Oxford University Press, 2010.

Bliss T. V. and T. Lomo. 'Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of the Perforant Path.' *Journal of Physiology* 232(2) (1973): 331–56.

Clark, A. and D. Chalmers. *The Extended Mind*. Analysis, 58 (1998): 7–19.

Corkin, S. 'What's New with Patient H.M.?' *Nature Reviews Neuroscience* 3 (2002): 153–60.

Craik, F. I. M. and E. Tulving. 'Depth of Processing and the Retention of Words in Episodic Memory.' *Journal of Experimental Psychology: General* 104 (1975): 268–94.

Craver, C. 'The Making of a Memory Mechanism.' *Journal of the History of Biology* 36 (2003): 153–95.

Davachi, L. 'Encoding as a Concept.' *Science of Memory: Concepts*. Eds. R. Roediger, Y. Dudai, and S. Fitzpatrick. New York: Oxford University Press, 2007.

De Brigard, F. Reconstructing Memory. PhD Dissertation. University of North Carolina, Chapel Hill, 2011a.

——. 'The role of attention in conscious recollection.' *Frontiers in Psychology* 3 (2011b): 29.

——. 'Predictive memory and the surprising Gap. Commentary on Andy Clark's "Whatever Next? Predictive Brains, Situated Agents and the Future of Cognitive Science".' *Frontiers in Psychology* 3 (2012): 420.

——. 'Is Memory for Remembering? Recollection as a Form Episodic Hypothetical Thinking.' *Synthese* 191 (2014a): 155–85.

——. The Anatomy of ~~Memory~~ Amnesia. *Scientific American Mind* (2014b) May/June: 33–37.

Eldridge, L. L., et al. 'Remembering Episodes: A Selective Role for the Hippocampus During Retrieval.' *Nature Neuroscience* 3 (2000): 1149–52.

Engert, F. and T. Bonhoeffer. 'Dendritic Spine Changes Associated with Hippocampal Long-Term Synaptic Plasticity.' *Nature* 399 (1999): 66–70.

Fodor, J. *The Language of Thought*. Cambridge, MA: Harvard University Press, 1975.

Frankland, P. W. and B. Bontempi. 'The Organization of Recent and Remote Memories.' *Nature Reviews Neuroscience* 6 (2005): 119–30.

Fransen, E., A. A. Alonso, and M. E. Hasselmo. 'Simulations of the Role of the Muscarinic-Activated Calcium-Sensitive non-Specific Cation Current I(NCM) in Entorhinal Neuronal Activity During Delayed Matching Tasks.' *Journal of Neuroscience* 22(3) (2002): 1081–97.

Furlong, E. J. 'Memory.' *Mind* 57 (1948): 16–44.

Garry, M., et al. 'Imagination Inflation: Imagining a Childhood Event Inflates Confidence That it Occurred.' *Psychonomic Bulletin & Review* 3(2) (1996): 208–14.

Gerard, R. W. What is memory? *Scientific American* 189 (1953): 118–26.

Hardt, O., E. Ö. Einarsson, and K. Nader. 'A Bridge Over Troubled Water: Reconsolidation as a Link Between Cognitive and Neurotraditions.' *Annual Review of Psychology* 61 (2010): 141–67.

Harvey, J. W. 'Knowledge of the Past.' *Proceedings of the Aristotelian Society* 41 (1940): 149–66.

Hasselmo, M. E. 'Encoding: Models Linking Neural Mechanisms to Behavior.' *Science of Memory: Concepts*. Eds. R. Roediger, Y. Dudai, and S. Fitzpatrick. New York: Oxford University Press, 2007.

Hebb, D. O. *The Organization of Behavior*. NY: Wiley, 1949.

Heil, J. 'Traces of Things Past.' *Philosophy of Science* 45 (1978): 60–72.

Hemmer, P. and M. Steyvers. 'A Bayesian Account of Reconstructive Memory.' *Topics in Cognitive Science* 1 (2009): 189–202.

Hume, D. *A Treatise of Human Nature*. Oxford: Oxford University Press, 1739/1975.

Husserl, E. *On the Phenomenology of the Consciousness of Internal Time*. Netherlands: Kluwer, 1938/1991.
James, W. *The Principles of Psychology*. New York: Henry Holt & Co., 1890.
Jensen, O. and J. E. Lisman. 'Hippocampal Sequence-Encoding Driven by a Cortical Multi-Item Working Memory Buffer.' *Trends in Neurosciences* 28(2) (2005): 67–72.
Kandel, E. R. *Cellular Basis of Behavior, an Introduction to Behavioral Neurobiology*. San Francisco :W. H. Freeman and Company, 1976.
Kant, I. *Critique of Pure Reason*. Cambridge: Cambridge University Press, 1787/1998.
Kida, S., et al. 'CREB Required for the Stability of new and Reactivated Fear Memories.' *Nature Neuroscience* 5 (2002): 348–55.
Knowlton, B. J. and M. S. Fanselow. 'The Hippocampus, Consolidation, and on-Line Memory.' *Current Opinion in Neurobiology* 8 (1998): 293–6.
Köhler, W. *The place of value in a world of facts*. New York: Liveright, 1938.
Laird, J. *A Study in Realism*. Cambridge: Cambridge University Press, 1920.
Lashley, K. In Search of the Engram. *Society of Experimental Biology* Symposium 4, 1950. 454–82.
Leibniz, G.W. *Monadology*. Indiana: Hacket, 1714/1991.
Locke, J. *An Essay Concerning Human Understanding*. NY: Oxford University Press, 1690/1975.
Loftus, E. F. and H. G. Hoffman. 'Misinformation and Memory: The Creating of new Memories.' *Journal of Experimental Psychology: General* 188(1) (1989): 100–4.
MacKay, D. G. and L. E. James. 'Visual Cognition in Amnesic H.M.: Selective Deficits on the What's-Wrong-Here and Hidden-Figure Tasks.' *Journal of Experimental and Clinical Neuropsychology* 31 (2009): 769–89.
Malcolm, N. *Memory and Mind*. Ithaca, NY: Cornell University Press, 1977.
Martin, C. B. and M. Deutscher. 'Remembering.' *Philosophical Review* 75 (1966): 161–96.
Maudsley, H. *The Physiology of Mind*. London: MacMillan and C., 1876.
McClelland, J. L. and N. Goddard. 'Considerations Arising from a Complementary Learning Systems Perspective on Hippocampus and Neocortex.' *Hippocampus* 6 (1996): 654–65.
McClelland, J. L., B. L. McNaughton, and R. C. O'Reilly. 'Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory.' *Psychological Review* 102 (1995): 419–57.
Merlau-Ponty, M. *Phenomenology of Perception*. New York: Humanities Press, 1962.
Michaelian, K. and J. Sutton. 'Distributed Cognition and Memory Research: History and Current Directions.' *Review of Philosophy and Psychology* 4(1) (2013): 1–24.
Mill, J. *Analysis of the Phenomena of the Human Mind*. London: Baldwin & Cradock, 1869.
Morris, R. G. M. 'Memory: Distinctions and Dilemmas.' *Science of Memory: Concepts*. Eds. R. Roediger, Y. Dudai, and S. Fitzpatrick. New York: Oxford University Press, 2007.
Moscovitch, M., et al. 'Functional Neuroanatomy of Remote Episodic, Semantic and Spatial Memory: A Unified Account Based on Multiple Trace Theory.' *Journal of Anatomy* 207 (2005): 35–66.
Mullally, S., D. Hassabis, and E. Maguire. 'Scene Construction in Amnesia: An fMRI Study.' *Journal of Neuroscience* 32 (16) (2012): 5646–53.
Nadel, L. and M. Moscovitch. 'Memory Consolidation, Retrograde Amnesia and the Hippocampal Complex.' *Current Opinion in Neurobiology* 7 (1997): 217–27.
Nadel, L., et al. 'Multiple trace theory of human memory: Computational, neuroimaging and neuropsychological results.' *Hippocampus* 10 (2000): 352–68.
Nader, K. and E. Ö. Einarsson. 'Memory reconsolidation: An Update.' *Annals of the New York Academy of Sciences* 1191 (2010): 27–41.
Penfield, W. The permanent record of the stream of consciousness. *Acta Psychologica* 11 (1955): 47–69.
Reid, T. *Essays on the Intellectual Powers of Man*. Edinburgh: McLachlan, Stewart, & Co., 1785/1849.
Robins, S. A capacity view of memory traces. PhD Dissertation, 2012.
Roediger, H. L. and K. B. McDermott. 'Creating false memories: remembering words not presented in lists.' *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (1995): 803–14.
Rosen, D. 'An Argument for the Logical Notion of a Memory Trace.' *Philosophy of Science* 42 (1975): 1–10.
Russell, B. *The Analysis of the Mind*. London, 1921.
Ryan, L., et al. 'Hippocampal Activation during Episodic and Semantic Memory Retrieval: Category Production and Category Cued Recall.' *Neuropsychologia* 46 (2008): 2109–21.
Schacter, D. L. *Forgotten Ideas, Neglected Pioneers: Richard Semon and the Story of Memory*. Philadelphia: Psychology Press, 2001.
Schacter, D. L. and D. R. Addis. 'The cognitive neuroscience of constructive memory: Remembering the past and imagining the future.' *Philosophical Transactions of the Royal Society B* 362 (2007): 773–86.
Schacter, D. L., S. A. Guerin, and P. L. S. Jacques. 'Memory distortion: an adaptive perspective.' *Trends in Cognitive Sciences* 15(10) (2011): 467–74.

Scoville, W. B. and B. Milner. 'Loss of recent memory after bilateral hippocampal lesions.' *Journal of Neurology, Neurosurgery and Psychiatry* 20 (1957): 11–21.
Semon. *The Mneme*. London: Allen & Unwin, 1904/1921.
Silva, A. J., et al. 'CREB and memory.' *Annual Review of Neuroscience* 21 (1998): 127–48.
——. 'Deficient hippocampal long-term potentiation in alpha-calcium-calmodulin kinase II mutant mice.' *Science* 257 (2002): 201–6.
Skinner, B. F. *Science and Human Behavior*. New York: Macmillan, 1953.
Sorabji, R. *Aristotle on Memory*. Chicago: University of Chicago Press, 2006.
Spinoza, B. *The Collected Works of Spinoza*. Vol. I. Ed. E. M. Curley. Princeton: Princeton UP, 1985.
Squire, L. R. 'Neuropsychology of memory.' *The Biology of Learning*. Eds. P. Marler and H. Terrace. Berlin: Springer-Verlag, 1984. 667–85.
Steyvers, M., T. L. Griffiths, and S. Dennis. 'Probabilistic inference in human semantic memory.' *Trends in Cognitive Sciences* 10(7) (2006): 327–34.
Stout, G. F. *A Manual of Psychology*. London: University Tutorial Press, 1898/1915.
Sutton, J. *Philosophy and Memory Traces: Descartes to connectionism*. Cambridge: Cambridge University Press, 1998.
——. Memory. *Stanford Encyclopedia of Philosophy*, 2009.
——. 'Observer Perspective and Acentred Memory: some puzzles about point of view in personal memory.' *Philosophical Studies* 148 (2010): 27–37.
Watson, J. B. *Behaviorism*. Chicago: University of Chicago Press, 1930.
Zemach, E. M. 'A Definition of Memory.' *Mind* 77(308) (1968): 526–36.