Forthcoming in The Harvard Review of Philosophy – Final Draft – 3/24/20

The explanatory indispensability of memory traces

Felipe De Brigard
Duke University

Doubt about traces is to be dispelled by reference to the sheer difficulty of psychological explanation without them.
(John Sutton, 1998: 300)

Abstract. During the first half of the 20th century, many philosophers of memory opposed the postulation of memory traces based on the claim that a satisfactory account of remembering need not include references to causal processes involved in recollection. However, in 1966, an influential paper by Martin and Deutscher showed that causal claims are indeed necessary for a proper account of remembering. This, however, did not settle the issue, as in 1977 Malcolm argued that even if one were to buy Martin and Deutscher's argument for causal claims, we still don't need to postulate the existence of memory traces. This paper reconstructs the dialectic between realists and anti-realists about memory traces, suggesting that ultimately realists' arguments amount to inferences to the best explanation. I then argue that Malcolm's anti-realist strategy consists in the suggestion that causal explanations that do not invoke memory traces are at least as good as those that do. But then, *contra* Malcolm, I argue that there are a large number of memory phenomena for which explanations that do not postulate the existence of memory traces are definitively worse than explanations that do postulate them. Next, I offer a causal model based on an interventionist framework to illustrate when memory traces can help to explain memory phenomena and proceed to substantiate the model with details coming from extant findings in the neuroscience of memory.

1. Introduction

The notion of memory trace is as old as our interest in understanding memory. It predates the distinction between philosophy and psychology; indeed, it predates the distinction between philosophy and *science*. In the *Theaetetus*, for instance, Plato talks of experiences leaving traces in our memory as seal rings leave impressions in a wax table (194c). These impressions—the analogy tells us—are representatives of the seal ring, just as memory traces are representatives of the experiences that created them. Zeno the Stoic and Aristotle also embraced the view that perception leaves traces, and that such traces give rise to the memories we later on recover during recollection.[1] The appeal to memory traces persists in many modern philosophers' writings, including Descartes, Hobbes, Locke, Hume, and Mill.[2] Aside from some notorious skeptics,[3] the use of memory traces to explain the phenomenon of recollection was so widespread during the 19th century that it became the received view by the time psychology established itself as an independent discipline around the 1870s.

By the end of the 19th century both philosophers and psychologists seemed to agree that, given the status of neuroscience at the time, memory traces were merely hypothetical. However, philosophers and psychologists disagreed as to *how* to interpret the scope of this hypothesis. On the one hand, philosophers saw the postulation of memory traces as a *theory-independent hypothesis*. Memory traces were hypothetical precisely because their acceptance within a theory of memory was at stake. As a result, they thought that the first step in order to know whether or not there are memory traces was conceptual: one needed to find out whether the notion of memory trace was at all required for our correct understanding of memory. On the other hand, psychologists

thought of memory traces as a *theory-dependent hypothesis*; indeed, they thought of it as a psychophysical hypothesis.[4] From the psychologist's point of view, memory traces were hypothetical, not because we were questioning whether they were required at all for a correct account of memory, but rather because we just didn't know what sort of physical (i.e., neural) entity they could be. Psychologists thought that the task of verifying the nature of memory traces was an empirical one; it had to do with finding out the nature and precise location of memory traces in the brain, not with whether or not we were justified in postulating them. Thus, while philosophers like Russell and Broad were interested in finding out whether or not we required the notion of memory trace in order to have a full-fledged account of remembering, psychologists like Semon—who coined the term "engram" to refer to memory traces—and James were in the business of devising theories about the biological and physiological nature of memory traces.[5]

As a result, during the first part of the 20th century, the existence of memory traces became the object of two different and independent inquiries: one scientific or empirical, and one philosophical or conceptual. Indeed, for many philosophers engaged in the latter, developments in the psychology and neuroscience of memory were often seen as irrelevant to answering the philosophical question about the existence of memory traces.[6] After all—they reasoned—if we are not justified in postulating their existence, then we have less reason to believe that scientists are warranted in taking memory traces to be the sort of entity that can be empirically discoverable. Against this background, a number of philosophers—including Russell, Ryle, Wittgenstein, and Benjamin [7]—strongly argued against the view that a successful account of remembering requires any reference whatsoever to a causal intermediary between the past experience and its subsequent recollection. This skepticism fueled an anti-realist view about memory traces that found its clearest expression in Malcolm's opposition to the realist stance offered by Martin and Deutscher.[8]

The purpose of this essay is to critically examine Malcolm's anti-realist strategy vis-à-vis Martin and Deutscher's proposal, and to offer an argument for a realist view about memory traces. More precisely, I argue that Malcom's argument against the need for invoking memory traces in our explanations of remembering is flawed, and that in fact we often need to postulate them. To that end, I will begin, in section 2, by going over Malcolm's influential characterization of memory traces. As I will explain, this characterization identifies three necessary conditions for something to be a memory trace. One of these conditions—the causal condition—constitutes the target of Malcolm's argument, as it employed by realists about memory traces to postulate their existence. Next, in section 3, I show that the use of this causal condition to postulate the existence of memory traces is a particular instance of an argumentative strategy in the philosophy of science whereby theorists postulate the existence of unobservable entities supposedly referred to by our theoretical terms. Specifically, I claim that arguments in favor of the reality of memory traces are usually inferences to the best explanation (IBE), motivated by the realist's rejection of causal explanations involving action at a temporal distance. Malcolm's anti-realist move is to argue that, for the case of remembering, causal explanations involving spatiotemporal gaps between cause and effect are at least as good explanations as those involving memory traces, thus undercutting the realist's motivation to postulate their existence.

In section 4, I argue against this anti-realist strategy, not by way of showing that action at a distance is not possible, but rather by suggesting that the mere acceptance of action at a distance still does not give us the best possible causal explanation for recollection. More specifically, I argue that even if one accepts the possibility of causal explanations involving action at a distance, there are still many causally related questions about recollection for which that sort of explanation is insufficient. I offer instead, in section 5, a model for the causal explanation of recollection using

an interventionist framework.[9] I explain how such an account would fare better than the mere action at a distance account—and, incidentally, than the mere causal account—when it comes to many causally relevant questions about recollection. As anticipated, this proposed model requires the existence of causally relevant memory traces. To substantiate the model, I offer—in section 6—a mechanistic interpretation of the postulated memory traces drawing from recent developments in the neuroscience of learning and memory. I explain what, according to these empirical findings, memory traces may be, and how they can behave causally as suggested by the interventionist model. Finally, in section 7, I briefly discuss three possible objections to my argument.

2. *The philosophical notion of memory trace*

Although there are some disagreements in the way philosophers of memory use the notion of memory trace,[10] for the current purposes I will endorse the characterization favored by Malcolm.[11] According to this view, in its most general form, "memory trace" is used in reference to an entity or a process (or a set of entities or processes), *M*, that exist 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*. Additionally, according to Malcolm, in order for *M* to be the memory trace of *x*, M_x, three conditions must obtain. First, there is *the causal condition*, which states that a memory trace must play a causal role in the recollection of the event it is a trace of; *Mx* needs to have been caused by *Ex*, and in turn it needs to be the cause of *Rx*. Second, there is *the retention condition*, which says that a memory trace must retain the intentional content entertained during the remembered event; that is, to be the memory trace of *Ex*, the intentional content acquired by *Mx* at t_1 need to be kept unchanged through t_2 until retrieved at t_3. Finally, there is *the isomorphism condition*, which claims that a memory trace must be structurally similar or isomorphic to that which is remembered; that is, the intentional content entertained during *Rx* must be isomorphic to that of *Ex* because its structural isomorphism is preserved through t_2 by *Mx*.

Many philosophers have argued against the last two conditions on conceptual and empirical grounds. Some have suggested, for instance, that it is not at all clear what it means to say that memory traces preserve or retain mental contents through time,[12] or that they need to do so in a format that, in any interesting way, is structurally isomorphic to the remembered experience or event (e.g., Rosen 1975). In fact, some have argued that memory traces cannot meet the isomorphism condition because experiences don't have structures, and thus there is no structure to preserve or reinstate at retrieval.[13] Likewise, on the basis of scientific findings, other philosophers have argued against the claim that if memory traces are neural entities, then there most likely cannot be a meaningful way of talking about memory traces *preserving* any sort of mental content—as opposed to, say, *reconstructing* it during retrieval.[14]

Although I will briefly mention the retention and the isomorphism conditions toward the end of the paper, my focus here will be on the causal condition, as it constitutes the backbone of the dialectic between realists and anti-realists about memory traces. More precisely, as I show in the next section, while both realists and anti-realists about memory traces seem to agree on a causal condition for remembering, they disagree as to whether or not it entails the need to postulate causal intermediaries, and this disagreement divides their ontological commitments regarding memory traces.

3. Memory traces as theoretical posits

As mentioned above, skepticism about memory traces gained some initial prominence in contemporary philosophy of memory with the publication of Russell's *The Analysis of the Mind* in 1921. There, Russell argued that given the precarious state of the science of memory at the time, if we were to take causal claims about past experiences causing present recollections to be meaningful, then we need to find a way to interpret such claims without the appeal to unobservable (and, thus, unverifiable) causal intermediaries. His proposal, very much in the spirit of positivism, was to introduce the notion of "mnemic causation," viz. the view according to which a past experience can be the direct cause of a subsequent recollection. In other words, according to Russell's theory of mnemic causation, there is nothing wrong with saying that *Ex* at t_1 is the direct proximal cause of *Rx* at t_3. Notice, incidentally, that this strategy is not that dissimilar from that employed by methodological behaviorists, prominently Watson and Skinner, for whom explanations about learning needed to include only stimuli and responses outside of the organism, with the former almost always diachronically separated from the latter.[15] Far from being a far-fetched proposal, I reckon, Russell's theory of mnemic causation was very much *a la par* with contemporaneous psychological theories of learning.[16]

Russell's criticism of the notion of memory trace inherent in his theory of mnemic causation was echoed, and amplified, by Wittgenstein, who was strongly opposed to the idea of invoking any kind of intermediary trace to explain memory. In his *Remarks on the Philosophy of Psychology*—written in the mid-1940s—Wittgenstein famously remarked that "whatever [an experienced] event does leave behind in the organism, *it* isn't the memory".[17] Moreover, he seemed to suggest that our tendency to think that memories are caused by traces speaks to the need of modifying our idea of causation itself:

> I saw this man years ago: now I have seen him again, I recognize him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored-up there in any form? Why must a trace have been left behind? Why should there not be a psychological regularity to which no physiological regularity corresponds? If this upsets our concepts of causality then it is high time they were upset.[18]

Wittgenstein's skepticism was then carried on by Ryle and Benjamin, but, I daresay, it found its most clear expression in his disciple Norman Malcolm. In fact, Malcolm not only denies the need to postulate memory traces to give an account of remembering, he even denies that we need to talk about experiences causing memories at all. For instance, in his influential work *Knowledge and Certainty*, he emphatically declares that "our use of the language of memory" carries no implication about the causes of our remembering or about the causal mechanisms involved in our recollections.[19] As a result, Malcolm's view—at least in 1963—was that we don't need to postulate the existence of memory traces as causal intermediaries between an experience, *Ex*, and its recollection, *Rx*, because we can dispose of the causal condition altogether; neither our concept of memory nor of remembering requires any reference whatsoever to causes or causal processes and, as a result, we can easily do away without them.

Shortly after, however, Martin and Deutscher's celebrated paper *Remembering*, published in 1966, put realism about memory traces back on the table. The main purpose of that paper was to argue against the claim that a causal condition is not required for a proper analysis of our concept of remembering. Their argument is based on cases—some real, some imaginary—of people that had a particular experience *x* at a certain time, t_1. Then, during some subsequent and arbitrary

interval of time, t_2, they forget the event. However, at a later time, t_3, these people do something "for which the only reasonable explanation" (Martin and Deutscher 1966: 176) is that they experienced *x* at t_1. The observation that a causal claim is required in order to make sense of their behavior at t_3 as a consequence of the experience at t_1 motivates Martin and Deutscher to claim that "if a person's account of what he saw is not due even in part to his seeing it, it cannot be said that he remembers what he saw" (Martin and Deutscher 1966: 175-176). Since the examples Martin and Deutscher discuss intuitively fall under our concept of remembering, then they conclude that a person's recollection of *x* must be due to her having experienced *x*. Thus, they formulate their causal condition (CC) for remembering:

(CC) *S*'s experience of a particular event *x*, *Ex*, causes—or, at least, is causally relevant to—*S*'s subsequent recollection of the event, *Rx*.

There are two points I would like to extract from Martin and Deutscher's analysis. First, their paper brought causation back into the analysis of memory by way of pointing out the explanatory indispensability of the causal condition. Unfortunately, this point isn't stressed enough, partly because Martin and Deutscher's own analysis suggests that CC is simply the conclusion of the following *modus tollens*:

Argument 1:

(P1) If *S*'s *Rx* is not caused by *Ex*, then S is not *Rx*-ing (Assumption)
(P2) But S is *Rx*-ing (as evidenced by their thought experiments)
(C) Therefore, *S*'s *Rx* is caused by *Ex*.

But notice that this interpretation may render their argument vacuous. As stated, P1 implies its contrapositive:

*(P1) If S is *Rx*-ing, then S is caused by *Ex*

which is precisely the conditional they want to prove. But, of course, Martin and Deustcher should not want that, for if P1 is supposed to be an *assumption*, then the argument shouldn't prove what they assumed to begin with.

I think a better interpretation is to treat their argument *inductively*, so that the thought experiments they discuss are to be accommodated by an IBE. Consider one of their well-known thought experiments. A painter is asked to draw an imagined rural landscape. When he's done, his parents recognize the painting as depicting the view from the house they used to live in many years ago. The painter, who has no recollections of the time they lived in such house, claims to have imagined the scene he painted. However, the intuition this thought experiment is supposed to elicit—and let's assume that it does for the sake of argument—is that the painter is actually painting the scene from memory and not from imagination. Notice that Martin and Deuschter's claim is that we take the painter's envisioning the scene while painting it as a case of remembering (as opposed to imagination) *because* the "only reasonable explanation" for his mentally entertaining that precise scene at t_3 is his having experienced it at t_1, even if he did not remember it at all during t_2. Thus, if we take their argument to be an IBE, then CC would enter as the

hypothesis that best fits the data provided by the thought experiment. Here is the argument, schematically:[20]

Argument 2:

(P1) The painter's case is clearly an instance of recollection.
(P2) The hypothesis CC, if true, would explain why the painter's case is an instance of recollection.
(P3) No other hypotheses can explain why the painter's case is an instance of recollection as well as CC does.
(C) Therefore, CC is (probably) true.

If this is the case, then the appeal to CC is the result of an IBE, as opposed to the conclusion of a deductively valid argument.

The second point I'd like to extract from Martin and Deutscher's analysis is that the existence of memory traces is supposed to follow, as a matter of course, from the acceptance of CC (Martin and Deutscher 1966: 189). Their argument, which is not terribly straightforward in their paper, can be reconstructed as follows:[21]

Argument 3:

(P1) *S*'s *Ex* is diachronically separated from *S*'s *Rx*.
(P2) A cause cannot be diachronically separated from its effect (i.e., there is no causation at a temporal distance).
(MTC) Therefore, there *must* be an intermediary causal connection, *Mx*, between *Ex* and *Rx* such that *Ex* is the proximal cause of *Mx* and *Mx* is the proximal cause of *Rx*.

But notice, once again, that the argument for MTC hinges on an IBE. The idea is that the postulation of memory traces as causal intermediaries between *Ex* and *Rx* allows us to preserve CC without having to accept the metaphysically uncomfortable claim that there is causation at a temporal distance. In other words: memory traces become theoretical posits postulated to help explain the causal connection between an experience, *Ex*, and its recollection, *Rx*, without having to accept action at a temporal distance. Thus, Malcolm tells us:

> Presumably the reader will know that memory traces are not entities, states, or processes that neural surgeons have discovered in the course of their investigations of the brain, as dentists discover cavities. The memory trace is what may be called "a theoretical construct". It is something that is inferred to exist from the presence of things that unquestionably exist, such as learned skills, habits, and occurrences of recognition and remembering.[22]

It is worth remarking that, at least for Malcolm, memory traces are postulated in virtue of our "abhorrence of 'action at a distance'—in this case, action at a temporal distance".[23] Memory traces are conceived as playing the role of "bringing about a memory response: [for] without the existence of a trace there would be a gap in a causal chain and causal action would occur at *a temporal distance*".[24] In brief, the motivation behind the postulation of a memory trace as a theoretical posit

is the fact that it constitutes a better explanation of how *Ex* causes *Rx* than the alternative action-at-a-temporal distance account in which *Ex* directly causes *Rx*.

The problem with this conclusion, however, is that it leaves open the following possibility: when it comes to explaining how *Rx* came about as a result of *Ex,* an explanation involving action at a temporal distance could be *at least as good* as an explanation involving memory traces. This is precisely Malcolm's important move in his 1977 book *Memory and Mind*. His argument, which is reminiscent of Russell's defense of mnemic causation, is that the kind of explanation we usually invoke when talking about remembering does not imply, in any way, that there should be a process mediating *Ex* and *Rx*. Suppose—to use one of his examples—that you tell someone that you saw a boat capsize last week. Now imagine that, for whatever reason, your interlocutor is in disbelief: 'How do you know that?' she asks, to which you reply, 'I know because I saw it happen'.[25] The thought here is that in explaining how it is that you remember the boat capsizing, you are applying a causal claim, just as Martin and Deutscher argued, but it makes no reference to any sort of causal process or state mediating the event perceived and your recollection of it. As a matter of fact, it makes no sense to ask whether you are certain that there was an ongoing causal process between your witnessing the boat's capsizing and your relating the story.

> The strange, irrelevant character of this question shows that there is a familiar use of causal language consisting of such ordinary locutions as "because of", "due to", "the cause of", and the more technical "necessary causal condition", which carries no implication of a causal process filling up the temporal space between the occurrence of a cause of x and the occurrence of x. We can agree with Martin and Deutscher that the language of memory does, in a sense, require a "causal interpretation", but not agree that memory as a causal concept entails the concept of causal process […] Eliminate the assumption of a causal *process*, and *the causal argument* for a memory trace collapses.[26]

In sum, Malcolm argues that one can accept CC without having to postulate memory traces. Causal explanations involving action at a temporal distance are—according to him—perfectly reasonable explanations for recollection, and nothing about intermediate causal processes is implied by our use of the concept of remembering. While CC may be an IBE as to how *Ex* causes *Rx*, one does not need to accept the second IBE in which memory traces are postulated; explanations involving action at a temporal distance are as good as those involving memory traces. In the next section, however, I argue that they are not.

4. The explanatory indispensability of memory traces

In the previous section I claimed that the appeal to memory traces stemmed from the realization that their postulation was required to come up with the best possible explanation as to how an experience, *Ex*, causes its recollection, *Rx*. After all, the assumption of causally mediating memory traces avoided the uncomfortable metaphysical pitfalls of causation at a temporal distance. Malcolm's anti-realist reply, however, was that one could accept the claim that *Ex* causes *Rx* without having to be committed to causally mediating memory traces, simply because explanations involving action at a temporal distance are equally good explanations for remembering. As a result, the postulation of memory traces as theoretical entities for an adequate account of remembering was thought to be unnecessary, and the idea of trying to find them empirically was deemed unwarranted. With this move, Malcolm made anti-realism about memory traces, once again, an attractive theory in the philosophy of memory.[27]

Notice, though, that Malcolm is not arguing in favor of the possibility of action at a temporal distance as a *metaphysical* claim. Whether or not a cause can bring about an effect after a temporal gap is irrelevant to Malcolm's argument. His point, just like Martin and Deutscher's, is about causal *explanation*. After all, he accepts the IBE motivated by 'Argument 2.' What he rejects is the IBE motivated by 'Argument 3.' And he rejects it, not because he denies P2 as a metaphysical claim, but rather because he denies that the acceptance or rejection of causation at a temporal distance has anything to do with successful causal explanations for the phenomenon of recollection.[28] In other words, he does not think that the postulation of intermediary causal processes adds anything to our account of how *Rx* was brought about as a result of *Ex*.

This, however, is what I think Malcolm gets wrong, for I am not sure how explanations involving action at a temporal distance can really satisfy our explanatory necessities when it comes to various causally relevant questions about memory and remembering. If we *only* care about experiences causing successful recollections, as Martin, Deutscher, and Malcolm do, maybe a case can be made to the effect that action at a temporal distance is all we need to accept in order to furnish satisfactory causal psychological explanations. But successful recollection is *not* the only thing we care about when we demand causal explanations for our memories. We often want to know, for instance, why is it that a person, having experienced an event, can nonetheless *fail* to remember it. Additionally, we may want to know why, given that a subject experienced a particular event, she only managed to remember *part* of it, or why she remembered it *distortedly*. Moreover, sometimes we wonder whether it is possible to *facilitate* or to *hamper* our subsequent recollection of an event after having witnessed it. To put it succinctly, we often wonder whether it is possible to intervene in the alleged causal connection between an experience, *Ex*, and is subsequent recollection, *Rx*.

Let me offer an analogy to drive my point home. Consider a case in which someone consumes cyanide at t_1 and then dies at t_3. A natural way of describing the event is to say that the person died as a result of her consuming cyanide—that the ingestion of cyanide caused her death. If all we want to know is why she died at t_3, alluding to her consuming cyanide at t_1 may be a sufficient explanation. The same goes for remembering. As I stressed, Malcolm's examples (as well as Martin and Deutscher's) only pertain to successful recollections of past events. After all, the motivation behind the IBE that leads to the acceptance of CC is simply that we cannot make sense of a particular *successful* memory retrieval behavior at t_3 unless we accept as its cause having the relevant experience at t_1. A similar IBE is at work when, upon seeing a dead body exhibiting the distinctive signs of cyanide poisoning, a coroner alludes to the person's prior ingestion of cyanide as a causal explanation of his death. In such a case, asking the coroner whether or not he's certain that a causal process was going on between the person's ingestion of the cyanide and his eventual death would seem as awkward as asking whether or not you are certain that a causal process was going on between your witnessing an event and your recalling it afterwards. Here, alluding to your having witnessed a boat capsizing—to borrow Malcolm's example—may be enough of an explanation as to why you remember it, the same way in which alluding to cyanide ingestion may be enough of an explanation for the person's death.

But suppose that, right next to the dead body, there is another person who also ingested cyanide but *failed* to die. Let's assume that she exhibited some of the symptoms—shortness of breath and pink skin color—but none of the lethal ones, like pulmonary edema and cardiac arrest. Again, in this case, cyanide ingestion can explain the person's symptoms. For example, if someone asks why her skin is pink, one can rightly say that it is due to her having ingested cyanide. But then an obvious question arises: given that both subjects ingested cyanide, why is it that only one

of them died while the other *failed* to die? Now, I take it, talk of intermediary causal processes becomes necessary. The only way in which one can explain why, given the same initial conditions, one person died while the other person failed to die, is by way of alluding to some difference in the causal process that occurred between the cyanide ingestion and the subsequent symptomatic behavior. One possibility is that the person who survived had increased levels of hydroxocobalamin in her blood due to, say, excessive consumption of vitamin B_{12}. As a result, the ingested cyanide preferentially bonded molecules of hydroxocobalamin, leaving the hemoglobin's cytochrome oxidase less affected—which would explain why her levels of blood oxygenation were enough to elicit shortness of breath and skin coloring but *not* lethal pulmonary or cardiac arrest. There are other possibilities too. The point, though, is that when it comes to explaining the differential effects of cyanide ingestion in these two people, any successful causal explanation is going to involve intermediary causal processes.

The same is true in the case of memory. Consider a small modification of Malcolm's example. Suppose that you weren't alone when you witnessed the boat capsizing. You were with your friend Mary. Both you and Mary were side by side when the event occurred, both of you were looking at the event, and both of you have roughly the same visual acuity. However, only you remember the event later on. Now, when you wonder why is it that you remember the event while Mary fails to remember the same event, even when both of you witnessed it, appealing to an intermediary causal process is the natural way to proceed. One may say, for instance, that Mary wasn't paying attention, or that she has difficulties consolidating information from short- to long-term memory, or perhaps that she has seen so many boats capsizing lately that she cannot remember just that one. Of course, one may allude to some more "organic" explanations; one may say, for instance, that Mary was given an amnestic drug right after she witnessed the event, or that she suffers from some kind of neurodegenerative disease, or even that her medial temporal lobes were damaged at some point after having witnessed the boat capsizing. Whatever the story we tell, it is going to involve a reference to intermediate causal processes that differed between her case and yours.

Notice that the point I am making does not hinge on our knowledge of the neural mechanisms by means of which memories get consolidated and further retrieved—that part of the story will come later. My point so far is about the necessity of alluding to intermediate causal processes in order to reach the best causal explanation of an *unsuccessful*—versus a *successful*—case of remembering. In other words, a causal explanation that does not make reference to intermediate causal processes won't be able to account for the differential effects between cases of successful and unsuccessful recollection. This means that, when it comes to accounting for differential effects during recollection, a causal explanation that does not involve intermediate causal processes won't be as good a causal explanation as one involving intermediate causal processes. Thus, Malcolm is wrong when claiming that, when it comes to recollection, explanations involving action at a temporal distance are explanatorily on par with those that posit intermediate processes.

To be sure, this argument can also be made when the differential effect involves *improved*—as opposed to impaired—recollection. Suppose that Mary did not fail to remember the witnessed event but she actually remembered it better than you did. Unlike you, she remembered—let's say—that there was a red fender on the starboard side when the boat capsized. Again, other things being equal, any successful explanation is going to involve some reference to intermediate causal processes that differed between you and Mary. These processes can be as simple as closely attending to the fender while witnessing the event, or as complex as having received a dose of

strychnine—shown to enhance memory retention in some mammals[29]—right after seeing the boat capsizing. I think the same goes for other differential effects, not only between subjects but also within subjects. For example, someone may adduce lack of sleep when trying to explain why she failed at a particular test that later on, after a good night of sleep, she can pass with no trouble. The fact of the matter is that we often allude to intermediate causal processes when we offer explanations of differential effects in recollection.

Malcolm is wrong, then, in thinking that causal explanations involving action at a temporal distance are explanatorily equivalent to those involving intermediate causal processing. He isn't entirely to blame, though. The root of the problem, I think, lies in interpreting CC as stating that a reference to *Ex* may be a sufficient condition for explaining how *Rx* came about. Martin and Deutscher also share this assumption, for they appeal to memory traces via the second IBE stated in Argument 2—the second premise of which Malcolm rejects. But this is the wrong way to introduce memory traces. What Martin and Deutscher should have said is that CC *and* memory traces are a package deal. More precisely, what they should have said is that appealing to the past event *Ex* alone does not constitute the best causal explanation for *Rx* (save, perhaps, in the very circumscribed and highly under-described cases of successful recollection that Malcolm discusses). The past event is *part* of the causal explanation, but on most occasions, as in the cases of differential effects just discussed, the appeal to memory traces is also required, not as a fallout of accepting the past event as the cause of *Rx*, but as a resource to explain the psychological phenomenon itself. I suggest, therefore, to modify CC in favor of a causal condition that incorporates memory traces:

(CC*) *S*'s experience of a particular event *x*, *Ex*, plus a memory trace of *x*, *Mx*, cause—or, at least, are causally relevant to—*S*'s subsequent recollection of the event, *Rx*.

At this point, it is then worth asking how and when do memory traces become causally relevant when it comes to explanations of differential effects on recollection. I suggest an answer to these questions in the next section.

5. Intervening memory

In the previous section I argued, *contra* Malcolm, that memory traces become indispensable to explain differential effects in recollection, and that causal explanations that do not appeal to intermediary causal mechanisms are not as good as those that do. I now want to suggest that a promising strategy to understand how and when memory traces become indispensable to account for differential effects on recollection; this strategy involves relying on Woodward's manipulability theory of causation.[30] According to his view, causes are considered devices for manipulating and controlling effects. Causal explanations explain because they convey information about the way in which one could potentially manipulate or control a certain effect by intervening on a previous event we take to be its cause. Thus, successful causal explanations are used to answer what Woodward calls "what-if-things-had-been-different" questions: "the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways".[31] In the case of cyanide poisoning, for instance, understanding why one person died while the other one survived requires understanding what it is that we could have done in the case of the person that died to affect the result that occurred in the case of the person who did not die. In other words, we want to know whether there was something one could have done between t_1 and t_3 to prevent her death.

Might it have been possible that even though she ingested cyanide at t_1 we could have done something at t_2 in order to prevent her death at t_3?

As it turns out, there are ways in which one can prevent death by cyanide poisoning. For instance, we know that cyanide, when dissolved in water, inhibits cytochrome oxidase blocking electron transport, which in turn decreases the amount of oxygen in the blood. This condition causes lactic acidosis, whereby the pH of the hemoglobin is reduced and it starts building up D-lactate, which rapidly damages organic tissue—especially in our lungs and stomach—thus leading to one's death. As a result, a person's death after ingesting cyanide is potentially preventable at several points during the process. Most typically, one could administer nitrites to turn hemoglobin into methahemoglobin, which is preferentially bonded by the cyanide. The bonding of cyanide and methahemoglobin creates cyanmethahemoglobin, which in turn can be treated with sodium thiosulfate to convert the cyanmethahemoglobin into hemoglobin, sulfites, and thiocyanate, the last of which can be secreted through urine without further damage to the organism. However, other possible manipulations could be potentially implemented, like the use of hydroxocobalamins to artificially increase the pH level in the hemoglobin while eliminating the cyanide, or by finding a mechanism to inhibit the creation of D-lactate.

The relevant point is that these interventions—some of which are in fact implemented in medical facilities (like the use of nitrites) and some of which are merely potential (like the use of some chemical agent that could reduce the hemoglobin's pH)—allow us to manipulate the buildup of D-lactate. When the levels of D-lactate in the blood reach a certain threshold, a body enters into the physiological condition known as lactic acidosis, which can be lethal. But if one can reduce the levels of D-lactate, lactic acidosis is then prevented and the chances of survival increase. Therefore, according to the manipulation account I am relying on, the immediate cause of death in the cyanide poisoning case just described is the amount of D-lactate in the person's blood. We can tell that because we know that had we intervened to reduce the level of D-lactate in the blood, the person would have merely experienced shortness of breath and skin discoloration.

Let me put it graphically. As I described the case above (see Figure 1A), there are two different events, *B* (death via cardiac arrest) and *C* (skin discoloration), which appeared to have been caused by the same event *A* (cyanide consumption). However, as I argued, if we limit our causal explanation to *A*, the differential effect would remain mysterious. So, we wonder whether some other event *D* happened between the time t_1 in which *A* occurred and the time t_3 in which both *B* and *C* occurred, such that it could explain the differential effect. As it turns out, there is: lactic acidosis. We know that *D* is the cause of *B* because we can intervene, *I*, on *D* and prevent the buildup of D-lactate, thus manipulating the effect and 'switching', as it were, the causal path from *B* to *C* (Figure 1B).

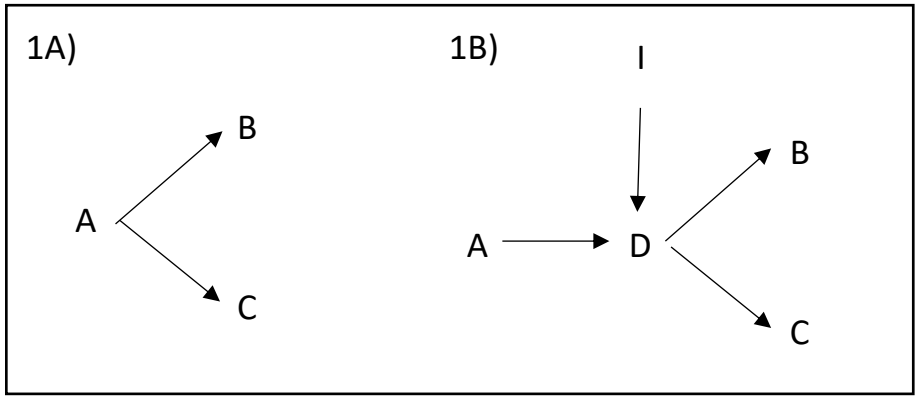

Figure 1: 1A) Graphical models of an acyclic causal graph with one cause and two possible effects. 1B) Graphical model of a manipulation on an intervening factor.

The same, I surmise, occurs with memory. The suggested variation on Malcolm's case unveils a differential effect between a situation in which one remembers the event, *B*, and situation in which one does not remember the event, *C*, despite the fact that in both situations one has experienced the initial event, *A*. As in the case of the cyanide poisoning, appealing merely to having witnessed the event does not explain the differential effect. So we wonder whether there is an intermediate event, *D*, such that a proper intervention, *I*, upon it could switch the causal chain from *B* to *C*. In other words, we want to know whether there is an intermediate causal mechanism that could explain why one person remembered or failed to remember a particular event. And we could know that with the appropriate intervention. Enter neuroscience.

6. Memory traces as multi-level neural mechanisms

In the last section, I suggested an interpretation of the role of memory traces in causal explanations of remembering in terms of Woodward's manipulability theory of causation. Now I would like to suggest a mechanistic model for memory traces that can provide a framework for understanding the way in which certain experimental manipulations conducted by cognitive neuroscientists have actually produced—and could possibly produce—differential effects in recollection. Since this interpretation is largely inspired by Craver's account of multi-level neural mechanisms,[32] it is useful to explain what he means by such terms.

According to Craver, a mechanism is a set of entities and activities arranged in specific ways to produce regular changes in a period of time.[33] By 'neural mechanism,' therefore, I will refer to the sorts of mechanisms studied in neuroscience. Neural mechanisms typically include entities such as neurons, neurotransmitters, oligodendrocytes, hippocampi, brains, etc. They also include activities such as neuronal firing, enzyme release, information processing, brain region activation, etc. The entities and the activities composing neural mechanisms have spatial and temporal organizations that are essential for the mechanism to perform its operations. Finally, the ways in which the mechanism's entities and activities are organized typically compose hierarchies. Each strata of the mechanistic hierarchy is usually called a 'level,' so mechanisms that can be decomposed into more than one level are 'multi-level mechanisms'.

As an illustration of a multi-level neural mechanism, consider Craver's example of a mechanistic decomposition of spatial memory in four levels (Figure 2). Each level is the object of study of a relatively independent sub-discipline in the neurosciences, as each level is investigated with a distinctive array of experimental methods. The top level includes entities such as organisms (e.g., mice, humans) and surrounding environments, as well as activities such as discrimination, button pressing and swimming. At this level, experimental psychologists, cognitive ethologists, and comparative psychologists study spatial memory using experimental methods like the Morris water maze, radial arm mazes and virtual reality computers. In the second level (one level down) we find entities such as the hippocampus and the entorhinal cortex, as well as computational activities such as informational transfer and spatial map formation. This level is usually studied by cognitive neuroscientists and neuropsychologists via experimental methods such as event related potentials (ERP), functional magnetic resonance imaging (fMRI), positron emission tomography (PET) scans, and several diagnosis assessment methodologies often implemented in clinical settings. The relevance of the hippocampus and the entorhinal cortex—that is, the entities of the

second level—is determined by their dependence on the entities and the activities of the third level. This level includes entities such as granule and pyramidal cells, and activities such as neuronal firing and depolarization. Neurophysiologists and, to some extent, neuroanatomists, study this mechanistic level with experimental methods such as intra- and extra-cellular recording, cell body staining, track tracing, and, sometimes, very localized neuropharmacological interventions, like optogenetics and microiontophoresis, whereby the researcher injects small dosages of particular chemical compounds directly into the neural tissue. Finally, the bottom level consists of molecular mechanisms that include entities such as N-Methyl D-aspartic (NMDA) receptors and Mg^{2+} ions, and activities such as binding and electron releases. Molecular neurobiologists study this level using experimental methods such as pharmacological interventions and gene knockouts.[34]

Although an oversimplification, Craver's spatial memory example highlights an essential feature of any multi-level neural mechanism—including, as I suggest, memory traces. Craver calls it *mutual manipulability,* and it basically specifies a condition of sufficiency for a component to be a part of a multi-level mechanism. According to the mutual manipulability condition, "a part is a component of a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the component *and* one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole."[35] For example, we can tell that LTP in the pyramidal cells in region CA1 of a rat's hippocampus is part of the multi-level neural mechanism of the organism's spatial memory because we can intervene upon its mechanistic operations—by removing NMDA receptors in this location, for instance—thus inhibiting the activity of the place cells and making it impossible for the hippocampus to form spatial maps. Conversely, we can alter the induction of LTP in CA1 by way of intervening higher neural levels, e.g., severing afferent neural tracts or modifying the rat's behavior.

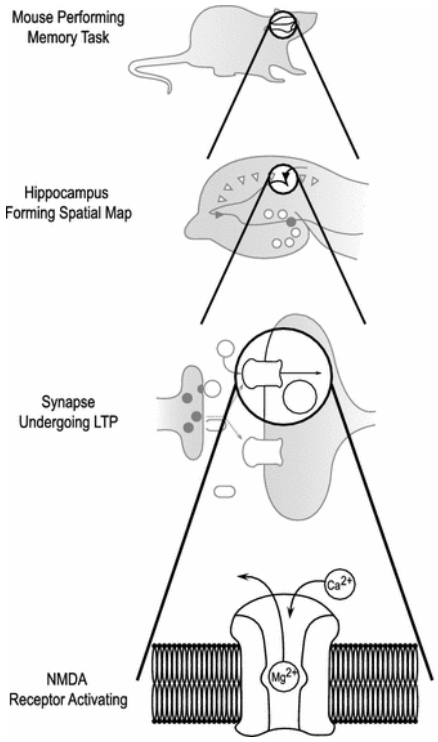

Figure 2: Multi-level neural mechanism of spatial memory (from Craver and Darden, 2013).

With the conceptual scaffolding of the mechanistic account, we can go back to our discussion about the existence of memory traces and ask whether there is a multi-level neural mechanism one can intervene upon in order to bring about differential effects in recollection. Since most interventionist techniques in neuroscience are relatively new—particularly those that afford controlled manipulations at specific mechanistic levels—the precise nature of such a mechanism is currently poorly understood. However, there are number of experimental results that can help to reveal its structure. First, consider manipulations at the molecular level. In a classic study, Flexner and colleagues (1963) injected intracerebrally several kinds of protein synthesis inhibitors in the hippocampi of stimuli-conditioned mice. They discovered that graded amounts of puromycin would impair the consolidation of recently acquired stimulus information. Unlike other pharmacological compounds used as control agents, Flexner and colleagues' discovery unveiled that when peptide transfer is disrupted in the ribosome of hippocampal cells, memory consolidation is impaired. This important experiment revealed part of the molecular level of memory traces by manipulating a specific component and bringing about a differential effect in recollection. Ever since, different pharmacological and genetic manipulations have been used, and although we are far from distinguishing the neural mechanism underlying an event-specific memory trace, recent manipulations looking at differential effects in content-specific memory traces suggest that the search is promising. Recently, Fellini and collaborators showed that NMDA receptors in another area of the hippocampus, CA3, are essential for pattern recognition tasks but not for spatial task, showing that controlled manipulations at the molecular level can illuminate the structure of content-specific memory traces.[36] Finally, parallel developments have occurred in the field of optogenetics, a revolutionary new technique that is enabling researchers to manipulate neuronal activity with light. Using optogenetics, researchers have been able to selectively turn neurons on and off in specific regions of the hippocampus, effectively switching the behavior of fear-conditioned rodents from active to inactive fear-conditioned behavior.[37]

Manipulations at the neurophysiological level have also shed light on the neural components of content-specific memory traces. The story begins with a widely cited study by Duncan (1949), in which he showed that electroconvulsive treatment (ECT) could impair the consolidation of recently acquired information—a finding that has been corroborated extensively.[38] Unfortunately, the effects of ECT are quite massive, and the precise reasons as to why they affect memory consolidation are unclear. Many neuroscientists hypothesize that ECT interrupts protein synthesis temporarily, which in turn affects the polarization of the cell membranes blocking the transport of neurotransmitters.[39] Fortunately, the depolarization component of the electroconvulsive shocks can now be isolated with the use of transcranial magnetic stimulation (TMS), a non-invasive experimental technique in which a rapidly changing magnetic field sends off a weak electric current to a specific region in the cerebral cortex in order to produce a localized depolarization. Although seldom used in the context of memory given the difficulty of stimulating the medial temporal lobes, recent studies have explored the way in which depolarization affects memory retrieval. In a recent study, for example, Kohler and collaborators (2004) employed repetitive TMS to stimulate regions in the left inferior pre-frontal cortex (LIPFC), which were previously associated with successful encoding of the studied material (using the subsequent-memory paradigm, which I explain below). Participants who were stimulated in the LIPFC showed higher accuracy for encoded words relative to both non-stimulated subjects and non-LIPFC stimulated subjects. Since it appears that repetitive TMS above 5 Hz transiently

increases cortical excitability[40]—an effect that parallels LTP—Kohler et al.'s study suggests that electric activity in the LIPFC is part of the mechanism underlying semantic memories.

Even more promising dissociations can be observed when we scale up a level. With the advent of non-invasive neuroimaging techniques, cognitive neuroscientists are starting to identify brain regions that are differentially involved during content-specific memory retrieval. Two important lines of evidence are of particular interest here. The first line of evidence pertains to findings employing the subsequent memory paradigm.[41] In this paradigm, participants are asked to memorize content-specific stimuli (e.g., words, pictures, etc.) while in the MRI scanner. The recorded brain activity during encoding is then compared with the participant's responses for subsequently remembered versus forgotten stimuli. The use of the subsequent memory paradigm in cognitive neuroscience has revealed a network of interrelated brain regions, whose engagement plays a critical role during the consolidation of memory traces effectively leading to the recollection of particular episodes.[42] The other line of research pertains to one of the most consistent results in the research on the cognitive neuroscience of memory: remembering re-activates the sensory areas that were involved during the encoding of the retrieved material.[43] The extent to which content-specific sensory cortices engaged during encoding are re-activated during retrieval has only recently started being studied. Still, the results from these studies consistently show that visual information selectively re-activates visual cortices, auditory information selectively reactivates auditory cortices, and olfactory information selectively reactivates the olfactory cortices.[44]

One final line of evidence that speaks to the nature of the top mechanistic level of memory traces comes from neuropsychology. Departing from the observation that visual cortices were engaged during retrieval of visual memories, cognitive neuropsychologists David Rubin and Daniel Greenberg studied the nature of memory deficits associated with selective damage in the visual cortex.[45] They observed that, consistent with the sensory reactivation hypothesis, patients with damage in the visual cortex have trouble remembering visual details of previously encoded events, leading to what is now called visual memory-deficit amnesia.[46] Importantly, the psychological manifestation of the visual memory-deficit amnesia differs from the typical medial-temporal amnesia—such as H.M.'s—in that it only affects visual information; episodic information encoded non-visually or amodally (e.g., names) is spared. Brain lesions do not constitute direct manipulations, however, for it is hard to say whether a particular patient would have remembered a specific stimulus had she not suffered the brain lesion. A more controlled experiment would be called for. For instance, combining the subsequent memory paradigm and the TMS techniques reviewed above, cognitive neuroscientists could localize those brain regions preferentially engaged during the successful encoding of different stimuli (say, faces and houses), and then, during retrieval, they could selectively TMS each region. One would expect, therefore, that if the brain region that gets activated during successful encoding of a particular face, *x*, is part of its memory trace, then by magnetically stimulating that very region one could control whether or not the subject successfully remembers having seen *x*. As such, this would be direct evidence to the effect that there is an intermediary causal mechanism between the successful encoding of an event (*Ex*) (i.e., seeing face *x*), and its subsequent remembering (*Rx*)—or failing to remember (not-*Rx*)—the event. Such an intervention—to go back to the discussion of the previous section—would allow the cognitive neuroscientist to "switch" the causal path from *B* to *C*, as illustrated in Fig. 1B.

In sum, the few studies I just surveyed offer us a picture of the ways in which neuroscientists have manipulated, and *could* manipulate, memory traces at different levels. The putative mutual manipulability of memory traces requires that interventions done at one level

affect the organization of the other levels. The fact that blocking NMDA receptors *but not* M2 receptors affects subsequent retrieval of content-specific memories [47] tells us that NMDA receptors, but not M2 receptors, are a part of that memory's trace. Likewise, if depolarizing the right occipital face area (rOFA) *but not* the right lateral occipital area (rLO) during recognition selectively impairs one's recognition of a particular face, this intervention would tell us that that the rOFA, *but not* the rLO, would be part of the memory trace of that face.[48] Taken together, the results of these studies are starting to give us a picture of the neural underpinnings of memory traces that resembles the multi-level structure of the causal mechanisms involved in cyanide poisoning. Suitable interventions at the right level of the neural mechanisms composing memory traces may give us the differential effects in recollection that the anti-realist was unable to explain.

7. Objections, and (quick) rebuttals

There are, of course, a number of ways one could challenge the current proposal. In this section I will briefly reply to three possible objections. A first objection stems from the fact that Woodward's manipulationist account of causation has difficulty accommodating preemption.[49] To understand how this concern applies to the case of memory traces, consider again Figure 1B. Here, the thought is that a proper intervention, *I*, would help to clarify whether or not *D* is part of the causal path to *B*. If we take *B* to be *Rx*, and *C* to be not-*Rx*, then showing that *I* switches from *B* to *C* suggests that *D* is (at least part of) *Mx*. But if *D* stands for a multi-level mechanism, then any intervention on *D* will really be an intervention on a putative causal path between variables inside *D*. Graphically, the actual model should look more like Figure 3A. Here, *X*, *Y* and *Z* are variables that stand in place of entities at some level of description of *D* (i.e., *Mx*). The intervention may block the causal path from *X* to *Y*, "switching it" instead to *Z*, which in turn causes *C*, rather than *B*. Suppose—to make this model more concrete and related to the previous section—that *Y* stands for NMDA receptors, *Z* for M2 receptors, and that *I* represents an intervention on the process of glutamate binding from afferent neurons projecting onto CA3. The problem, however, is that there could be a common factor, *P*, that can directly cause *X*, *Y*, and *Z*, and which the intervention, *I*, cannot rule out, even though it could have brought about *Y*, and thus *B* (Figure 3B). Such a result, therefore, would undercut our reason to believe that *X* is part of the causal path—the memory trace—for *B*. More concretely, it is possible that there could be a background factor that could enable the passing of positively charged ions through the cell membrane even if one were to block glutamate binding to NMDA receptors. If this were the case, we would be less compelled to say that NMDA receptors are part of the multi-level neural mechanism that composes the memory trace for *Rx*.

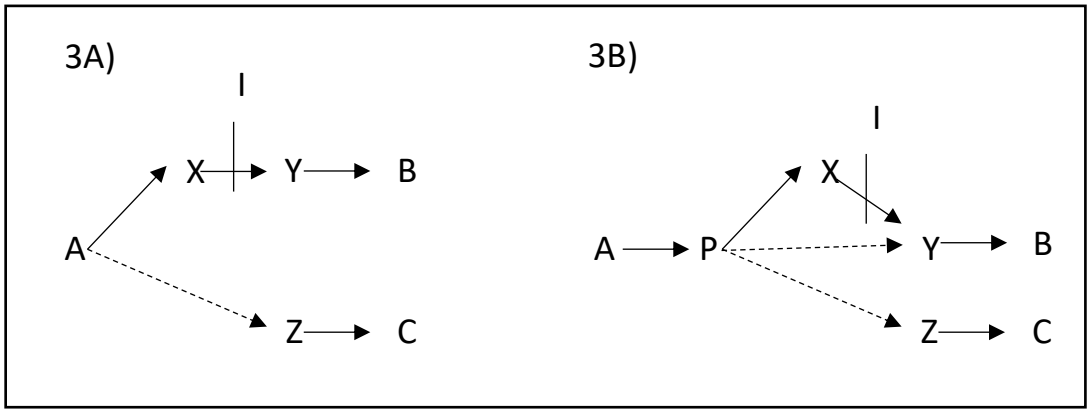

Figure 3: Graphical depiction of the problem of preemption for an interventionist model of memory traces.

I can think of two replies to this objection. A first, easy reply is simply burden-shifting: preemption is a problem for pretty much every theory of causation, so it is no more a problem for my preferred framework than it is for any other causal view. But a second, perhaps more interesting, reply is to point out that this is a case in which a difficulty in principle turns out to be an opportunity in practice. More precisely, while I agree that, in principle, the causal role of a common factor such as *P* would be missed by the intervention on the path from *X* to *Y*, for the practical purpose of identifying the memory trace that mediates *A*—i.e., *Ex*—and *B*—i.e., *Rx*—the question as to whether or not *P* is part of the multi-level neural mechanism simply becomes an empirical hypothesis. After all, our goal as scientists seeking to explain the causal structure of a memory trace, *Mx*, is precisely to uncover the causal net that links *Ex* to *Rx*. If our inquiry suggests that there is a reason to believe that *P*, and not *X*, is really part of *Mx*, then a suitable intervention on *P* would be the best way to test that hypothesis out. And since my proposal here is that, as a matter of empirical inquiry, we can think of memory traces as multi-level neural mechanisms within a manipulationist framework, concerns about common factors can be then assuaged by pointing out that they, too, can be subject to experimental manipulation.

A second objection is to suggest that I have misinterpreted Malcolm, and that one should read him as being concerned with a notion of remembering that covers *only* cases of successful recollection. As such, cases of unsuccessful recollection would not—according to this reading—fall under the concept of remembering Malcolm is concerned with. My response is that it is unclear how to delimit cases of successful, as opposed to unsuccessful, recollection. Consider another modified version of Malcolm's case of the boat capsizing. Al, Bea, and Cory witness the boat capsizing. Upon retrieval, Al remembers that the boat capsized and that the tip of the bow was red but can't recall whether the vessel flipped on its starboard or port side. Bea recalls that the boat capsized and that the tip of the bow was red, but falsely remembers that it flipped portside—it actually flipped on its starboard side. Finally, Cory remembers the boat capsizing, that it flipped on its starboard side, but falsely remembers that the tip of the bow was purple, not red. Who remembers the event successfully? One possibility is: nobody! They all got some detail either missing or wrong. Alas, if this was the case, most of our memories would be unsuccessful. We always miss some details of past events. Another possibility is to say that only Al remembers the event successfully, as his memory did not include falsehoods, only gaps. But doesn't it look like failing to conjure up the fact that a whole boat flipped on one of its sides is worse than misremembering that it did it rightwards versus leftwards, and isn't this mistake worse than misremembering the little detail that the bow was red and not purple? True, unlike Al's, both Bea and Cory's memories are false, but in some way they seem more accurate than Al's. Of course, what I'm trying to do with this case is to highlight well-known difficulties with the notions of truth, accuracy, and fidelity as they apply to episodic memory.[50] The fact that our episodic memories are not exact replicas of past experiences, and that a certain amount of distortion and forgetting is perfectly normal, is a well-accepted fact in the science of memory and, thankfully, in contemporary philosophy of mind. Unfortunately, this very fact makes the distinction between successful and unsuccessful remembering unclear, and thus unlikely that people operate with two entirely distinct concepts of remembering.

Finally, someone may offer a third objection to my manipulationist framework on account that many memory interventions that result in differential effects on recollection occur at the psychological and/or behavioral rather than at neural level. For instance, effects of divided

attention during encoding, or imagination inflation and misinformation at retrieval,[51] are well known cases of psychological interventions that bring about differential effects on recollection, even though they are not directly intervening the causal chain between *Ex* and *Rx* at the neural level. *Prima facie*, this objection is easily dismissible: given that I embrace the principle of mutual manipulability for multi-level mechanisms, all these effects ultimately are due to some change in the underlying neural structure. The fact that these interventions can be easily describable at the psychological or behavioral level does not mean that the underlying neural mechanisms are explanatorily inert. However, there is another concern in the vicinity of this objection, which has been recently articulated by Sarah Robins.[52] The concern here is reminiscent of the old philosophical distinction between the content and the vehicle of a mental representation. Memory traces, as we saw in section 2, are supposed to carry intentional content; *Mx* is about *x*. The vehicles of these contents, I have argued, are multi-level neural mechanisms. The problem is that current neuroscience tells us that neural mechanisms are in constant change; they undergo all sorts of dynamic changes through time.[53] If so, how can we attribute stability to an intentional content whose vehicle is dynamically changing through time? Moreover, how can an intervention in an everchanging vehicle predictably produce changes in the intentional content it supposedly carries? I think this concern is spot on, and my take is that it should force us to reconsider the *retention* and the *isomorphism* conditions mentioned in section 2. Doing so is, alas, beyond the scope of the current paper. Yet, it is a critical task if we want to understand the nature of memory traces, since, as I have argued, they continue to be explanatorily indispensable.

Acknowledgments

When working on this paper, many people gave me great advice. I want to particularly thank Santiago Amaya, Dorit Bar-On, Carl Craver, Daniel Dennett, Shenyang Huang, Bill Lycan, Maximiliano Martinez, Edouard Machery, Kevin O'Neill, Jesse Prinz, Sarah Robins, and Michael Strevens for their comments on previous drafts. I also want to dedicate this paper to John Sutton, whose book *Philosophy and Memory Traces* was pivotal in helping me decide to pursue research in the philosophy of memory. His work continues to be a source of inspiration.

Notes

1. Gomulicki 1953; Sorabji 2006.
2. De Brigard 2014.
3. For instance, Reid 1785/1849
4. James 1890.
5. Russell 1921) and Broad (1925) Semon 1904/1921
6. This is not to say that psychologists accepted the existence of memory traces by fiat. On the contrary, with the advent of methodological behaviorism, the notion of memory trace fell in disrepute (Watson 1930; Skinner 1953). Any mention of 'memory traces'—indeed, any mention of 'memory' as opposed to 'learning'—was practically 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" (Watson 1930; Skinner 1953). Many interpreted Lashley's results as evidence against the existence of memory traces. Nevertheless, subsequent developments—e.g. H.M.'s case (Scollville and Milner 1957), the discovery of LTP (Bliss and Lømo 1973)—reinvigorated the scientific quest for engrams, to the point that, today, most neuroscientists likely accept their existence (Thomson 2005).
7. Russell (1921), Ryle (1949), Wittgenstein (1953), and Benjamin (1956)

8. Malcolm 1963; Malcolm 1977; Martin and Deutscher 1966.
9. Woodward 2003.
10. see Robins 2017 for a recent review.
11. Malcolm 1977. See also, Bernecker 2008.
12. Squires 1968.
13. Heil 1978.
14. Sutton 1998; Michaelian 2011; De Brigard 2014; Robins 2016.
15. Watson 1930; and Skinner 1953.
16. Pincock 2006.
17. Wittgenstein 1980: 220.
18. Wittgenstein 1980: 905.
19. Malcolm 1963: p. 237. See also Munsat 1966.
20. Lipton 1991.
21. See Malcolm 1977.
22. Malcolm 1977: p. 171. Malcolm was not alone in thinking of memory traces as theoretical posits. Here is Heil, writing at around the same time: "It is important to see that the existence of traces is not supported by independent psychological or physiological evidence. Traces are postulated just because it is thought that their postulation provides an explanation for the phenomenon of memory—and perhaps other psychological processes as well. [Memory traces] are what once were called theoretical entities, devices introduced in the context of a theory to explain some more accessible phenomenon" (Heil 1978: 62). Importantly, the idea that memory traces are theoretical entities is still alive and well in contemporary philosophy of memory. Here is, for example, a recent excerpt from Bernecker (2008): "Memory traces are employed as theoretical entities, that is, as devices introduced in the context of a theory to explain some more accessible phenomenon. The status of the concept of *memory trace* is like that of *equator* or *center of gravity*. And when memory traces are taken to be theoretical constructs, to find fault with the theory of memory traces is to cast doubt on our need to postulate traces in order to account for remembering."
23. Malcolm 1977: p. 174.
24. Malcolm 1977: p. 179, emphasis in the original.
25. Malcolm 1977: p. 183.
26. Malcolm 1977: 185.
27. Coincidentally, the idea that memory traces may not be required for a successful explanation of how *Rx* can be brought about by *Ex* also received some attention in psychology, as it constituted the backbone of the ecological approach to remembering (Gibson 1979; Turvey & Shaw 1979; Michaels & Carello 1981). Much of what I say here could easily apply to this view. For a nice criticism of the ecological approach to memory, which I find very congenial to the spirit of this paper, see Sutton 1998, part IV.
28. To put it à la van Fraassen 1980: for Malcolm, an explanation involving action at a temporal distance is all one needs to save the phenomenon of recollection, so there is no reason to believe in the reality of the intermediary unobservable events supposedly referred by our notion of memory trace.
29. McGaugh and Krivanek 1970.
30. Woodward 2003. Woodward's view, of course, is not the only view about causal explanations. I will not defend his view against the usual contenders, but the reader is welcome to check Woodward and Hitchcock 2003, for that purpose.
31. Woodward 2003: p. 11.
32. Craver 2002; 2007.
33. See Craver 2001. For the original formulation, see Machamer et al. 2000.
34. Craver 2002; 2007.
35. Craver 2007: 141.
36. Fellini and collaborators 2009.
37. E.g., Liu et al. 2012; Ramirez et al. 2013. For a recent review, Josselyn et al 2015. For a philosophical view, Robins 2018.
38. Fraser et al. 2008.

39. Fink 1990.
40. Hallett 2000.
41. Wagner et al. 1998.
42. Paller and Wagner 2002.
43. Wheeler et al. 2000; Nyberg et al. 2000.
44. Wheeler and Buckner 2003; Gottfried et al. 2004; Woodruff et al. 2005. See Danker and Anderson 2010 for a recent review. Neuroimaging techniques such as fMRI and ERP are detection rather than intervention techniques. However, as I am go on to explain, when combined with intervention techniques, imaging methods can provide us with valuable causal information that we wouldn't have been able to gather otherwise.
45. Rubin and Greenberg 1998.
46. See Greenberg et al. 2005 for a review of 11 cases.
47. Patterson et al. 1990.
48. See Pitcher et al. 2009.
49. Strevens 2006; 2008.
50. See, for instance, Bernecker 2017; Robins 2016.
51. Brainerd and Reyna 2005.
52. Robins 2020.
53. De Brigard 2017.

References

Benjamin, B.S. 1956. "Remembering." *Mind* 65: 312-331.
Bernecker, S. 2008. *The Metaphysics of Memory*. Dordrecht: Springer.
Bernecker, S. 2010. *Memory*. Oxford: Oxford University Press.
Bliss, T.V. and Lømo, T. 1973. "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): 331-56.
Brainerd, C. J., Reyna, V. F. 2005. *The science of false memory*. New York: Oxford University Press.
Broad, C.D. 1925. *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.
Craver, C. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68: 53-74.
Craver, C. 2002. "Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory." *Philosophy of Science* Supplemental 69: S83-S97.
Craver, C. 2003. "The Making of a Memory Mechanism." *Journal of the History of Biology* 36: 153-195.
Craver, C. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.
Danker, J. F., Anderson, J. R. 2010. "The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding." *Psychological Bulletin* 136: 87- 102.
De Brigard, F. 2014. "The nature of memory traces." *Philosophy Compass* 9(6): 402-414.
De Brigard, F. 2014. "Is memory for remembering? Recollection as a form episodic hypothetical thinking." *Synthese* 191(2): 155-185.
De Brigard, F. 2017. "Cognitive systems and the changing brain." *Philosophical Exploration.* 20(2): 224-241.
Duncan, C.P. 1949. "The retroactive effect of electroshock on learning." *Journal of Comparative Physiological Psychology* 42:32–44.
Fellini L., Florian C., Courtey J., Roullet P. 2009. "Pharmacological intervention of hippocampal CA3 NMDA receptors impairs acquisition and long-term memory retrieval of spatial pattern completion task." *Learning & Memory* 16, 387-394.

Fink, M. 1990. "How does convulsive therapy work?" *Neuropsychopharmacology* 3(2): 83-7.
Flexner J.B, Flexner L.B, Stellar E. 1963. "Memory in mice as affected by intracerebral puromycin." *Science* 141:57–59.
Fraser, L.M., O'Carroll, R.E., Ebmeier, K.P. 2008. "The effect of electroconvulsive therapy on autobiographical memory: a systematic review." *The Journal of ECT* 24, 10–17.
Gibson, J.J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
Gomulicki, B.R. 1953. *The Development and Present Status of the Trace Theory of Memory*. New York: Cambridge University Press.
Gottfried, J.A, Smith A.P., Rugg, M.D., Dolan, R.J. 2004. "Remembrance of odors past: human olfactory cortex in cross-modal recognition memory." *Neuron* 42:687– 695.
Greenberg, D.L, Eacott M.J., Brechin D, & Rubin D.C. 2005. "Visual memory loss and autobiographical amnesia: A case study." *Neuropsychologia* 43(10):1493–1502.
Hallett, M. 2000. "Transcranial magnetic stimulation and the human brain." *Nature* 406(6792): 147-50.
Heil, J. 1978. "Traces of Things Past." *Philosophy of Science* 45: 60–72.
James, W. 1890. *The Principles of Psychology*. New York: Henry Holt & Co.
Josselyn S.A., Kohler S., Frankland P.W. 2015." Finding the engram." *Nature Review Neuroscience.* 16:521–34.
Köhler S., Paus T., Buckner R.L., Milner B. 2004. "Effect of left inferior prefrontal stimulation on episodic memory formation: a two-stage fMRI-rTMS study." *Journal of Cognitive Neuroscience* 16: 178-188.
Lipton, P. 1991. *Inference to the Best Explanation*. New York: Routledge.
Liu X., Ramirez S., Pang P.T., Puryear C.B., Govindarajan A., Deisseroth K., Tonegawa S. 2012. "Optogenetic stimulation of a hippocampal engram activates fear memory r ecall." *Nature* 484: 381–385.
Machamer, P.K., Darden, L., Craver, C. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67(1): 1-25.
Malcolm, N. 1963. *Knowledge and Certainty*. Ithaca: Cornell University Press.
Malcolm, N. 1977. *Memory and Mind*. Ithaca: Cornell University Press.
Martin, C.B., Deutscher, M. 1966. "Remembering." *Philosophical Review* 75: 161–196.
McGaugh, J. L., Krivanek, J. A. 1970. "Strychnine effects on discrimination learning in mice: effects of dose and time of administration." *Physiol. Behav.* 5, 1437–1442.
Michaelian, K. 2011. "Generative memory." *Philosophical Psychology* 24(3): 323–342.
Michaels, C. F., Carello, C. 1981. *Direct Perception*. Englewood Cliffs, NJ: Prentice-Hall.
Munsat, S. 1966. *The Concept of Memory*. New York: Random House.
Nyberg, L., Cabeza, R. 2000. "Brain imaging of memory." In E. Tulving & F. I. M. Craik (Eds.), *The Oxford Handbook of Memory*. New York: Oxford University Press.
Paller, K. & Wagner, D. 2002. "Observing the transformation of experience into memory." *Trends in Cognitive Science* 6(2): 93-102.
Patterson T. A., Lipton J. R., Bennett E. L., Rosenzweig M. R. 1990. "Cholinergic receptor antagonists impair formation of intermediate-term memory in the chick." *Behav. Neural Biology* 54, 63–74.
Pincock, C. 2006. "Richard Semon and Russell's Analysis of Mind." *The Journal of Bertrand Rusell Studies* 26:101-25.
Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., Duchaine, B. 2009. "Triple Dissociation of Faces, Bodies, and Objects in Extrastriate Cortex." *Current Biology* 19(4), 319-324.
Ramirez, S. Liu, X., Lin, P., Suh, J., Pignatelli, M,, Redondo, R.L., Ruan, T.J., Tonegawa, S. 2013. "Creating a false memory in the hippocampus." *Science* Vol. 341, Issue 6144, pp. 387-391
Reid, T. 1785/1849. *Essays on the Intellectual Powers of Man*. Edinburgh: McLachlan, Stewart, & Co.
Robins, S.K. 2016. "Representing the Past: Memory Traces and the Causal Theory of

Memory." *Philosophical Studies* 173, 2993–3013.
Robins, S.K. 2017. "Memory Traces". In S. Bernecker and K. Michaelian (Eds.) *Routledge Handbook of the Philosophy of Memory* (pp. 76–87). New York: Routledge.
Robins, S.K. 2018. "Memory and Optogenetic Intervention: Separating the engram from the ecphory." *Philosophy of Science* 85 (5), 1078–1089.
Robins, S. K. 2020. "Stable Engrams and Neural Dynamics." *Philosophy of Science* Forthcoming.
Rosen, D.A. 1975. "An argument for the logical notion of memory trace." *Philosophy of Science.* 42(1):1-10.
Rubin, D. C., Greenberg, D. L. 1998. "Visual memory deficit amnesia: A distinct amnesic presentation and etiology." *Proceedings of the National Academy of Sciences* 95, 5413-5416.
Russell, B. 1921. *The Analysis of Mind*. London: George Allen and Unwin.
Ryle, G. 1949. *The Concept of Mind*. Oxford: Oxford University Press.
Scoville, W.B., Milner, B. 1957. "Loss of recent memory after bilateral hippocampal lesions." *Journal of Neurology, Neurosurgery, and Psychiatry* 20:11–21.
Semon, R.W. 1904/1921. *The Mneme*. London: Allen & Unwin.
Skinner, B.F. 1953. *Science and Human Behavior*. New York: Macmillan.
Sorabji, R. 2006. *Aristotle on Memory*. London: Duckworth.
Squires, R. 1969. "Memory Unchained." *Philosophical Review* 78(2): 178–196.
Strevens, M. 2008. "Comments on Woodward, 'Making Things Happen.'" *Philosophy and Phenomenological Research* 77:171–192.
Strevens, M. 2007. "Review of Woodward, 'Making Things Happen.'" *Philosophy and Phenomenological Research*, 74:233–249.
Sutton, J. 1998. *Philosophy and Memory Traces: Descartes to connectionism*. Cambridge: Cambridge University Press.
Thomson, R.F. 2005. "In Search of Memory Traces." *Annual Review of Psychology* 56(1): 1-23.
Turvey, M.T, Shaw, R. 1979. "The Primacy of Perceiving: An Ecological Reformulation of Perception for Understanding Memory." In: L-G. Nilsson (ed.), *Perspectives on Memory Research: Essays in Honor of Uppsala University's 500th Anniversary*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
van Fraassen, B.C. 1980. *The Scientific Image*. Oxford: Clarendon Press.
Wagner, A. D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A., Rosen, B.R., Buckner, R.L. 1998. "Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity." *Science* 281: 1188-1191.
Watson, J.B. 1930. *Behaviorism*. Chicago: University of Chicago Press.
Wheeler, M.E., Petersen, S.E., Buckner, R.L. 2000. "Memory's echo: vivid remembering reactivates sensory-specific cortex." *Proceedings of the National Academy of Sciences of the United States of America* (97): 11125–11129.
Wheeler, M. E., Buckner, R. L. 2003. "Functional dissociation among components of remembering: control, perceived oldness, and content." *The Journal of Neuroscience* 23: 3869-3880.
Wittgenstein, L. 1953. *Philosophical Investigations.* New York: Blackwell.
Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology, Vol. I.* Edited by G. E. M. Anscombe and G. H. von Wright; translated by G. E. M. Anscombe. Oxford: Basil Blackwell.
Woodruff, C.C., Johnson, J.D., Uncapher, M.R., Rugg, M.D. 2005. "Content specificity of the neural correlates of recollection." *Neuropsychologia* 43, 1022–1032.
Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, J., Hitchcock, C. 2003. "Explanatory Generalizations, Part I: A Counterfactual Account." *Nôus* 37: 1–24.