Causation in Memory: Necessity, Reliability and Probability

Causação na Memória: Necessidade, Confiabilidade e Probabilidade

Abstract: In this paper, I argue that causal theories of memory are typically committed to two independent, non-mutually entailing theses. The first thesis pertains to the necessity of appropriate causation in memory, specifying a condition token memories need to satisfy. The second pertains to the explanation of memory reliability in causal terms, and it concerns memory as a type of mental state. Post-causal theories of memory can reject only the first (weak post-causalism) or both (strong post-causalism) theses. Upon this backdrop, I examine Werning's (2020) causalist argument from probabilistic correlation. I argue that it doesn't establish the necessity of appropriate causation, and thus it can only target strong post-causalist theories. I end up by presenting some general considerations, suggesting that memories may not always be causally linked to past experiences.

Keywords: memory; causation; reliability; probability; Reichenbach Common Cause Principle

Introduction

Causal theories of memory are motivated by a pair of closely related intuitions. Memories seem to *depend* on past experiences. To remember an event, you must have experienced it before.[1] Thus, you can remember a trip you took last summer but not a trip your family took without you. Moreover, past experiences seem to play a role in *bringing about* memories. You can remember the trip *because* you've experienced it before: your experiences then contribute to your experience of remembering now. The core idea of causal theories – that memories are *caused* by past experiences – aims to do justice to these widely shared intuitions. Appealing to a causal connection, the theories purport both to specify what it is to say that memories are brought about

[1] In this paper, I will deal only with *episodic* remembering, which is almost universally characterized as a state/process of entertaining quasi-sensory representations of particular past events. (See Andonovski (2020) for an introduction to, and criticism of, this dominant view.) For convenience, I will often drop the "episodic" and speak only of "memory" or "remembering".

by past experiences *and* to explain why they depend on them. Accordingly, in the context of different theoretical pursuits, the core idea can be developed in a variety of ways. In fact, even in the development of particular causal theories, the idea has given birth to two distinct theses that are rarely distinguished. The first pertains to the necessity of appropriate causation for memory, while the second concerns the explanation of memory reliability in causal terms. It is the wager of this paper that the two theses are *not* mutually entailing, and that an appreciation of this fact is required for a proper understanding of what is at stake in the debates about memory causation.

I proceed as follows. In section 1, I introduce the two causalist theses and argue that they are independent. I then distinguish two kinds of post-causal theories of memories: weak post-causalism rejects only the necessity of appropriate causation, while strong post-causalism rejects both. In section 2, I examine Werning's (2020) recent argument for the necessity of causation in memory. I argue that, even if sound, the argument does not establish the necessity of appropriate causation. Hence, it can only target strong post-causalist theories. In section 3, I present some general considerations to support the intuition that, at least in some normal cases, our memories may not be causally linked to past experiences.

1. Causation in Memory: Between Analysis and Explanation

In this opening section, I introduce the two causalist theses: the first concerning the necessity of causation in memory, the second the causal explanation of memory reliability. After establishing that the theses are independent, I use this to distinguish two kinds of post-causal theories of memory.

The necessity thesis is often characterized as the defining feature of causal theories of memory,[2] and can be seen as a natural development of the core causalist idea.[3] Bracketing some of the idiosyncrasies of the theories, we can characterize it in the following way:

(NECESSITY) A subject's memory of a particular past event must be appropriately causally connected to their past experience of the event.

[2] For example. Michaelian & Robins (2018, p.14) tell us that Martin & Deutscher's classic theory qualifies as causal *in virtue* of stipulating that a causal connection between a memory and a relevant past experience is necessary.
[3] See, e.g., Martin & Deutscher (1966, pp. 173-183); Bernecker (2010, Ch.4); Michaelian (2011, pp.330-336).

The thesis made its way to philosophy of memory in the context of conceptual analysis, traditionally understood as the pursuit for sets of constitutive – individually necessary and jointly sufficient – conditions (Von Leyden 1961; Martin & Deutscher 1966). It aims to specify a condition *token* mental states have to satisfy in order to belong to the relevant mental state *type* – memory. The condition, as Robins (2021) points out in a recent essay, is important for identifying all and only states of genuine remembering, weeding out "borderline and/or degenerate cases." (p.215). It is supposed to weed out states of mere imaginings in disguise (*not* causally linked to relevant past events) as well as states of relearning (not causally linked in an *appropriate* way).[4]

The truth of NECESSITY has been a central concern of recent philosophy of memory. Despite this, there has been very little attention paid to the notion of necessity appealed to in the thesis. At first glance, theorists like Martin & Deutscher (1966) and Bernecker (2010) seem to appeal to a *conceptual* necessity, holding simply in virtue of the meanings of our concepts. On this reading, the proposal that a memory must be appropriately causally linked to a past experience is like the one that a bachelor must be unmarried. Second glances complicate the picture, however. As Zemach (1983) convincingly argues, by introducing specific conditions for appropriateness – e.g., that a causal connection is appropriate only if sustained by a memory trace that is a structural analog of the past experience (1966, pp.186-191) – Martin & Deutscher seem to venture an empirical hypothesis about the kind of world we inhabit. This commitment is explicitly acknowledged by Bernecker (2010), who considers the linguistic intuitions guiding his method "empirical working hypotheses" (pp.31-34).[5] For causalists like Werning (2020), interested in the natural functioning of memory systems, it is even clearer that the appeal is to some notion of *(meta)physical* necessity, holding in virtue of the meaning of our concepts *and* the laws/regularities governing such systems. On this reading of NECESSITY, the proposal that a memory must be appropriately causally linked to a past experience is – in some *to-be-determined* sense – like the one that a person jumping on Earth must come down. (Michaelian (2016a), who is the principal critic of the thesis, likely has a similar notion in mind.) Hence, while it is probable that the various notions will stand in relevant entailment relations, we should nevertheless proceed with care when

[4] On most causal theories, a causal connection between a memory and a past experience is appropriate just in case it is sustained by a *memory trace* – a stored mental representation that carries content about the past event and is causally operative in producing the memory (Martin & Deutscher 1966; Bernecker 2010). For causal theories that do *not* posit content-bearing memory traces, see Perrin (2018) and Werning (2020).

[5] Interestingly, Bernecker (2010) considers the conditions he posits for remembering *"at best* necessary conditions" (p.31, emphasis original). By this, however, he only seems to mean that the necessities appealed to will not be of the logical/conceptual kind (ibid). If the "at best" is meant to refer to a more comprehensive methodological principle, he doesn't tell us what this principle is.

discussing the status of NECESSITY.[6] In addition, a good case can be made that the thesis should be understood as involving a *ceteris paribus* clause and be taken to hold only in conditions in which the subject's memory systems functions normally (cf. Andonovski 2021a). With all the caveats and qualifications on the table, proponents of NECESSITY will undoubtedly agree about one thing.[7] At least in the actual world, every memory of a particular past event, entertained by some subject whose memory functions normally, is appropriately causally connected to the subject's past experience of the event.

The causal reliability thesis is likewise a natural development of the core causalist idea, and it appears in some form in all major causal theories.[8] Unlike NECESSITY, however, the thesis does not aim to specify a constitutive condition token memories must satisfy. It is rather an *explanatory* thesis about the reliability of memory processes:[9]

(CAUSAL RELIABILITY) The reliability of memory processes is best/only explained in causal/informational terms – i.e., by positing a causal connection between memories and past experiences.

Following Goldman (1986) and Michaelian (2011), we can characterize a memory process as reliable just in case it tends to produce accurate representations in large proportion – i.e. (significantly) greater than 50% – of the cases.[10] Since accuracy is a graded notion, the assessment of memory reliability requires the introduction of a threshold, which specifies how closely the content of a memory representation must resemble a represented event for the memory to count as accurate, along with a characterization of the relevant dimensions of accuracy. In a paper that will occupy us in the next section, Werning (2020) provides the necessary tools. On his view, which we can adopt here, a memory representation is sufficiently accurate just in case the aggregate –

[6] For a discussion of the relations between different notions of necessity, as well as an important criticism of Kripke's arguments for the distinctness of *metaphysical* necessity, see Priest (2021).
[7] If you are skeptical that proponents and critics of NECESSITY understand the thesis in similar ways, you and I will get along well. Please try to suppress your skepticism for the duration of this paper, however. As I hope to show, theorists who are after a *(meta)physical* NECESSITY thesis don't have good *a priori* reasons to accept even the weakest version of it.
[8] See, e.g., Martin & Deutscher (1966, pp. 175-177); Bernecker (2017, pp.8-9); Werning (2020, pp.303-309 & 321-323).
[9] Hence, this explanatory thesis should be distinguished from Goldman's (1986)'s *epistemological* thesis, which aims to analyze the justification of beliefs in terms of the reliability of the process that generates them.
[10] As noted above, Goldman (1986) defines reliability in terms of the tendency of the process to produce *true* beliefs. I have opted for accuracy here because I agree with Sant'Anna (2018) that this graded notion is a better fit for episodic memory. However, since the assessment of reliability requires the introduction of an accuracy threshold (see the main text), this choice doesn't significantly affect the character of the resultant thesis.

spatial, temporal and qualitative - resemblance between its content and the event exceeds some specified, and potentially context-sensitive, threshold.[11]

The motivation behind CAUSAL RELIABILITY is straightforward and is often presented in intuitive terms. If I keep on winning the lottery, getting the numbers right in a large proportion of the cases, it is *very* likely that there is a causal process – presumably one involving the *transmission* of information – linking my guesses to the generation of the numbers. Either I have somehow rigged the generating system, got access to its outputs, or someone is simply telling me the numbers. Similarly for memory: our memories tend to be accurate in large proportion of the cases (only) *because* they are connected to the events they represent by some information-bearing causal process.[12] (This is indeed why remembering is, *ceteris paribus,* much more reliable than imagining the future.) The more reliable the memory process is, and the more accurate memories tend to be, the more likely a causal connection between memories and past experiences is.

As the definition above may intimate, CAUSAL RELIABILITY comes in two flavors. On the weak version, the reliability of memory is *best* explained in causal/informational terms, yet reasonable alternative explanations may exist. On the strong version, in contrast, the causal explanation of reliability is the *only* such explanation. We see a nod to this thesis in Martin & Deutscher's characterization of the painter who, without knowing, paints a detailed scene from his childhood. Given such accuracy, they tell us, "the only reasonable explanation…is that he [is] remembering" (1966, p.167), by which they mean that the painting is causally/informationally linked to the painter's childhood experience. Yet, it is Bernecker (2017) who presents the strong thesis most clearly:

> [H]ow does our memory system manage to reliably produce accurate representations? If this question is not answered in terms of a causal process connecting the past and present representation, then, as far as I can see, *we are left with a picture whereupon there is a remarkable correlation between our memory representations and past events but nothing to explain the correlation* (p.9, emphasis added).

[11] In his paper, Werning talks of *truth*-closeness ("verisimilitude") between the representational content of a memory and *some* actually occurring event, which he defines in terms of similarity greater than a specified threshold. The differences in terminology don't matter for my purposes here.

[12] Causal theories typically consider the causal connection between memories and past experiences to involve transmission of *content* (e.g., Martin & Deutscher 1966; Bernecker 2010). But even theorists skeptical of attributions of content accept that mnemonic causal connections involve transmission of *information* (e.g., Werning 2020).

In the absence of any reasonable alternatives, positing causal connections between memories and past experiences becomes *explanatorily* necessary. Causal explanations of memory reliability, on this view, are the only game in town. No reliability without causation.

Yet, this formulation of CAUSAL RELIABILITY tends to obscure an important aspect of the thesis. On CAUSAL RELIABILITY, what is explained in causal terms is the *tendency* of memory processes to produce accurate event representations in large proportion of the cases. If memory is highly reliable, then there must be *some* underlying causal mechanism that explains this fact. In addition, in purportedly many cases, the accuracy of a particular memory will be due to a causal/informational link to the past. The more accurate the memory is – or the more "unusual" the remembered features of an event are (Martin & Deutscher 1966, p.177) – the more likely is that there is some causal connection to the event.[13] Yet, even in its strong form, CAUSAL RELIABILITY does not entail that the accuracy of *every token* memory will require positing a causal connection to a (specific) past experience. Winning the lottery legitimately is, after all, (meta)physically possible, even if it is vanishingly unlikely. So is, analogously, representing an event highly accurately without drawing on information from one's past experience. In normal circumstances, however, we frequently accept as accurate memories whose contents are *very* "imperfectly" similar to the represented events.[14] In cases of this kind, which arguably constitute a large class, the likelihood of accurate memories not causally linked to the past events is significantly higher. (The comparison with imagining the future may again be illustrative. Given a sufficiently low, yet realistic, accuracy threshold, one *can* (sometimes) accurately represent future events; cf. Williamson 2016; Michaelian 2016b.) Martin & Deutscher are indeed aware of this, clarifying that in *some* cases of memory accuracy, alternative explanations do exist, even if they are "comparatively unlikely" (1966, p. 177).

The preceding discussion should hint at why CAUSAL RELIABILITY does *not* entail NECESSITY. Three key points are worth highlighting. First, the causal explanation of memory reliability has no direct bearing on our concepts of remembering. So, on a conceptual understanding of necessity, the entailment clearly does not hold. Second, CAUSAL RELIABILITY also does not entail the (meta)physical necessity of causation in memory. As was illustrated above, it is (meta)physically possible that even a "perfectly" accurate memory is not

[13] Note, however, that this causal link may not be of the "appropriate" kind. See the discussion in the main text below.
[14] The accuracy threshold, in many contexts, cannot be set too high since many ordinary memories will be weeded out. I'll return to these issues in the next section when I discuss Werning (2020)'s argument from probabilistic dependence.

causally linked to a past experience. For "imperfectly" accurate memories, which probably constitute the vast majority of genuine memories, the possibility may indeed be very realistic. (I will return to this issue in the sections 2 and 3.) Finally, and perhaps crucially, even if CAUSAL RELIABILITY were to entail the necessity of causation in memory, it would still not entail the necessity of *appropriate* causation. The reason is simple. Many causal chains that *could* explain the accuracy of an event representation will be of the inappropriate, "deviant" sort. A classic case involves writing down your experiences, forgetting all about them, and then reading about them years later (cf. Robins 2017). In this case, there *is* an information-bearing causal connection that can explain the subject's representational accuracy, yet it is not of the right sort. Indeed, cases of *relearning,* as traditionally defined, will typically have this structure.[15] The existence of explanatorily relevant deviant chains illustrates most clearly that CAUSAL RELIABILITY does not entail NECESSITY. Perhaps less interestingly, the converse entailment also does not seem to hold. For example, one may insist that, as a matter of conceptual necessity, memories are appropriately caused by past experiences, yet refuse to take a stance on the explanation of memory reliability. In fact, such an explanatory agnosticism may be available even on a (meta)physical reading of NECESSITY. It is, as far as we know, possible that memories are necessarily causally linked to past experiences, yet that their accuracy is not explained in causal terms. While available in principle, however, this position is unlikely to be held by a causal theorist.

With these points on the table, we can briefly survey the landscape of causal and post-causal theories. It is clear that traditional causal theorists – like Martin & Deutscher (1966) and Bernecker (2010) – endorse *both* NECESSITY and CAUSAL RELIABILITY. It is less clear, however, whether neo-causalists like Werning (2020) do so as well.[16] In the next section, I will argue, on by then familiar grounds, that Werning's detailed argument for CAUSAL RELIABILITY does not establish NECESSITY. Yet, this does not show a lack of commitment to the latter thesis, which may be defended in other ways. A more meticulous exegesis is required to ascertain Werning's stance on the necessity of appropriate causation.[17]

---

[15] Deviant causal chains, incidentally, need not involve externalization of information. There may be cases in which a subject accurately represents a past event but relies solely on general information she has acquired on different occasions. I will examine these in section 3.

[16] Other neo-causalists, like Perrin (2018, 2021), are more clearly committed to NECESSITY, even if they do "not yet provide a description of what it is for such a causal connection to be *appropriate*" (Michaelian & Robins 2018, p.22).

[17] A careful reading of Werning (2020) doesn't help *much* in this regard. The word "appropriate" appears only once in a relatively long paper, and only in the course of criticizing the classic causal theory (p.316). For that theory, of course, the appropriateness of a causal connection is secured by a *content*-bearing memory trace. The question then is whether Werning's "minimal" traces, which carry sub-categorial information but *not* content (p.321), can secure a causal connection's appropriateness. My money is on *no*.

THEORIES OF MEMORY	(NECESSITY)	¬ (NECESSITY)
(CAUSAL RELIABILITY)	**Causal Theories of Memory** (Martin & Deutscher 1966; Bernecker 2010; Werning 2020?)	**Weak Post-Causalism** (FTM - Fernández 2019; STM? - Michaelian 2016a, 2016c)
¬ (CAUSAL RELIABILITY)	/	**Strong Post-Causalism** STM? – (Michaelian 2016a, 2016c)

Table 1. Causal and Post-Causal Theories of Memory.

In the other camp, things are also more complicated than they may appear. Michaelian & Robins (2018) characterize as *post*-causal theories of memory that reject NECESSITY. The discussion above, however, allows to distinguish two kinds of such theories. *Weak* post-causal theories reject NECESSITY but *not* CAUSAL RELIABILITY. *Strong* post-causal theories, in contrast, reject both theses. Fernández's (2019) functional theory of memory (FTM) presents a seemingly paradigmatic example of weak post-causalism.[18] On FTM, a token mental state is a memory just in case it plays the proper functional role, defined only in terms of the *tendency* to be caused in the appropriate way. While memories are typically (appropriately) caused by past experiences, a *particular* mental state need not be caused in such a way to qualify as a memory (2019, pp.47-53). Despite the apparent rejection of NECESSITY, however, Fernández's discussion makes it quite clear that the causal explanation of memory reliability is a virtue of causal theories that FTM seeks to inherit (see, e.g., pp.36-37). The functionalist thus endorses CAUSAL RELIABILITY.

[18] In Andonovski (2021a), I propose that Fernández's FTM may only reject the necessity of appropriate causation in *abnormal* circumstances. Depending on whether one takes NECESSITY to include a *ceteris paribus* clause, then, FTM may turn out not to be a post-causal theory at all. I am bracketing this issue here. If you agree with the proposal in (Andonovski 2021a), however, you can place FTM in the causalist camp.

What about Michaelian's (2016a; 2016c) simulation theory of memory (STM)? On STM, a mental state is a memory just in case it is produced by a properly functioning – read: reliable – episodic system "aiming" to represent an event from the subject's personal past (2016a, p.107). Full stop. No appropriate causal link to a past experience is necessary; in fact, *no* causal link to a past experience is necessary (pp.110-113). Michaelian has expressed this commitment frequently, and with much verve, so it may be tempting to simply assume that STM belongs in the strong post-causalist camp. While the textual evidence is scant, there are three reasons to suspect that this is not the case. First, with the definition above, STM aims only to specify the conditions a token mental state must satisfy in order to *qualify* as a memory – i.e., the constitutive conditions for memory. This does not entail that STM cannot use the same explanatory resources causal theories do in explaining memory accuracy. Second, Michaelian has indeed gestured towards such use on multiple occasions – e.g., in explaining cases of "misremembering" (Michaelian 2016c). In such cases, he accepts, the accurate representation of details from past events is due to retention of information, presumably through an appropriate causal chain (pp.9-10). In fact, STM may favor similar explanations in *many* cases of genuine remembering (2016a, p.103). Third, STM does not rule out *deviant* chains as explanatorily relevant. In fact, many cases in which testimonial information is incorporated – a go-to simulationist example (see 2016a, Ch.7) – will involve causal chains that *can* explain a memory's accuracy. With all this said, there are also reasons to be careful. For one, Michaelian does entertain the possibility of explaining reliability without appealing to retention (see, e.g., 2016c, pp.7-9), even if he doesn't tell us much about it. For another, some of his arguments – like the "anti-necessity" argument (2016a, pp.103-104) – rely on the premise that the episodic system functions in a *very* similar way in remembering and imagining the future. It is not clear whether this premise is compatible with weak post-causalism. But we need not settle this issue here. What matters is that the two theses, determining whether STM is weakly or strongly post-causal, are *independent.* This will matter for assessing arguments against the theory.

Summing up, causal theories of memory are typically committed to two non-mutually entailing theses: the first posits the necessity of appropriate causation in memory, the second the need to explain memory reliability in causal terms. Post-causal theories reject only the former or both theses.

2. The Common Cause Principle, Reliability and Necessity

In this section, I examine Werning's (2020) recent argument for the necessity of causation in memory. I argue that, even if sound, the argument only establishes CAUSAL RELIABILITY, not the necessity of causation – and thus *not* NECESSITY. Accordingly, it can only target strong post-causalist views. I end the section by discussing the relation between probabilistic correlation and memory causation.

Werning (2020) situates the discussion of memory reliability in the established tradition of explaining statistical relevance in causal terms. He shows that the reliability of memory processes amounts to a probabilistic correlation between a memory and the event it represents. Then, he appeals to a general principle – the famous *Reichenbach Common Cause Principle* (RCCP) – to argue that such correlation *requires* a causal connection between the memory and the event. The core idea of RCCP was introduced by Reichenbach (1971), but the principle was articulated more clearly and explicitly only later, most notably by Salmon (1984, 1998). In its most general formulation, it simply tells us that every correlation is due to some causal link connecting the correlated entities. No correlation without causation.

Yet, the notion of correlation appealed to in RCCP is, strictly speaking, only meaningful in the framework of a specified probability space (see Hofer-Szabó et al. 2013). Given a *classical* probability measure space *(X, S, p),* where X is a set of elementary events, S is a Boolean algebra of some subset of X, and p is a probability measure, the events $A, B \in S$ are positively correlated if: [19]

$$p(A \cap B) > p(A)p(B) \tag{1}$$

In words, A and B are positively correlated if the probability that *both* events occur is greater than the product of the individual probabilities.[20] According to the Common Cause Principle:

[19] RCCP applies to both positively and negatively correlated events. Here, I only focus on positive correlations.
[20] Reichenbach (1971, p.123) defined objective probability as the limit of a relative frequency. RCCP, however, does not depend on this conception. It only requires that objective probabilities can be assigned to events in *some* way.

(RCCP) If *A, B* are positively probabilistically correlated, then one of the following causal relations exists: *A* is the cause of *B, B* is the cause of *A,* or both *A* and *B* are caused by a third event *C,* which occurs prior to *A* and *B* and satisfies a set of independent conditions.[21]

Essentially, RCCP tells us that probabilistic correlations between events are ultimately derived from causal relationships. If *A* and *B* are (positively) correlated, then either they are directly causally linked or they are both caused by a third event *C.* An example, first presented by Reichenbach (1971, p.157) and developed further by Hitchcock (1998), can illustrate the principle's intuitive plausibility. Imagine a theater troupe travelling the country. Occasionally, the leading man falls ill with gastric distress – call this event *M.* The same thing sometimes happens to the leading lady – event *L.* Even though both *M* and *L* are rare, they tend to happen together. That is, the events are probabilistically correlated: $p(M \cap L) > p(M)p(L)$. In such circumstances, RCCP tells us, we are justified in inferring that: either one of the actors falls ill and infects the other or both actors fall ill for the same reason (e.g., eating the same tainted food at a restaurant). In any case, what *explains* the correlation between the events is an underlying causal relation. This points to the significance of the Common Cause Principle. RCCP allows us to infer causal relationships from observable statistical correlations between events (Hofer-Szabó et al. 2013; Hitchcock & Rédei 2021).

Werning's (2020) argument, based on RCCP, is intuitively easy to grasp. If our memory system is generally reliable – a proposition that both causalists and post-causalists accept – then an event representation produced by the system has a high probability of being accurate.[22] So, an act of entertaining such a representation – of *seeming* to remember – increases the probability that the event occurred as represented.[23] (If you have a seeming memory of a fun hike in the mountains, then it is likely that the fun hike took place.) On the flip side, the occurrence of an event increases the probability that the event will be the target of a memory representation. Hence, events and

[21] The event *C* – the Reichenbachian *common cause* – must satisfy four such conditions. The first two specify that *A* and *B* are conditionally independent given *C* as well as given the absence of *C*. The second two specify that *A* and *B* are more probable, conditional on *C,* than conditional on the absence of *C.* The conditions do not matter for my purposes, so I will not discuss them.
[22] I will use "memory system" and "memory process" interchangeably in what follows.
[23] With "seeming to remember", I refer to the act of entertaining an event representation produced by a normally functioning memory system. Thus, all instances of genuine remembering are instances of seeming to remember. Yet, in some cases when we seem to remember, we are not genuinely remembering (i.e., the representation we entertain is not accurate). I am not committed to the view that the subject that seems to remember must, in any strong sense, *feel* like she is remembering (unlike Robins 2020).

states of apparent remembering are positively probabilistically correlated; indeed, they are likely *strongly* positively correlated. Given, RCCP, then, there must be some underlying causal relation that explains the correlation. Barring retro-causality and unlikely common causes, there is only one option on the table: events must cause (seeming) memories.

To establish the correlation, however, we first need to characterize the events whose probability is measured.[24] Consider a type of occurrence *E* that can be represented in memory (a hike, a trip, a flash of light, an itch etc.). We will call entertaining a memory representation with content ⌜E⌝ - event $R⌜E⌝$.[25]A token mental state will belong to the type of event $R⌜E⌝$ iff the representation entertained is produced by a normally functioning memory system *S and* has content ⌜E⌝.[26] Recall that a memory representation is accurate just in case its content sufficiently resembles a past event. Werning (2020, pp.306-307) makes this idea precise, introducing a satisfaction condition for memory accuracy, i.e. $A(R⌜E⌝)$:[27]

$$A(R⌜E⌝) \Leftrightarrow \exists x.Occur_{@}(x) \;\&\; d(e,x) \geq \vartheta_{S,E.} \qquad (2)$$

In (2), *e* is an event that belongs to the type *E*, $\vartheta_{S,E}$ $(0.5 < \vartheta_{S,E} \leq 1)$ is a (context-sensitive) accuracy threshold, and $d : E \times E \rightarrow [0.1]$ is a similarity measure that takes into account relevant spatial, temporal and qualitative (STQ) properties of events. In words, a memory representation is accurate just in case "some event actually occurs/occurred whose STQ-similarity to the event represented exceeds the [accuracy] threshold" (p.307). Let's call an event that is sufficiently STQ-similar to the content ⌜E⌝ - event *E'*.

Since accuracy requires occurrence, the probability that *E'* has occurred given an *accurate* memory is 1. This is the case whether or not memory processes are reliable. What the reliability

[24] In what follows, my exposition diverges from Werning's. One change is important: Werning doesn't define a proper event space and ends up comparing the probabilities of elementary events and *non*-Boolean relations between them (e.g., on page 322, he compares the probabilities of existentially quantified propositions). I have amended the exposition to remedy this problem. I nevertheless end up at the same place, however, with *one* major difference. Namely, Werning assesses the probability of an *occurrent* STQ-similar event *e'* (ibid). But that is a mistake. What needs to be assessed is the probability that such an event – I use the notation for event types – *E' has* occurred given an episode of seeming remembering.

[25] Following Werning, I will use corner quotes to signify the representational content of a state. Thus, ⌜E⌝ refers to event *E* (cf. 2020, pp.306-307). Werning doesn't tell us (a) how he understands the notion of representational content as well as (b) how the representational content of a memory is determined. If you think the silence on these issues is bound to raise a number of downstream problems, you are in good company. I am trying not to ruffle *too many* feathers here, however, so I will be as faithful to Werning's presentation *as possible* (see the previous note).

[26] Please keep in mind that the episode of seeming to remember is itself an event in *S* (with some probability).

[27] As I noted before, Werning talks of *truth*-closeness, or "verisimilitude". As far as I can see, the difference is only terminological.

of memory tells us, however, is that the probability of a particular memory representation *being* accurate is high:

$$p(E' | R\ulcorner E\urcorner) \geq \rho_{S,E} \quad (3)$$

That is, the probability that *E' has* occurred given an apparent memory with content $\ulcorner E\urcorner$ is higher than a certain threshold probability ($\rho_{S,E}$). By definition, this threshold probability is significantly higher than .5. Accordingly, it is almost certain that:

$$p(E' | R\ulcorner E\urcorner) > p(E') \quad (4)$$

In words, the occurrence of $R\ulcorner E\urcorner$ – i.e., a subject entertaining a representation, with content $\ulcorner E\urcorner$, produced by a memory system – *increases* the chance that an event *E'* has occurred. Given the definition of conditional probability, the inequality in (4) is logically equivalent to (5):

$$p(E' \cap R\ulcorner E\urcorner) > p(E')p(R\ulcorner E\urcorner) \quad (5)$$

And, in (5), we have the familiar inequality that features in the antecedent of RCCP. *E'* and $R\ulcorner E\urcorner$ are positively probabilistically correlated. If RCCP is true, then, this correlation ultimately derives from an underlying causal relation between *E'* and $R\ulcorner E\urcorner$. Assuming that $R\ulcorner E\urcorner$ does *not* cause *E'* (a reasonable assumption), then either *E'* causes $R\ulcorner E\urcorner$ or *E'* and $R\ulcorner E\urcorner$ have a common cause. Werning surveys some common cause scenarios and argues – in my book, convincingly – that a common cause is unlikely (pp.323-324). So, the only option left on the table is that *E'* causes $R\ulcorner E\urcorner$. That is, if memory is reliable, then there must be a causal connection between mental states of type $R\ulcorner E\urcorner$ and past events. Since, by definition, remembering is a state of type $R\ulcorner E\urcorner$, it must be caused by (a) past event(s).

Werning concludes from this that "a causal connection to [a past event] is still *necessary* to fulfill even the minimal requirement…of reliability" (p.301). On the face it, this looks like a defense of CAUSAL RELIABILITY. The Common Cause Principle tells us that causal explanations of memory reliability are indeed the only game in town. If memory is reliable, then there must be a causal mechanism connecting memories and past experiences. Yet, Werning offers

additional remarks which suggest that the argument is not only intended as a defense of CAUSAL RELIABILITY. Thus, he tells us that his own "trace minimalist" view shares a core commitment with traditional causal theories: that "there is no episodic memory without a causal link to experience" (p.325). Similarly, in describing the cognitive machinery that undergirds the production of memories, he is adamant that "a causal link to experience is *necessary*" (p.328, emphasis added). Here, Werning appears to specify a condition *token* mental states need to satisfy in order to be memories. To put appropriateness to the side just for a moment, he appears to be defending NECESSITY.[28] In fact, Werning uses the RCCP argument to point to an alleged instability of the simulation theory:

> Simulationism turns out to be an unstable position. It tries to hold on to the reliability condition on memory… and at the same time rejects the need for a causal link between experience and remembering… [Yet], the reliability of an episodic memory…requires a direct or indirect causal connection between the remembered event and the episodic memory (p.329).

In short, STM can't afford to deny the necessity of causation in memory, since the reliability of memory systems/processes requires it. Even the weakly post-causalist STM, then, is not feasible.

With the accumulation of putative counterexamples to the Common Cause Principle (e.g., Sober 1988; Cartwright 1988; Schurz 2017), it may be tempting to simply reject its truth, thus undercutting Werning's argument. Yet, recent discussions have made it clear that assessing the status of RCCP is "a *very* subtle matter requiring a careful investigation of both the principle itself and the evidence for/against it provided by our best scientific theories" (Hofer-Szabó et al. 2013, p.2; see also Wroński 2014).[29] So, here I will simply assume that RCCP is true. I will nevertheless argue that Werning's RCCP argument does not establish the necessity of causation in memory. The reason should be obvious to careful readers of the previous section. RCCP does license inferences from probabilistic correlation to the existence of a *type*-level causal relation. But it does not license inferences about *token* causal relations between particular events. More precisely: it only licenses inferences about the probability of such relations. Token events, which belong to event-types that are probabilistically correlated, need not be causally related.

[28] Responding to an earlier paper of mine (Andonovski 2018), Werning does talk of episodic memories having a prototype structure (2020, p.313n). Yet, importantly, he takes memories that do not fit this prototype to be ""deficient" cases of episodic memory (e.g., when no longer reliable as in certain patients with dementia)" (ibid). RCCP is, then, taken to guarantee that in *normal* cases, memory reliability requires a causal connection (see note 24). I will return to the issue of normality and "deficiency" in section 3.

[29] After all, RCCP puts forward a purely *existential* metaphysical claim, and philosophical tradition has taught us to be skeptical of alleged falsifications of such claims.

Hitchcock & Redei (2021) use the theatre troupe example to illustrate this point for common causes. Suppose the probabilities of the events are as follows:

$$p(F) = .1$$
$$p(L|F) = p(M|F) = .8$$
$$p(L|\neg F) = p\ (M|\neg F) = .1$$

where $p(L)$ is the probability of the leading lady getting sick, $p(M)$ is the probability of the leading man getting sick, $p(F)$ is the probability that they ate tainted food, and $p(\neg F)$ is the probability that they didn't. We can then calculate:[30]

$$p(L \cap M|F) = p(L|F)p(M|F) = .64$$
$$p(L \cap \neg M|\ F) = p(L|\neg F)p(M|\neg F) = .01$$
$$p(L) = p(L|F)p(F) + p(L|\neg F)p(\neg F) = .17 = p(M)$$
$$p(L \cap M) = p(L \cap M|F)p(F) + p(L \cap M|\neg F)p(\neg F) = .073$$

Hence, L and M are probabilistically correlated:

$$.073 = p(L \cap M) > p(L)p(M) = .17^2 = .0289.$$

Suppose that the causal relation that explains the correlation is indeed the one to the common cause F. The two actors tend to get sick at the same time because they eat at the same restaurants and occasionally share tainted food. If, on a *particular* day, both actors get sick, can we infer that they have eaten tainted food? We can calculate:

$$p(F|L \cap M) = .064 / .073 \cong .877.$$

While it is very probable that they did eat tainted food, there is a not an insignificant chance that they did *not*. This is despite the fact that F is a Reichenbachian common cause of L and M. In about

[30] The first two calculations show that M and L are conditionally independent, given F as well as given $\neg F$. Since the assignment of probabilities shows that M and L are more probable, given F (than given $\neg F$), F satisfies the four conditions for being a Reichenbachian common cause (see note 21).

12% of the cases, token events that belong to the event types *L* and *M* respectively, are *not* causally related. As Reichenbach (1971) himself points out, the "existence of a common cause is…in such cases is not absolutely certain, but only probable" (pp.157-158).

Let's turn back to memory. For reliable memory systems, we saw, the probability of a particular representation being accurate is high ($p(E'|R\ulcorner E\urcorner) \geq \rho_{S,E}$). This is explained by the existence of some causal mechanism – a process that connects seeming memories and past experiences. Let's call some internal event that belongs to this process – event *T.* This can be the event of forming a content-bearing memory trace, per the traditional causal theories of memory. But it can also be the event of transmitting sparse information – Werning's "minimal" trace – or even an event not involved in information transmission at all (a "purely neural" event, perhaps). What matters is that *T* is that intermediary event in the process that causally connects *E'* and $R\ulcorner E\urcorner$. In a simple model of a causal process, for example, *T* is directly caused by *E'* and directly causes $R\ulcorner E\urcorner$; thus, *E'* causes $R\ulcorner E\urcorner$ via *T.* Figure 1 presents a Lewis-style "neuron diagram" of such a causal process. The circles ("neurons") represent events that may (not) occur, with shading a circle (a neuron "firing") indicating that the event *has* occurred. The arrows represent stimulatory causal connections ("synapses"). The temporal order of the events is represented by reading left to right. In Figure 1, *E'* caused *T,* which then caused $R\ulcorner E\urcorner$.

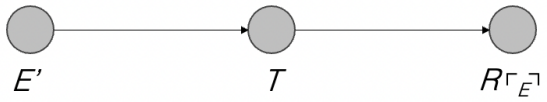

Figure 1. A causal process connecting *E', T* and $R\ulcorner E\urcorner$.

Suppose, then, that the probability threshold for reliability is set pretty high (e.g. $\rho_{S,E} = .85$) and that the probabilities are set to the values in Table 2:

	T		¬ *T*		Totals
	E'	¬ *E'*	*E'*	¬ *E'*	
$R\ulcorner E\urcorner$	.22	.02	.03	.02	.29
¬ $R\ulcorner E\urcorner$	.1	.01	.05	.55	.71
Totals	.32	.03	.08	.57	1.00
$p(R\ulcorner E\urcorner) = .29$ $p(T) = .35$ $p(E') = .4$			$p(R\ulcorner E\urcorner \mid T) \cong .685$ $p(T\mid E') = .8$ $p(R\ulcorner E\urcorner \mid T \cap E') \cong .687$		

Table 2. Hypothetical probabilities for a model of a reliable memory system.

We can calculate the probability of a memory representation being accurate:

$$p(E'|R\ulcorner E\urcorner) = p(R\ulcorner E\urcorner|E')p(E') / p(R\ulcorner E\urcorner) = .625 \text{ x } .4 / .29 \cong .862$$

So, the memory system is reliable and E' and $R\ulcorner E\urcorner$ are positively correlated:

$$.862 \cong p(E'|R\ulcorner E\urcorner) > \rho_{S,E} = .85 > p(E') = .4$$

Suppose that, on a *particular* occasion, both E' and $R\ulcorner E\urcorner$ have occurred. We can ask how likely it is that the intermediary event T, connecting E' and $R\ulcorner E\urcorner$, has occurred. From the probabilities above, we can calculate:

$$p(T|E' \cap R\ulcorner E\urcorner) = .22 / .25 = .88$$

Even when memory is highly reliable, then, there is a reasonable chance that the intermediary event T – i.e. the event that connects E' and $R\ulcorner E\urcorner$ in the model of the memory process – has *not* occurred. As above, in 12% of the cases, token events that belong to the event types E' and $R\ulcorner E\urcorner$ respectively, are *not* causally related. Not all memories produced by a reliable memory system are causally linked to the events they represent.

At this point, the causalist may be tempted to fiddle with the probabilities and the reliability threshold. Such amendments may indeed help but they cannot get us all the way. Given the *foundational* assumption that memory is imperfectly reliable, we will always end up with a (sizeable) class of memories not causally linked to the events they represent.[31] To be very clear, the point here is *not* that such a class of memories exists. It is rather that we cannot infer – *solely* from the probabilistic correlation between E' and $R\ulcorner E\urcorner$ – that it doesn't. Probabilistic dependence is not sufficient for token causation. The system architecture proposed by Menzies (1996), and depicted in Figure 2, illustrates this nicely. Suppose that the process connecting A to E is *very* reliable, in the following sense: if A were to fire, it would be very probable that the intermediary

[31] If memory were *perfectly* reliable, of course, this whole discussion would be moot. Alas, it is not.

neurons *B and C* fire. Suppose then that the process connecting *D* and *E* is much less reliable: if *D* were to fire, it would be improbable that *F* and *G* fire. Finally, suppose that there is a moderately reliable inhibitory connection between *D* and *B* (depicted by the line ending with a black dot): if *D* were to fire, it would be moderately probable that *B* would *not* fire. A neuron that is both stimulated and inhibited does not fire.

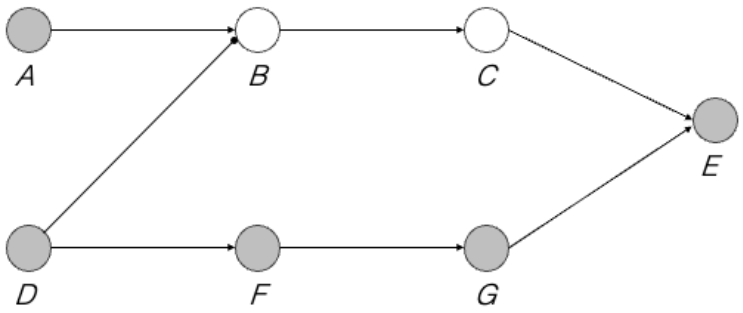

Figure 2. The causal process $D \rightarrow F \rightarrow G \rightarrow E$ "making it the hard way" in a probabilistic system. From Menzies (1996).

When the system functions normally, the occurrence of *A* typically entails that *B, C* and *E* will also occur. Sometimes, however, the improbable happens. In the case depicted in Figure 2, the unreliable process connecting *D* to *E* goes to completion, "making it the hard way".[32] At the same time, *D's* firing inhibits *B,* so that the reliable process from *A* to *E* doesn't go through, despite *A* firing. As Menzies (p.89) points out, this example shows that there can be token causation *without* probabilistic correlation. Thus, *D*'s firing causes *E* to fire even though it doesn't increase the chance of *E* doing so. In fact, it lowers it, since it increases the chance that the reliable process is inhibited. Even more importantly, the example shows that there can be probabilistic correlation without token causation. *A*'s firing does *not* cause *E* to fire, despite increasing the chance of it.[33] The example thus illustrates that probabilistic correlation is neither necessary nor sufficient for token causation.

The bottom-line is that we cannot infer the necessity of causation from the probabilistic correlation between memories and the events they represent. Even if memory is highly reliable, it

[32] The use of "making it the hard way" in this context apparently originated with Wesley Salmon, even though the expression doesn't appear in his writings (see Hitchcock 1995, p. 287, note 14).
[33] As Menzies points out, this is the case even if we take into account the moderate probability that *B*'s firing is inhibited. If *A* had not fired, the chance of *E* firing would have been lower; i.e. the unreliable process from *D* would be the only way *E* could fire.

doesn't follow that "there is no episodic memory without a causal link to experience" (Werning 2020, p.325). Establishing the necessity of causation – in the (meta)physical sense – requires showing that our memory systems do *not* function like the system in Menzies' example. And, whether this is the case is an empirical question that cannot be settled by only consulting statistical data. This is, of course, another way of saying that CAUSAL RELIABILITY does not entail NECESSITY. Werning's RCCP argument establishes the truth of the former but not the latter thesis.[34] Thus, it can only constitute an argument against *strong* post-causalist theories of memory. As we saw in the last section, it is not clear whether there *are* any such theories on the market, with even the simulationist likely to endorse CAUSAL RELIABILITY. But this exegetical issue is not terribly important for our purposes. What matters more is that the two theses are independent. So, while simulationism may not be true (on the strong version), it is not – in any theoretically interesting sense – an "unstable position". The reliability of memory does *not* require a causal connection between a remembered event and a memory of it.

3. Memory Systems, Probability and Necessity

The take-away lesson from section 2 is that the (meta)physical necessity of causation cannot be established without looking at how memory systems *actually* work. If, after reading the section, you have become reasonably convinced of this, then my job here is more-or-less done. Nevertheless, you may still harbor a suspicion that there has been much ado about nothing. In normal circumstances, you may think, memories *are* appropriately causally linked to the events they represent. (After all, the reliability of memory is explained by positing such a causal connection.) All that has been shown is that *ab*normal instances of remembering – "deficient" cases of episodic memory (cf. Werning 2020, p.313n) – sometimes occur. The short answer, of course, is that we can't know *a priori* whether these cases *are* abnormal, however intuitive that conclusion may seem. In this final section, I will nevertheless attempt a slightly longer answer. I will present some general considerations to pump the intuition that, at least in *some* normal cases, our memories may not be (appropriately) causally linked to past experiences. These are familiar considerations – see Schacter 2012; De Brigard 2014; Michaelian 2013, 2016a; Andonovski 2020,

[34] Note, by the way, that I have completely bracketed the issue of *appropriateness* in this section. The inferred probability of appropriate causal links in reliable memory systems will arguably be significantly lower compared to our estimates.

2021b – so I will not be breaking any ground. I will, however, dress them in the language of probability to establish continuity with the themes of this paper and to solidify the common ground for future discussion.

Let's start with a simple case of causal "preemption" (cf. Lewis 1986).[35] Suppose that you have observed some event *E'* and have thus encoded and stored a (content-bearing) memory trace of it – event *S.* At a later point, the memory trace is activated (*A*) which causes you to *accurately* represent the observed event ($R\ulcorner_E\urcorner$). But you were not alone when you observed the event; your partner was there, too. They also stored a trace of the event (S_P) and later represented the event successfully ($R_P\ulcorner_E\urcorner$) upon the activation of it. In fact, they did so slightly before you did. (The episodes were likely occasioned by the presence of some external cue, which I am omitting for convenience.) Both of you, then, represented *E'* on your own, using your "internal" resources – as uncontroversial cases of remembering as they come. In your case, *E'* caused $R\ulcorner_E\urcorner$ via *S* and *A.* As it happens, however, if something had gone wrong along that causal route – e.g. if you had not been able to activate your memory trace ($\neg A$) – your partner would have told you about the event *(T)* and you would have represented it accurately anyway. It was thus only the occurrence of *A* that "inhibited" *T*. This scenario is depicted in Figure 3.

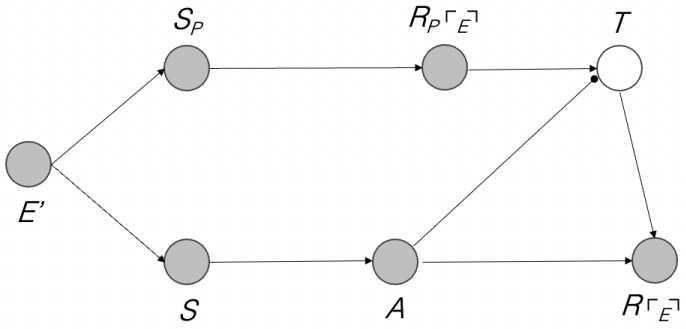

Figure 3. Causal preemption in memory.

Preemption cases have generated philosophical interest primarily because they illustrate that (token) causation and dependence come apart (Lewis 1986; Paul & Hall 2013, Ch.1). In the example, $R\ulcorner_E\urcorner$ is caused by *A* but it is *not* counterfactually dependent on it. Had *A* not occurred, the "backup" process would have gone through, with *T* causing $R\ulcorner_E\urcorner$. *Ex hypothesi,* the scenario depicted in Figure 3 is deterministic. But as we saw, the point extends to probabilistic scenarios.

[35] You can find cases of *remembering* preemption in, e.g., Martin & Deutscher (1966, pp.178-179) and Bernecker (2008, p.53).

We can amend our example slightly, so that the "backup" process is more reliable than the primary one – that is: $p(R\ulcorner E\urcorner \mid S_P) > p(R\ulcorner E\urcorner \mid S)$. (There can be plenty of reasons for this: you have a poor memory, your partner is more attentive than you are etc.) We then get a familiar result: *S* and *A* cause $R\ulcorner E\urcorner$ despite *lowering* the probability of $R\ulcorner E\urcorner$'s occurrence.

This last case is of interest to us here, however, because it illustrates another point. There is an, arguably sizeable, class of scenarios in which "external" processes do more reliably lead to accurate representations of past events. (Think of distant memories, benevolent partners and friends, diaries, photographs, video cameras etc.) If the goal of memory is to *maximize* reliability, then relying on such processes – at least in the relevant class of scenarios – is likely a good idea. Incorporating testimonial information is the most natural way of doing this. The evidence that we, routinely and systematically, incorporate testimonial information in our memories is now overwhelming (for reviews, see Loftus 2005; Wylie et al. 2014). For obvious reasons, experimenters have primarily investigated the incorporation of *mis*information – i.e. misleading or erroneous information – presented to subjects after they have experienced some event. Hence the name "the misinformation effect" (Loftus 2005). But as Michaelian (2013) has convincingly argued, the effect should occur relatively rarely in natural settings. Given the *typical* competence and benevolence of (trusted) testifiers – Michaelian speaks of an "honesty bias" – the incorporation of testimonial information will often lead to the formation of accurate representations. In some circumstances, indeed, it will lead to the formation of accurate representations that the subject would not have otherwise formed – a *positive* information effect (pp.2448-2552).[36]

A number of theoretical accounts of the (mis)information effect have been offered throughout the years (see Ayers & Reder 1998). To paper over some of the nuances, a common assumption is that the memory trace and the testimony – in some to-be-determined way – *jointly* cause the resultant representation. Figure 4 illustrates one way of understanding such joint causation. As before, the "internal" process goes through the storage (*S*) and activation (*A*) of a memory trace. But testimony (*T*) also plays a role in causing $R\ulcorner E\urcorner$. (For convenience, I've not depicted the steps from *E'* to *T* that featured in Figure 3.) Information from the memory trace and

[36] Michaelian (2013) brings in a variety of considerations in defending this claim. Two are worth mentioning here. First, there is evidence that we tend to allow testimony to influence our memories when we believe that testifiers are likely to be more *accurate/competent* than us. Second, there is also evidence that testifiers do tend to be *honest.* The possibility of a positive information effect depends on both the competence and honesty/benevolence of testifiers. For details, see pp.2446-2552.

the testimony are incorporated in a new, combined representation (*C*), which then is activated in the (seeming) remembering.[37]

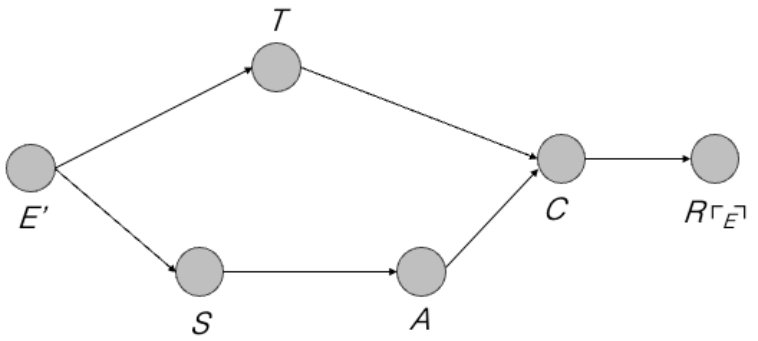

Figure 4. Testimonial incorporation as joint causation in memory

On this way of understanding the phenomenon, the key question is what the "rules" for the firing of *C* are. The traditional, causalist proposal is that *A* is necessary, even if not sufficient, for $R\ulcorner_E\urcorner$. Accordingly, *C* only fires if *A* fires but *C* may (not) require *T* firing.[38] In any case, $p(A \mid C \cup R\ulcorner_E\urcorner) = 1$.

The discussion above hints at why this proposal may turn out false. As Werning (2020, pp. 315-316) correctly points out, given the number of events we can remember, storing and preserving all "experiential" information independently would be information overkill.[39] Moreover, there are good reasons to think that the "loss" of such information is beneficial in the long run. (You can find *my* reasons for thinking this in Andonovski 2020 and 2021b.) Consider then situations in which *A* does *not* occur but there is a *very* reliable "backup" process that preserves the relevant information. For example, an event happened to you a long time ago, and *all* your experiential information is lost, but your friends have repeatedly told you about the event. Are there principled reasons for thinking there are *no* normal cases of (seeming) remembering in such situations? On the empirical front, there is nothing conclusive. If anything, some of the results – like those from the famous "lost in the mall" studies (Loftus & Pickrell 1995; Hyman et al. 1995) – point in the post-causalist direction. After all, if misinformation can lead to false memories of entire episodes, correct testimonial information can seemingly lead to true memories. Moreover,

[37] You don't have to accept this account to appreciate the general point. As far as I can see, as long as *some* kind of joint causation is at work, the relevant questions below can still be posed.
[38] When there is no relevant testimonial information to incorporate, for example, the trace content is simply preserved in *C*.
[39] "Experiential" information is the information that is acquired via first-hand experience and stored in a trace (*S*).

Werning's probabilistic considerations are moot here. If *T* and *C* are probabilistically correlated *and T* causes *C* in an actual scenario, then everything is in order from the perspective of the Common Cause Principle. Hence, there are no decisive *explanatory* reasons to accept NECESSITY. It is possible that at least in some normal cases, memories are not causally linked to past experiences via "discriminatory" memory traces. That is: $p(A \mid C \cup R\ulcorner E\urcorner) < 1$. Figure 5 depicts such a case. Due to the loss of experiential information, *A* does not occur. Nevertheless, the reliable "backup" process goes to completion.[40]

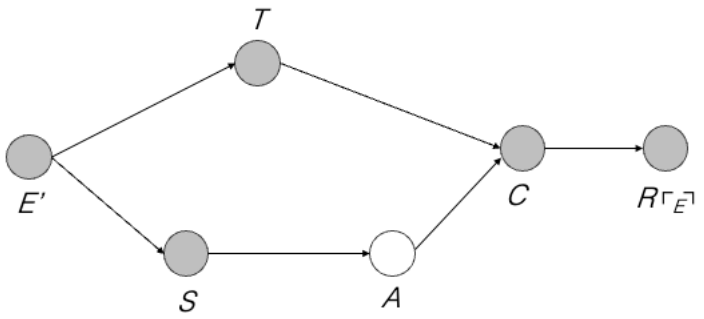

Figure 5. "Non-appropriate", testimonial causation in memory

What about memories not causally linked to past events *at all*? In the last section, I argued that these are not ruled out by Werning's application of the Common Cause Principle. But what then is their status and how should we think of them in relation to RCCP? (Are they just occasional flukes?) To appreciate the complexity of this question, consider a case in which you've had many different experiences of qualitatively similar events. For example, for over a year, you've gone to the same coffee shop every day, at approximately the same time of day, you've ordered the same thing, you've been served by the same barista etc. As a frequent visitor, you have slowly but steadily "picked up" many important details and regularities. You have stored information about unique events but have also formed models and schemas of varying generality – a cognitive map of the coffeeshop (Behrens et al. 2018), a personality model of the barista (Hassabis et al. 2014), a "procedure for buying coffee" schema (Ghosh & Gilboa 2014) etc.[41] When you attempt to recollect your experiences, the task and context dictate what information should be retrieved and how it should be put to use (see Andonovski 2021b). Even when you remember a singular episode,

[40] If your response to all this is "Yes, but this is not a memory", then you are probably operating with a different notion of necessity.
[41] I use "representation talk" here partly for convenience and party because I don't know *how* to talk about these issues in non-representational terms.

however – a paradigmatic case of "episodic" memory – the content of the memory will be dynamically (re)constructed from elements that may or may not originate in the specific experience (De Brigard 2014; Michaelian 2016a). (E.g. we typically rely on knowledge of the target location abstracted from multiple experiences.)

Suppose then that, during one of your visits (let's call it V_1), you were preoccupied with some problem at work. (Perhaps you were thinking about the RCCP the whole morning.) As a result, you were not able to pay sufficient attention to your surroundings and thus to encode a (detailed) trace of your dealings in the coffeeshop. In normal circumstances, can you then rely solely on your accumulated knowledge about the visits to accurately remember elements from V_1? If you aim to remember an element *unique* to V_1, then the answer is probably no. To be more precise: it may be (meta)physically possible to get things right, but it *would* be just a fluke. (Memory doesn't normally work like that.) Yet, if you aim to remember more common elements of V_1, – and the accuracy threshold is not set unreasonably high – then the chance of producing an accurate representation is much higher. This is nicely illustrated by Steyvers & Hemmer (2012), who studied the contribution of prior knowledge to the recall of particular scenes. In a clever setup, subjects were asked to name the objects they would expect to find in a scene of a given type (e.g. an office). These expectations were then used as "reasonable guesses" about the objects that might be present in a *particular* scene of the type (to which the subjects had not been exposed). Remarkably, the subjects' performance in this condition was pretty high. For the first guessed item, the subjects were 85% accurate, "even though the response [was] not based on any episodic information of the presented scenes" (p.138). (For comparison, subjects who were presented with the scene for 10 seconds were 90% accurate.) For multiple items, the accuracy decreased but even for as many as 16 items, the cumulative accuracy was still higher than 55%. The result demonstrates that prior knowledge "can *greatly* contribute to the accuracy of recalling… natural scenes" (p.139, emphasis added).

It also allows us to situate our discussion in the context of CAUSAL RELIABILITY. When I introduced the thesis, I motivated it with a lottery analogy. If I win the lottery many times, it is very likely that something is fishy – i.e. that there is some causal/information link that explains my success. Werning's use of the RCCP relies on a similar intuition: if our memory system keeps on producing accurate representations of past events, it is very likely that there is some underlying causal mechanism that explains *its* success. While this intuition is generally correct, we can now

see that the lottery analogy is not entirely apt. What needs to be "guessed" in memory is very rarely a unique set of elements, drawn at random. It is rather a set whose elements often appear together and are typically organized in a similar way. In Werning's idiom: a target event (e.g. V_1) is often qualitatively and spatially (the "Q" and "S" in the "STQ" metric), even if not temporally ("T"), similar to a number of other events (V_2, V_3, V_4 etc.). As the rememberer acquires more knowledge about the relevant class of events, their ability to "reconstruct" details from a particular event steadily improves (cf. De Brigard 2014). And, while information from the target event will presumably often be used in such reconstruction – perhaps as an "error signal" for memory's predictive machinery (Werning 2020, pp.326-328) – there is no *a priori* reason to think that such information is *necessary.* (RCCP certainly doesn't require this.) We should thus expect, at least some, normal cases in which an (accurate) memory of an event is not causally linked to it. Figure 6 uses the coffeeshop scenario to depict a case of this kind. The subject fails to draw on information for their experience of the target event V_1 – hence, A does *not* fire. Yet, they manage to rely on their prior knowledge (P_V) – accumulated from qualitatively similar events V_2, V_3 and V_4 – to construct a sufficiently accurate representation (C). As a result, they can successfully remember the target event $(R\ulcorner V_1 \urcorner)$. In this case, $p(A \mid R\ulcorner V_1 \urcorner) < 1$.

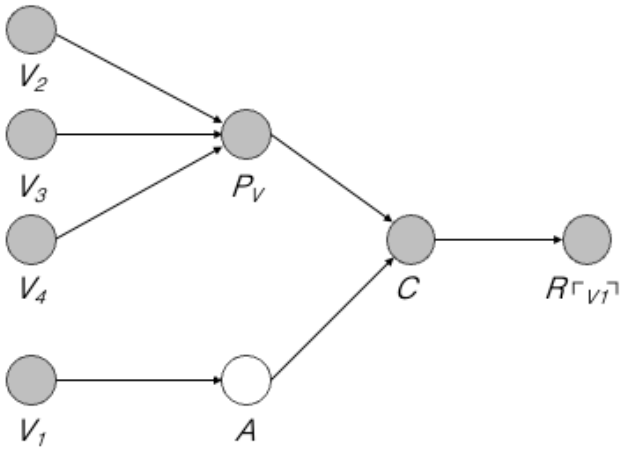

Figure 6. Remembering without a causal link to a past event

I'll end this section with a point from further leftfield. So far, I have assumed that the long-term "goal" of memory is to maximize reliability in the representation of particular past events. But this may not be a safe assumption. Indeed, much of the recent discontent with traditional causalism is due to the rejection of this assumption (see Schacter et al. 2012; Michaelian 2016a,

Ch.6). In a recent essay (Andonovski 2021b), I have expressed *my* discontent, arguing that memory is a faculty performing a kind of "cognitive triage": management of information for a *variety* of uses under significant computational constraints.[42] One consequence of this view are the potential tradeoffs between accuracy and relevance. For illustration, let's slightly modify our example. Imagine you have recently moved to a new city. It's an exciting place with a vibrant coffeeshop scene. You've decided to try a few of the more peculiar shops, and it turns out they boast roasts from across the world and attract interesting crowds. Visiting, you try to "take in" as much as possible, yet you are interested in many things: the quality of the coffee but also the atmosphere in the neighborhood, the eccentric visitors, the paraphernalia on the walls and so on. (At this stage, all of this can be relevant.) You want to remember these first experiences but you also have other goals, not the least of which is to learn more about the neighborhood and the kind(s) of lifestyle it affords. Importantly, your goals *modulate* what information is encoded and how it is stored and consolidated in your memory for future use (see 2021b, pp. 237-247). Later, when you call your old friends to report, you may inform them of a peculiar incident but you'll also likely discuss your general impressions of the new city.

Importantly, the different goals and uses will often place *conflicting* demands on the selection and organization of information in memory. If you are really interested in the coffeeshop dwellers, you'll often miss many of the less riveting details (apron colors, counter placements and the like). More pertinently, if you are intent on putting your finger on the pulse of the neighborhood, you'll try to extract important lessons about it – combining and integrating your impressions so that many of the singular details (gradually) wash away. While this doesn't always happen – consolidation can facilitate both the preservation of exemplar-specific information *and* its integration in general representations (see, e.g., Landmann et al. 2014) – it presumably often does. Given limited resources and shifting relevance patterns, there may be significant tradeoffs – most notably between accuracy in the representation of individual episodes and usefulness for learning (about the past). The tradeoffs, in turn, may lead to the formation of more inaccurate memories, and false memory beliefs, than we would perhaps expect.[43] Indeed, we arguably see an effect of this kind in some classic experimental paradigms – e.g. the DRM memory task (Roediger

[42] "Memory" here refers only to *declarative* memory. On the standard accounts, declarative memory is a kind of long-terms memory that involves the encoding, storage and retrieval of information. *Very* roughly, it supports the remembering of events ("episodic memory") and the remembering of facts ("semantic memory").

[43] Nevertheless, such beliefs are "epistemically innocent", delivering significant epistemic benefits that could not be (easily) attained otherwise (Puddifoot & Bortolotti 2019).

& McDermott 1995). Trying to remember theme-related words they've heard, participants generate a number of false positives – "recalling" semantic lures - *precisely* because they have successfully extracted the relevant commonalities between them. More boldly: they have unearthed the "generative principle" of the list.

Yet, in *naturalistic* settings, accurate memory representations will still be common. For example, in most "normal" environments, relying on information about the semantic associations between individual words will be helpful in reconstructing details that are not retained. The reason for this is simple: in everyday conversations, words like "sweet" *do* often co-occur with words like "sugar" and "candy". Making such associations "will often lead you to successfully predict features of a new environment and adopt accurate beliefs about items likely to be found in that environment" (Puddifoot & Bortolotti 2019, pp.759-760). This principle generalizes, with schematic/associative knowledge typically facilitating navigation in complex, dynamic environments. Sometimes, of course, relying on such knowledge can lead one astray. Yet, it *typically* won't. If this account is on the right track, the organization of information in memory requires a delicate balance between the maximization of accuracy and the maintenance of long-term relevance. Given such balance, relying on *general* information when representing particular events will – in some cases, at least – be the best we can get. But it may still be good enough.

In Figure 7, I illustrate the coffeeshop scenario again, factoring in the main insight of the memory-as-triage account. As before, the subject draws on their prior knowledge *(P_V)* – accumulated from events V_2, V_3 and V_4 – to construct a sufficiently accurate representation *(C)* of the qualitatively similar event V_1. Hence, they can successfully remember the target event *($R\ulcorner V_1 \urcorner$)*. The prior knowledge can also be used for another purpose, however – e.g. to describe the neighborhood (*D*). In fact, the goal of succeeding in activities like *D* may drive much of the schematization in the accumulation of prior knowledge (e.g. the "washing out" of experiential information, leading to $\neg A$). As a result, the probability of success in *D given P* may be higher than probability of success in $R\ulcorner V_1 \urcorner$ given *P*. Nevertheless, in cases like the one depicted in the Figure, the subject still manages to accurately remember V_1.

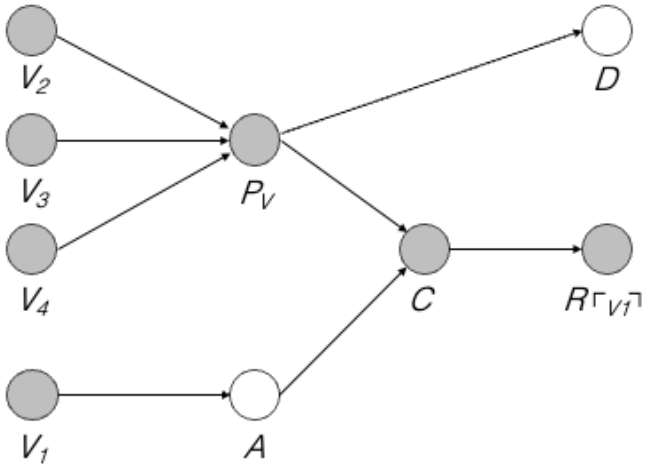

Figure 7. Remembering particular events is one of the "uses" of memory

I hope I have managed, in this brief discussion, to at least gesture at why post-causalists are *not* hung up on flukes. For all we know, some normal "non-deficient" memories are not causally linked to the events they represent.

Conclusion

I have argued that causal theories of memory are committed to two independent theses. The first thesis concerns the necessity of appropriate causation and specifies a condition *token* memories must satisfy. The second concerns the explanation of memory reliability in causal terms and identifies a *type*-level relation between memories and past events. There is a familiar, well-established way of accommodating this result. Scientific psychology, you may insist, offers only type-level causal generalizations. So, it is hardly surprising that explanatory concerns will not point to the necessity of appropriate causation. The notion of necessity simply doesn't play a role in the sciences of the mind.

As the paper hopefully makes clear, this is only part of the story. In reality, we are not only confounded by the challenge of reconciling conceptual, metaphysical and explanatory insights about memory. We are also mostly in the dark about how memory actually works. When it comes to understanding even the *typical* etiological profiles of memories, our predicament is very much like the one described by Ovid: the effects are visible to all but the causes are hidden.

Elucidating them will require a lot of work. And, while *a priori* considerations – probabilistic or otherwise – will provide some important constraints, it is the empirical psychologist that will end up doing the brunt of the work.[44]

[44] Thanks to André Sant'Anna for the invitation to contribute to this special issue. Thanks also to my comrades at the Centre for Philosophy of Memory and to audiences at Issues of Philosophy of Memory 2.5 and the Sofia-Grenoble Workshop on memory.

Bibliography:

Andonovski, N. (2018). Is Episodic Memory a Natural Kind?. *Essays in Philosophy,* 19(2), 178-195

Andonovski, N. (2020). Singularism about Episodic Memory. *Review of Philosophy and Psychology,* 11(2): 335-365

Andonovski, N. (2021a). Causation and mnemonic roles: on Fernández's Functionalism. *Estudios de Filosofía,* (64), 139-153

Andonovski, N. (2021b). Memory as Triage: Facing Up to the Hard Question of Memory. *Review of Philosophy and Psychology,* 12(2), 227-256

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review,* 5(1), 1-21

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron,* 100(2), 490-509

Bernecker, S. (2008). *The metaphysics of memory* (Vol. 111). Springer Science & Business Media.

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

Bernecker, S. (2017). A causal theory of mnemonic confabulation. *Frontiers in psychology*, 8, 1207

Cartwright, N. (1988). How to Tell a Common Cause: Generalizations of the Conjunctive Fork Criterion. In *Probability and Causality*, J. H. Fetzer (ed.), Dordrecht: Springer Netherlands, 181–188.

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

Fernández, J. (2019). *Memory: A self-referential account.* Oxford University Press, USA

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia,* 53, 104-114

Goldman, A. I. (1986). Epistemology and cognition. *Harvard University Press*

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cerebral Cortex,* 24(8), 1979-1987

Hitchcock, C. (1995). The mishap at Reichenbach fall: Singular vs. general causation. *Philosophical Studies,* 78(3), 257-291.

Hitchcock, C. (1998). The common cause principle in historical linguistics. *Philosophy of Science,* 65(3), 425-447

Hitchcock, C. & Rédei, M. (2021). Reichenbach's Common Cause Principle. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/physics-Rpcc.

Hofer-Szabó, G., Rédei, N. & Szabó, L.E. (2013). *The Principle of the Common Cause.* Cambridge University Press.

Hyman, I. E., & Pentland, J. (1996). The role of mental imagery in the creation of false childhood memories. *Journal of Memory and Language,* 35, 101–117

Landmann, N., M. Kuhn, H. Piosczyk, B. Feige, C. Baglioni, K. Spiegelhalder, L. Frase, D. Riemann, A. Sterr, and C. Nissen. (2014). The reorganisation of memory during sleep. *Sleep Medicine Reviews* 18 (6): 531–541

Lewis, D. (1986). Causation. Reprinted with postscripts in his *Philosophical Papers, Vol. II.* Oxford University Press, pp. 159-213.

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

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

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

Menzies, P. (1996). Probabilistic causation and the pre-emption problem. *Mind,* 105(417), 85-117

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

Michaelian, K. (2013). The information effect: Constructive memory, testimony, and epistemic luck. *Synthese,* 190(12), 2429-2456

Michaelian, K. (2016a). *Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past.* The MIT Press

Michaelian, K. (2016b). Against discontinuism: Mental time travel and our knowledge of past and future events. In Michaelian, K. et al (eds.) *Seeing the Future: Theoretical Perspectives on Future-Oriented Mental Time Travel.* Oxford University Press

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

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

Paul, L. A. & Hall, N. (2013). *Causation: A user's guide.* Oxford University Press

Perrin, D. (2018). A case for procedural causality in episodic recollection. In Michaelian, K., Debus, D. & Perrin, D. (Eds.) *New Directions in the Philosophy of Memory*. Routledge

Perrin, D. (2021). Embodied Episodic Memory: A New Case for Causalism? *Intellectica* 1, 74: 229-252

Priest, G. (2021). Metaphysical necessity: a skeptical perspective. *Synthese,* 198(8), 1873-1885.

Puddifoot, K., & Bortolotti, L. (2019). Epistemic innocence and the production of false memory beliefs. *Philosophical Studies,* 176(3), 755-780

Reichenbach, H. (1971). *The Direction of Time.* University of California Press. Reprint of the 1956 edition.

Robins, S. K. (2017). Contiguity and the causal theory of memory. *Canadian Journal of Philosophy*, 47(1), 1-19

Robins, S.K. (2020). Defending discontinuism, naturally. *Review of Philosophy and Psychology,* 11(2), 469-486

Robins, S. K. (2021). The failures of functionalism (for memory). *Estudios de Filosofía,* (64), 201-222

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition,* 21(4), 803

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. C. (1998). *Causality and explanation.* Oxford University Press

Sant'Anna, A. (2018). Episodic memory as a propositional attitude: A critical perspective. *Frontiers in Psychology,* 9, 1220.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, 76(4), 677–694.

Schurz, G. (2017). Interactive Causes: Revising the Markov Condition. *Philosophy of Science* 84(3): 456–479

Sober, E. (1988). The Principle of the Common Cause, in *Probability and Causality: Essays in Honor of Wesley C. Salmon*, James H. Fetzer (ed.), Dordrecht: Springer Netherlands, 211–228

Von Leyden W. (1961). *Remembering: A Philosophical Problem.* New York: Philosophical Library, Inc

Werning, M. (2020). Predicting the past from minimal traces: episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology* 11: 301–333

Williamson, T. (2016). Knowing by imagining. In Kind, A. & Kung, P. (eds.) *Knowledge Through Imagination,* 113-23. Oxford University Press

Wroński, L. (2014). R*eichenbach's Paradise: Constructing the Realm of Probabilistic Common "Causes".* Berlin: De Gruyter.

Wylie, L. E., Patihis, L., & McCuller, L. L. (2014). Misinformation effect in older versus younger adults: A meta-analysis and review. In Toglia, M.P. et al. (eds.) *The elderly eyewitness in court* (pp.52-80). Psychology Press

Zemach, E. M. (1983). Memory: What it is, and what it cannot possibly be. *Philosophy and Phenomenological Research*, 44(1), 31-44