Illusions of memory:
What referential confabulation can tell us about remembering

Recent philosophy of memory tends to treat confabulation as a distinctive type of representational error, marked by reference failure, often via direct analogy with the traditional conception of sensory hallucination. I argue that this model misrepresents the phenomenon. Drawing on the empirical possibility of referential confabulation—wherein confabulators mnemically refer to events in their past—I argue that mnemic reference and genuine remembering come apart. This, in particular, challenges causalist theories for which one element—appropriate causation—purports to secure reference and to separate genuine remembering from confabulation. Acknowledging referential confabulation requires causalists to complicate their story in a way that has implications on what remembering is. More generally, referential confabulation prompts a broader rethinking of memory error debates. Rather than being a distinctive type of content-level error, confabulation is better characterised as a processing malfunction: a breakdown in strategic retrieval and monitoring, but not necessarily in referential success. Appreciating this calls us to aim at a more nuanced conception of remembering and its frailties.

1 Introduction

In recent philosophy of memory, there is thought to be a peculiar type of error known as *(mnemic) confabulation*.[1] It is standardly contrasted with misremembering in a manner analogous to the traditional illusion-hallucination distinction in the philosophy of perception. Very simply, in perceptual illusion, there is an object one perceives but one misperceives it *as* having a property *F* (a property it does not in fact possess). In hallucination, one fails to perceive an object at all. Instead, one has an experience as of an object with certain properties, but there is no such object (or instantiated properties) to which one's experience consists in being perceptually related.[2] Accordingly, the standard view is that to misremember is for there to be an event one remembers but to which one misattributes certain features. In confabulation, by contrast, there is no event to which one is mnemically related in the relevant sense at all. In short, confabulations are "mnemonic hallucinations" (Robins 2019: 2149); they are errors because "there is no relation between

[1] I use the modifier 'mnemic' to mean memory-related (following Michaelian (2020)). Others quoted here use 'mnemonic', though this has the disadvantage of connoting tools or strategies for *aiding* memory.

[2] For discussion of the adequacy of the traditional distinction, see Macpherson & Batty (2016). It is enough for my purposes that this is the traditional conception on the basis of which the analogy has been made.

a person's seeming to remember a particular event or experience and any event or experience from their past" (Robins 2020: 125–6).[3]

In this paper, I first argue that there is an extensive class of mnemic confabulations suggested by the empirical literature that have been underappreciated for their philosophical implications. This is crucially not the traditional and unconvincing set of cases some have called 'veridical confabulations', which should fail to persuade us. Instead, mnemic confabulations often successfully *refer* to events in the subject's personal past. To what extent they are then accurate is less important than the presence of referential success: such confabulations are not (fully) illuminated by an analogy with hallucination. The lesson is that there is no general type of representational error that is—or that coincides with—mnemic confabulation.

Overlooking referential confabulation is not only a product of over-analogising, however. Causalist theories, as they have been stated, do not predict the existence of this class of confabulation. Indeed, their existence raises a *prima facie* challenge for causalists. Causalists rely on a single notion of appropriate causation to do two things: (i) explain how memory representations refer to past events; (ii) distinguish genuine remembering from confabulation. Specifically, they appeal to the *presence* of appropriate causation via a memory trace to explain mnemic reference and to its *absence* to explain mnemic confabulation. So what of referential mnemic confabulations? I argue that these two explanatory roles come apart. Causalists should abandon the idea that confabulation characteristically involves reference-failure. This has interesting and unacknowledged implications on the causalist's account of remembering and its critical contrast with confabulation. One dialectical upshot is that causalists can no longer claim theoretical advantage over non-causalist (e.g., simulationist) theories on grounds that they have a straightforward account of the distinction between remembering and confabulation.

A broader moral is that there is an instability in the use of the term 'confabulation' to both denote a distinctive type of content-level error (to be contrasted with misremembering as per the traditional illusion-hallucination paradigm) and, at the same time, to track the memory-related malfunction described in memory science. By revealing referential confabulation, achieving stability (in favour of the latter disambiguation) is instructive for thinking about what confabulation is and what it can tell us about memory.

[3] Comparison between confabulation and hallucination is endemic in recent philosophy of memory (Barkasi & Sant'Anna 2022; Bernecker 2017; Copenhaver 2024; Michaelian 2016b; 2020; Michaelian & Sutton 2017; Robins 2019; 2020; Sant'Anna & Michaelian 2019; Sant'Anna 2022; Werning & Liefke 2024).

§2 begins by introducing the concept of mnemic confabulation via its scientific origins (distinguishing it from confabulation in a broader sense). The concept is not folk-psychological, and it is unlikely that we can tap direct intuitions about its extension. §3 contrasts this with a philosophically popular conception of mnemic confabulation as being constituted by the content-level feature of reference-failure. §4 introduces the standard causalist picture, which—due to its implication that mnemic confabulation must coincide with reference-failure—partly explains the popularity of this conception. §5 exhibits the phenomenon of referential confabulation, which appears to be not only empirically possible and arguably instantiated in a number of cases, but a prediction of dominant empirical approaches to mnemic confabulation (namely, temporal context and strategic retrieval approaches). §6 considers several objections on behalf of the causalist, and §7 draws out the implications for what I take to be the most defensible line of response, before §8 concludes.

2 Mnemic confabulation

In the late 19th century and early 20th century, neurologists such as Kraepelin, Korsakoff, Bonhoeffer, and Wernicke began using the term 'confabulation' to describe a behaviour in patients exhibiting severe memory deficits, specifically in those suffering from what would come to be known as Wernicke-Korsakoff's syndrome:

> Only after a long conversation with the patient, one may note that at times he utterly confuses events and that he remembers absolutely nothing of what goes on around him. […] on occasion, such patients invent some fiction and constantly repeat it, so that a peculiar delirium develops, rooted in false recollection (pseudo-reminiscences) (Korsakoff (1891), in Victor & Yakovlev (1955: 397–9)).

The term is etymologically related to the Latin *fabulari*, i.e., 'fable', characterising the sense in which the patients' reports were both narrative-like in structure and ill-grounded (in a sense to be refined). Yet such claims to remember are literal, honest, and sincere in intent: the patient believes that they are genuinely remembering. Much of the time, as Talland later put it, the source of such confabulation was thought to be "the patient's actual experiences in an earlier phase of his life", the behaviour reflecting "disruption of his temporal frame of reference" rather than pure fabrication (1965: 56). Korsakoff (1891) sometimes called the

phenomenon 'pseudo-reminiscence', "the feeling of a reminiscence, […] rooted in the memory of some true events" (in Victor & Yakovlev (1955: 400f.)).

> Pseudo-reminiscences are very often, if not always, rooted in real memories and thus appear as *illusions of memory*. As a consequence, their character differs between the patients. Patients who have been passionate hunters will talk about their adventures, like hunting, fishing and the like (Korsakoff (1891), in Victor & Yakovlev (1955: 394); emphasis added).[4]

Over time, the word 'confabulation' came to be used for a much broader array of behavioural phenomena. By 1972, psychiatrist Neville Berlyne complained that the term, while 'widely employed', was 'poorly defined and variously interpreted' (1972: 31). In the intervening 50 years, its usage has only proliferated further. It is therefore helpful to distinguish a broad sense of the term that reflects this diversity of usage from the narrower and specifically *mnemic* sense of term in which I will be interested in this paper, one which finds its origins in the work of Korsakoff and contemporaries.[5]

In its more recent, *broad* sense, 'confabulation' may be characterised as follows:

Confabulation$_B$ The honest, confident, and literal production of an ill-grounded narrative, typically about oneself.

The narrative produced in confabulation$_B$ may consist in a verbal report, but it may also be 'silent', consisting in a judgment, or in an action with the narrative as its rational basis. 'Ill-grounded' here should be distinguished from *inaccurate*. It is neither strictly necessary nor sufficient that the report given by the subject be literally false. Rather, the report is defective in being somehow unreliably sourced or lacking a proper evidential basis.

In its broad sense, the term has been used by the likes of Bortolotti (2018) to describe everyday phenomena exhibited, for instance, in Nisbett & Wilson's (1977) famous study. Though the details are not important for our purposes, the study has been taken to suggest that we often lack conscious access to decision-making processes and 'confabulate' inaccurate (albeit more recognisable) rationalisations *post hoc*. The broad notion been characterised by Hirstein (2005) as an epistemic phenomenon that can involve

[4] It is with some foresight that Korsakoff (1891) speculated that the behaviour was to be explained by impaired associative processes operating over unconscious traces.

[5] In distinguishing the two notions, I do not mean to prescribe a strategy of explicit disambiguation in the sciences. The *conscious* use of polysemous terms can of course be productive and useful.

errors stemming from various sources not limited to memory (e.g., perception). Across these cases, it is far from clear that it has an interesting or deep connection to *memory*. A particularly good example (as Robins (2020) notes) is the application of the term to behaviours sometimes observed in Anton syndrome (e.g., Cao et al. (2020)). Patients with Anton syndrome are cortically blind but insist on being visually unimpaired. To cope with evidence to the contrary, such as when they manifestly bump into things, they may suggest, e.g., that the room is dark. The syndrome does not appear to require any distinctively *mnemic* deficit at all.

In general, confabulation$_B$ need not indicate a cognitive *malfunction* and, where it stems from malfunction, it is not the malfunction of a uniform, underlying mechanism or process. Rather, 'confabulation' in this usage refers to a behaviour with many distinct and non-overlapping causal explanations.

Confabulation in its *narrow* sense, by contrast, is closely related to the phenomena observed by the late 19th and early 20th Century neurologists mentioned earlier. Very roughly, we can think of confabulation in the narrower, mnemic sense as crucially involving the production of apparent memories as of events "which never actually happened, or which occurred in a different temporal-spatial context to that being referred to by the patient" (Dalla Barba 2002: 28), but which "are experienced as real memories" (191). Only somewhat less loosely, we might characterise confabulation in the narrow sense as follows:

Confabulation$_N$ The honest, confident, and literal production of an ill-grounded narrative, characteristically concerning one's past and with apparent epistemic first-handedness of source, the ill-groundedness of which is at least partly attributable to error of some distinctively *mnemic*, retrieval-related process.

Once again, the narrative produced need not be verbalised; it may consist in having an apparent memory without any downstream behavioural effects.[6]

Whether this loose description tracks something meaningful at a psychological level remains to be seen. No doubt the characterisation is loose and imperfect.[7] And, to be clear,

[6] Some distinctions, such as whether confabulation is 'spontaneous' or 'provoked', may matter for clinical psychology—the two may dissociate (Gilboa & Verfaellie 2010)—but are neglected for simplicity here.

[7] One might object that this downplays confabulations tied more closely to semantic memory than to episodic memory, or the possibility of future-oriented confabulation (Dalla Barba 2002: 197ff). 'Personal semantic' confabulations would remain characteristic, however. It is less clear that confabulation-like phenomena concerning purely impersonal, general world knowledge should be clustered with typical

this is no attempt at a conceptual analysis. But this characterisation will suffice for our purposes, saying enough to be contentful without saying so much as to fail in attempting to capture what much of the literature is plausibly committed to.[8]

A number of brief clarifications concerning confabulation$_N$. First, ill-groundedness should again be distinguished from literal inaccuracy. We will revisit this issue in subsequent sections. The basic idea is that confabulation is not characterised by inaccuracy *per se* but by errors in processing.[9] Second, by 'apparent epistemic first-handedness of source', I mean to specify that the form of memory I have in mind is autobiographical in character, perhaps in part grounded in the activity of an episodic system (Schacter & Tulving 1994). Confabulation's "area of reference is principally, even though not exclusively, the self" (Talland 1961: 365). Such representations come with the sense that the recollected content *comes from* some experience(s) in one's own personal past, rather than having been learned second-hand. On the way of developing this idea favoured by Mahr & Csibra (2018), such representations are meta-represented as having been obtained first-hand. Finally, 'mnemic retrieval-related process' is intended broadly. Dalla Barba's (2002) hypothesis that confabulation results from distortions of 'temporal consciousness' may also be included here.

To help make these descriptions more concrete, consider the following interview with patient H.W., who suffered bilateral infraction of the frontal lobes:

> Q. How long have you been married?
> A. About 4 months. […]
> Q. How many children do you have?
> A. Four. (He laughs.) Not bad for 4 months.
> Q. How old are your children?
> A. The eldest is 32, his name is Bob, and the youngest is 22, his name is Joe.
> Q. […] How did you get these children in 4 months?
> A. They're adopted. […]
> Q. Immediately after you got married you wanted to adopt these older

cases, but see Nahum et al. (2012). To the extent that they are characteristic cases, the 'strategic retrieval' approach favoured later in the paper can accommodate them.

[8] Though analogies to delusion have sometimes been made, disanalogies appear to run deeper. Delusion may be seen as a belief-formation disorder which, if it has a central mnemic component, involves biased encoding, not malfunction of retrieval (Gilboa 2010; Kopelman 2010).

[9] Robins (2020) suggests that mnemic confabulations are unlike confabulations in the broad sense because they are '*well*-grounded', i.e. justified in an internalist sense, since the subject blamelessly takes themselves to be remembering. This internalist conception of 'groundedness' is not what I have in mind.

children?
A. Before we were married we adopted one of them, two of them. The eldest girl Brenda and Bob, and Joe and Dina since we were married.
Q. Does it all sound a little strange to you, what you are saying?
A. (He laughs.) I think it is a little strange (Moscovitch 1989: 136).

Though H.W. certainly seems to confabulate in our narrow sense, this may also be combined with so-called secondary confabulation (e.g., 'They're adopted'), the *post hoc* rationalisation of an inappropriate response that *was* seemingly a memory error.

It is an open question to what extent it is fruitful to extend the use of the term 'confabulation' in this narrow sense beyond the clinical population. Consider the so-called Mandela effect. Is a neurotypical individual who seems to remember seeing news reports about Nelson Mandela's death during his imprisonment in the 1980s mnemically confabulating (Michaelian & Wall, forthcoming)? Are many of the subjects of the *lost in the mall* paradigm (Loftus & Pickrell 1995) mnemically confabulating when they seem to recall being lost in a mall as a child (Robins 2019: 2148–9)? My discussion will not hinge on this. Cases from the clinical population offer less contestable instances of the phenomenon. By focusing my attention to such cases, fewer assumptions need to be made.

A final, methodological remark. Bernecker (2023) notes that it is tempting to see some as using the term 'confabulation' stipulatively to refer to merely apparent recollections lacking the distinctive feature in virtue of which, they hypothesise, a representation counts as a genuine recollection. Unsurprisingly, this way of defining the term will lead to verbal disputes between parties who disagree on what the distinctive feature is. Insofar as Bernecker's interpretation is correct, the obvious antidote is to adopt a more explicitly (if not entirely) 'bottom-up', naturalistic approach, on which confabulation—in the narrow sense—is a theoretical concept inherited from memory science, relative to which the prospects for conceptual analysis via thought experiments or intuitions about 'cases' are dim. I will be adopting such an approach in this paper.[10]

3 Illusions of memory

In the rest of the paper, I will be exclusively concerned with confabulation in the narrow, *mnemic* sense (confabulation$_N$). Recent philosophy of memory has referred to

[10] As Bernecker puts it: "If the method of conceptual analysis works at all, it works only for terms that are entrenched in ordinary discourse. But this is not the case for 'confabulation'" (2023: 115).

confabulation in the narrow sense as (i) a symptom; (ii) a disorder; (iii) a type of error; and, (iv) a malfunction (sometimes by one author on a single page). The majority have characterised it, first and foremost, as a type of error. However, two ways of articulating the error should be conceptually distinguished, even if they are closely related. The first conception sees confabulation as a *representational* or content-level error relating to the *output* of memory processes. The second sees confabulation as "an error in the *process* by which a[n apparent] memory is generated" (Robins 2020: 126), albeit one that *coincides* with the first error.

3.1 Confabulation as output error

On the first conception, confabulations, misrememberings, etc., are distinctions at the level of the representational outputs of systems and processes. Confabulations are, first and foremost, representational outputs that fail to mnemically refer. Or, as Copenhaver (2024) puts it, a confabulating subject "fails to be acquainted with an event in their personal past" (22). This parallels the classical conception of the illusion-hallucination distinction, a distinction, first and foremost, at the level of experience. Depending on one's view, the illusion-hallucination distinction might *entail* certain things about how the experience was produced (e.g., on a causal theory). But the distinction itself applies, in the first instance, at the level of the experience.

On one version, there is a highest common representational factor across cases of remembering, misremembering, and confabulating, and it is when those genuineness or success conditions fail to be met in a particular way that there is confabulation: "sometimes we fail to perceive or remember anything at all, even when we have an experience of an object as present or an event as having happened, as in hallucination and confabulation" (Copenhaver 2024: 24). The same basic kind of memory output can be veridical, falsidical, or fail to refer. And it is in the latter case that we have the distinctive failure of confabulation. Robins' (2020: 122) idea that the common factor is a mental state of 'seeming to remember' is amenable to this reading. It is on this sort of view that it could make sense to say that

> [e]ven in a *functioning* memory system, a set of highly particular circumstances could lead to the production of a confabulation. The possibility is not restricted to malfunctioning systems; it is a possibility that is live in all instances of seeming to remember (Robins 2020: 128).

If confabulation is a *product* that might result from even properly functioning processes, then it is not itself constitutively tied to processing errors.

This conception is difficult to reconcile with the characterisation of mnemic confabulation provided in §2. It makes sense, however, on a view where confabulation is, first and foremost, a distinctive type of representational error in memory output, namely, one consisting in reference-failure. There is something dissatisfyingly incomplete and stipulative about this proposal. I will articulate the worry more fully in §4. For now, notice at least that this is in tension with the characterisation of confabulation (in the narrow sense) provided in §2.[11]

3.2 On apparent cases of 'veridical' confabulation

It is worth briefly remarking on a type of case often adverted to in this debate: confabulations that are, it is said, 'veridical' (Bernecker 2017; Michaelian 2016b).[12] In the sort of case that is inevitably described, the subject's report or representation was not properly produced such as to constitute a genuine instance of remembering, and yet the report or representation nonetheless *coincidentally happens to match* the description or features of some event in the subject's personal past. Suppose a subject S once experienced an event *e* but later lost all memory of *e* due to an amnesic trauma. Years later, suffering from dementia, S fabricates many stories, one of which, by sheer serendipity, "matches exactly" their experience of *e*, though this "does not come about because of any information [retained]" (Robins 2020: 125). As Robins puts it, "confabulations lack a connection to any event in the confabulators past, even if there is *a surface-level similarity* between what's represented and something that did occur, as in cases of veridical confabulation" (2020: 130). Even in these cases, "[t]he hallmark of confabulation is that if there is a correspondence between the recalled content and the past content, it lacks the modal stability required for [genuine] remembering" (Bernecker 2023: 113). In other words, even if there is a past representation which the current memory representation 'matches', the current representation fails to *track* its content: at close worlds where the past content is different, the 'recalled' content is not.

[11] If we want an analogy for confabulation, a better one may be dreaming. There are relevantly analogous issues that arise in cases of sensory incorporation here concerning whether one mishears one's alarm clock as a siren or auditorily hallucinates a siren, and whether one could, in principle, successfully perceive in some aspect 'through' a dream (see Macpherson (2024) for discussion).

[12] On so-called 'veridical hallucination' and its role in causal theories of perception, see Grice (1961) and Lewis (1980).

We should be reluctant to accept that these cases of mnemic confabulation involve genuine veridicality. The absence of *any* causal connection—or counterfactual stability—between one's representation and an event in one's past is at odds with the representation being *about* such an event at all. And if the representation fails to refer, there is just no fact of the matter whether the representation is veridical or not (Openshaw 2023: 298–9). There is no privileged event the properties of which are germane to accuracy evaluation. In other words, 'matching' is cheap and uninteresting. Such cases of entirely serendipitous 'veridical' confabulation should leave us unmoved, for they do not involve genuine semantic relations, nor therefore genuine veridicality.[13]

3.3 Confabulation as process error

On a second conception, confabulation is a *process error*. There are many ways to formulate such an account, and many ways to therefore cash out the characterisation of mnemic confabulation I offered in §2. However, on the sort of view I wish to target for discussion here, the processing error is not output-neutral but rather *coincides* with—and perhaps explains—the sort of error described in §3.1. The term denotes a process that fails to suffice for remembering, but one which coincides with a particular type of representational output, one not found in mere misremembering. For instance, according to Robins, while misrememberings "result from distortions of retained information",

- "Confabulation errors […] are *wholly inaccurate*, reflecting no influence of information retained from a particular past event" (2019: 2148; emphasis added).
- "Confabulations are *not simply false memory reports*, but reports that lack any substantive contact with information retained from a particular past event. To confabulate is to claim memory of a past experience that is *false in its entirety, not only in detail*" (2019: 2149; emphasis added).[14]

There are two elements here, reflected in the italicised and underlined portions of the text. As I understand Robins' claim, the underlined text is the proposed *explanans* of the error

[13] For a similar attitude to certain 'veridical hallucination' cases, see Schellenberg (2018: 95; 181). I will suggest in §5 that mnemic confabulations that are to some degree veridical *are* possible, but they involve a genuine semantic relation—a reference/aboutness relation—rather than mere descriptive 'matching'.

[14] It is worth acknowledging that Robins (2019) takes her account to be 'initial speculation'. In her (2020), Robins retracts the claim that confabulations must be false (126), in light of the possibility of so-called 'veridical confabulation'.

partly described by the italicised text. In a later paper, Robins (2020) develops an account of confabulation that is more explicitly *causalist* in this way. To such views we now turn.

4 The causalist account of genuineness *and* mnemic reference

A dialectical advantage causalists have claimed over their opponents is that causalists can neatly explain the putatively distinctive type of error in mnemic confabulation. According to Robins (2020), confabulations are "errors because they lack a causal connection between the event and its representation" (126). More fully, confabulation occurs

> when there is no relation between a person's seeming to remember a particular event or experience and any event or experience from their past—either because there is no such event in their past or because any similarity to such an event is entirely coincidental (2020: 125–6).

Similarly, Bernecker (2017) claims that mnemic confabulations differ from genuine rememberings in that "they fail to be suitably causally connected to the corresponding past representations, either because there are no corresponding past representations or because the causal connection has been severed" (12). Though their primary focus is the semantics of memory reports, Werning & Liefke (2024) echo this same idea: "in misremembering, the mnemonic attitude is grounded in a past experience", but "in the case of confabulation, the mnemonic attitude is not grounded in a past experience, either because there was no such experience or because the mnemonic attitude does not depend on that experience" (122).[15]

The driving motivation behind these accounts is commitment to causalism. Confabulations are errors *because* they lack an appropriate causal connection between an event in the subject's past and their present memory representation. So, absence of an appropriate causal link is *sufficient* for an apparent memory representation to be a mnemic confabulation. It follows from natural assumptions that the *presence* of such a causal link is *necessary* for remembering.

Given the historical orthodoxy of broadly causal metasemantic theories of proper names (Kripke 1980; Evans 1973) and of reference-fixing in singular thought (Devitt 1981),

[15] Moreover, confabulations do not belong to the same natural kind as genuine (mis)rememberings: confabulation differs "with regard to the underlying causal mechanism"; namely "a causal connection to a foregoing experience […] is explicitly negated" (Werning & Liefke 2024: 147).

it is unsurprising that causal theories of mnemic reference-fixing have enjoyed similar dominance. If reference is always—or at least paradigmatically—determined by a causal chain linking the tokening of a singular term or thought-vehicle to its referent, the appeal of a causal theory of mnemic reference-fixing is easy to see. The pattern is familiar: a causal link, suitably mediated and sustained under the right conditions, ensures reference.[16] Werning and Liefke (2024) are particularly clear in adopting this explanatory strategy: "the primary experience that underlies a particular episodic memory must be uniquely identifiable. This can be achieved by a minimal [memory] trace alone" (149). It is the causal aetiology of an episodic memory trace that not only uniquely determines the referent of a memory representation but also *makes it a genuine memory* representation.

As such, causalists employ the notion of appropriate causation to do two distinct things: (i) distinguish genuine (mis)remembering from confabulating; (ii) explain mnemic reference. Appropriately caused memory traces not only distinguish remembering from confabulating, they also constitute the means of mnemic reference-fixing. Insofar as causalists make use of one notion of appropriate causation to do (i) and (ii), confabulation is not merely a *processing error*, it is a distinctive form of *content error* (namely, one in which there is reference-failure). The possibility of referential confabulation is an apparent blind spot. Causalists face a dilemma: either they must give up their hallmark causal theory of reference, or they must deny that referential confabulators are genuine confabulators.[17]

Empirical evidence suggests that mnemic reference can survive in the absence of genuine (mis)remembering. I will address several objections to the challenge in §6. But it is worth anticipating a knee-jerk response at this stage, namely that the very notion of referential confabulation is incoherent. In light of the methodological remarks made in §2, the causalist should have no interest in making claims that are *true by stipulation*. Mnemic confabulation is not a philosophers' invention; it is an elaboration of a phenomenon discussed in the memory sciences.[18] The causalist account of mnemic confabulation is a

[16] Soteriou (2018) likewise suggests: "which particular past event is represented [...] is determined by the causal ancestry of the memory" (308).

[17] There are independent reasons a causalist may need to give up a *purely* causal story of reference-fixing, for instance the alleged aetiological promiscuity of traces. Causalists might adopt a hybrid story that also appeals to 'consumer-side' factors (Barkasi 2024: 18). I set this aside here for simplicity: the additions to the story will not provide the means to distinguish genuine remembering from mnemic confabulation.

[18] Interestingly, however, Robins suggests: "Mnemonic confabulation became important to philosophers of memory because consideration of this possibility influences the requirements on remembering, not because there are actual cases of confabulation being reported as symptomatic of various clinical conditions" (2020: 127). Insofar as Robins is right that, for some theorists, confabulation is a consideration by analogy with hallucination as a kind of 'case' for ascertaining necessary conditions, the methodology is unlikely to yield insights into memory as it actually works.

substantive (explanatory) claim, not a triviality. If there is a tendency to sometimes use the term 'confabulation' to refer to any case in which a subject has an apparent memory representation which was not appropriately caused via a memory trace, this should be resisted in the interests of terminological clarity. And we can perfectly well discuss mnemic event representations suffering reference-failure without succumbing to a temptation to call them 'confabulations'. To continue using the notion of mnemic confabulation in that way risks depriving us of interdisciplinary coordination on a phenomenon of great interest.

5 Referential confabulation

Historically, confabulation has sometimes been characterised as *compensatory fabrication:* "the emergence of memories of events and experiences which never took place" (Wernicke 1900: 139; cited in Schnider (2018)), wherein gaps, "caused by severe memory failure, are filled spontaneously or upon incitation with false memories" due to what is at some level a "desire to create an image of the past which has left no real traces" (Kraepelin 1909: 263; cited in Schnider (2018)).[19] This view is no longer empirically tenable. Over time, *temporal displacement* has come to be considered a central feature of the phenomenon.

Moll (1915: 428ff) distinguished five types of *fabrication*, contrasting, *inter alia*, 'random fabrications' in which the patient attempts to fill an amnesic gap through *ad hoc* invention, and more or less *true* recollections, incorrectly oriented in place and time. Van der Horst (1932) also emphasised the interest of those cases of confabulation in which "[t]he event itself is well remembered, but the temporal label has been lost" (68; cited in Schnider (2018)). After lamenting the lack of clarity with which the term has been used (see §2), Berlyne (1972) proposed that we distinguish "temporally displaced true memories" from "wish-fulfilling fantasies". Talland (1961) went so far as to suggest that whereas confabulation proper "draws on some actual incidences in the patient's past, displacing them in their temporal sequence or in their social and geographical context" (365), cases going beyond this are to be terminologically distinguished ('fabrications'). What sets confabulation apart, for Talland (1961: 366), is narrative incorrectness grounded in *temporal disorientation*.

Leading off from these developments, a cluster of modern theories we may call *temporal context* approaches propose that confabulation is the result of deficits in temporal consciousness (Dalla Barba 2002) or, more generally, in the ability to retrieve or identify–

[19] "Confabulations are free inventions" that "fill a void in memory […] devoid of any connection with a real event" (Bleuler 1923: 80ff; cited in Schnider (2018)).

and monitor the identification of—the temporal context of an apparent memory. On one incarnation of this approach, Schnider & Ptak (1999) suggest that (spontaneous) confabulations result from "an inability to suppress activated memory traces and associations at the right time" (680). Schnider (2018) qualifies this account, though the basic idea remains influential in the broader literature (see, e.g., Servais et al. (2023: 1708)).

Temporal context approaches have frequently been criticised for overemphasising chronological deficits. Such deficits, it is said, are likely to be symptoms of a deeper disorder than confabulation's general and principal cause. The explanation favoured by such approaches is more difficult to reconcile with the genuinely 'fantastical' reports one sometimes finds in the literature,[20] and with those that are more 'semantic' in character. Moreover, the compensatory secondary confabulations (observed in Moscovitch's interview (see §2)) do not seem like otherwise healthy reactions to conflicting beliefs.

However, temporal displacement is also considered a central feature of confabulation on increasingly dominant, *strategic retrieval* approaches (Burgess & Shallice 1996; Gilboa & Moscovitch 2002; Kopelman 1999). Like temporal context approaches, strategic retrieval approaches also de-emphasise fabrication. But what they emphasise in its place are breakdowns in retrieval and monitoring processes which also frequently result in temporal displacement. The approach builds on a widely accepted contrast between *associative* and *strategic* retrieval. The former is automatic, specific, and cue-dependent, such that a target memory (so to speak) is elicited immediately. For instance, this is likely to occur if one is asked 'Where is Paris?' In *strategic* retrieval, by contrast, a target memory is not elicited immediately and must be recovered through complex search processes, akin to a form of problem solving. Strategic retrieval is more likely to be needed if one is asked, e.g., 'When were you last in Paris?' These processes effectively form and refine a query, then guide the search via proximal cues and associative memory processes. The output is then monitored against the query, task demands, and other knowledge, to verify its plausibility. The output is also attributed to a source by heuristic processes based on contextual detail, sensory information, etc. (Johnson et al., 1993).[21] Confabulation may be constituted by a breakdown at any of these stages.

---

[20] It may be possible to strengthen temporal context approaches against this particular charge by combining insights from the source/reality monitoring literature, suggesting, e.g., that the patient who reported having met a woman with a bee's head (Turner et al., 2010) was remembering a merely imagined event. Likewise for Damasio et al.'s (1985) subject who reported he had been a 'space pirate' (another subject who suffered AcoA aneurysm).

[21] Source/reality monitoring errors can be contributing factors in mnemic confabulation but are on their own neither necessary nor sufficient.

On the strategic retrieval approach to confabulation, the offending breakdown might be not the failure to retrieve a trace but to retrieve a *task-relevant* trace (and to monitor success or failure in doing so). We can then expect that, as with temporal context approaches, "confabulations often refer to actually experienced events that are chronologically distorted" (Moscovitch 1995: 244). To illustrate this point, and to underscore the deep difference between referential confabulation and so-called 'veridical confabulation', I will present a few cases from the empirical literature which appear to involve manifestations of it. However, I would like to emphasise that although many cases of confabulation one finds in the empirical literature have this character, I do not mean to suggest that this is always or even most often the case. Confabulation is a heterogeneous phenomenon, and its features vary across disorders, for instance between Korsakoff's, Alzheimer's, and schizophrenia (see, e.g., Shakeel & Docherty (2015)).[22] It is enough for my purposes here that these cases regularly occur. With that said, the following cases are intended to help to incline the reader toward thinking subjects can mnemically refer without genuinely remembering. So the causalists' two uses of appropriate causation must be prised apart.

Case 1. Dalla Barba et al. (1990)'s 67-year-old patient C.A. was diagnosed with Korsakoff's syndrome and exhibited severe memory impairments. These appeared to be selectively episodic in nature. C.A. was severely impaired in autobiographical recollection for recent or remote events in her past. However, she performed reasonably well on tests typically thought to tap semantic memory. At times she refused to admit she was in a hospital or claimed that she was there visiting a friend. Dalla Barba et al. (1990) note that her confabulations were always provoked and never spontaneous. In a qualitative analysis on C.A.'s confabulatory responses, Dalla Barba et al. (1990) suggest that their content was never markedly bizarre and often "consisted of real memories framed in a wrong temporal context" (533). For instance, C.A. was asked what she did last Christmas. She replied, "I went to church and came back home to help my mother cook Christmas lunch for me and my brothers" (530). While such an event is thought to have taken place in C.A.'s remote past, C.A.'s mother had died 15 years prior to the interview, and one of her two brother 30 years prior. Dalla Barba et al. (1990) suggest:

[22] An investigation in one patient of the extent to which her confabulations were "'real memories' jumbled up and confused in temporal sequence" was carried out by Kopelman et al. (1997: 705ff). With the aid of the patient's brother, they identified some confabulations that had this character, alongside many that apparently did not (1997: 702).

> This constant inability to provide an adequate answer to the same type of questions in the different test sessions, suggests more a degraded representation of AB [autobiographical] memories than a failure in the access to AB traces (533).

It is natural to conclude that, in some such cases, C.A. is indeed confabulating, but that she is also mnemically referring to (contextually irrelevant) events in her personal past.

Case 2. Dalla Barba's (1993) patient S.D. was a regular runner who suffered a serious head trauma following a 200m fall in the mountains. Upon examination by a neurology department 10 months later, S.D., aged 37, was found to be densely amnesic and impaired across episodic and semantic memory tests. Asked what he did yesterday, S.D. replied that he had won a running race, and that he had been awarded with a piece of meat which was put on his right knee. Albeit implausible, Dalla Barba & Boissé note, in the context of this case, that "[c]onfabulations often described as fantastic or implausible […] are often made of autobiographical elements put together in an inappropriate semantic structure", for it "was actually during a running race in the mountains that he fell", suffering "a severe head trauma and an open wound in his right knee" (Dalla Barba & Boissé 2010: 97). Of course, there is again some speculation involved, but it is somewhat plausible that S.D. is both confabulating and mnemically referring to a specific event in his past, namely his accident.

Case 3. Finally, a case to suggest that confabulations vary in their apparent temporal specificity and stability (Burgess & McNeill 1999).[23] Confabulations can exhibit recurrent themes reflecting the patient's personal concerns or experiences. 51-year-old patient B.E. suffered 'content-specific confabulations' for 12 weeks while recovering at home from major surgery. B.E. would wake up each day and seem to remember, having apparently spoken to his business partner the previous day, committing to carry out a stocktake. Exactly where, he was never sure, and he came to different conclusions on different days. His wife, initially alarmed to find him getting ready for work, would ask him to call his business partner to clarify the situation. On subsequent occurrences, B.E. would remember the previous day's (real) telephone call, but each new day would believe something new had since been agreed. B.E.'s daily confabulation was stable and isolated: he did not confabulate about other areas of his life. Burgess & McNeil (1999) explain B.E.'s case in two parts: (i) a stable and selective 'generic confabulation' (a displaced generic memory, 'created by repetition and in transition between 'pure episodic' and 'pure semantic'

[23] La Corte et al.'s (2011) T.A. also suffered confabulations that "invariably consisted of habits or true memories misplaced in time" (312), though he is not reported as having acted on them.

representation'); (ii) inability to self-initiate retrieval of specific event memories that would conflict with the confabulation. It is quite plausible, then, to suggest that B.E. is mnemically referring yet mnemically confabulating.[24]

We should sometimes credit confabulators with successfully *mnemically referring* to events in their personal past.[25] But we should not thereby, in all such cases, credit them with *remembering*. This poses a distinctive challenge for causalists, for they appeal to the *presence* of appropriate causation via a memory trace to explain mnemic reference and to its *absence* to explain mnemic confabulation. So what of referential confabulations?[26]

Confabulation is not always quite like classic cases of sensory hallucination, for there is often the analogue of sensory input.[27] Often, as we have seen, confabulation involves the activation of episodic traces. It is, to that extent, more like some forms of perceptual illusion. To put it one way, while visual reference might entail *seeing* (though cf. Lande (2023)), mnemic reference does not entail *remembering*. So while aspects of a sensory experience that are (in the traditional sense) hallucinatory cannot involve visual reference, aspects of seeming to remember that are confabulatory *can* involve mnemic reference.

Insofar as causalists use appropriate causation to distinguish remembering from confabulating *and* to explain mnemic reference, referential confabulation will be a blind spot. For causalists, if one has an apparent memory representation that is appropriately caused by one's experience of the event, one *just is remembering*. Were the causalist to grant, for example, that patient C.A. mnemically refers to an event in her past, their account would entail that C.A. is simply remembering.[28] While the causalist's simple, readymade

---

[24] Do so-called 'generic memories' involve mnemic reference? On one view, they refer to *event-types* (Entwistle 2024). On another, they are referentially indeterminate (Openshaw, forthcoming). Whatever the details, it is implausible to think only recollections of specific, temporally discrete events involve reference.

[25] Note that the mnemic reference in these cases is not merely to event-*constituents* (individuals, places, etc.). Even the wildest confabulations will feature such reference. Rather, the mnemic event representation mnemically refers to some *event(s)* in the subject's personal past. I take it to be this sort of reference that is supposed to require appropriate causation via an episodic trace, and, as such, is my target.

[26] Referential confabulation might also raise challenges for other accounts, such as Bernecker's (2023) explanationist model, for they are surely not *coincidences*, and to the extent that they are genuinely accurate such accuracy can be explained by the corresponding facts.

[27] I do not mean to imply that I think hallucination *can* be understood as a distinctive sort of content-level perceptual error (namely, reference-failure), or that the traditional illusion-hallucination distinction is correct. For some relevant discussion, see Macpherson & Batty (2016) and James (2014).

[28] An anonymous reviewer notes that the causalist might reply as follows: for genuine remembering, an apparent memory representation must be not only appropriately caused by one's experience of an event but must also sufficiently 'match' it. In that case, the entailment in the main text might be blocked. The notion of matching is unclear without further elucidation, but the move holds some plausibility. Yet it does not get to the heart of the problem: the entailment would still go through were it granted that (e.g.) C.A.'s apparent memory representation *does* sufficiently 'match' some experience(s) in her past. And I think it is

account of mnemic reference is often considered an advantage, perhaps it is not. In the next section, I consider some responses to the argument from referential confabulation, before using these to reflect on the implications for the causalist theory once it is suitably modified.

6 Two objections

The first objection to consider denies that such cases are really confabulations at all, assimilating them to misremembering. The second allows that confabulation occurs but insists that it can coexist with mnemic reference via misremembering. I argue that neither succeeds in defusing the challenge. To anticipate, I suspect that there is nothing motivating these objections beyond *a priori* commitment to the view described in §3.

6.1 Objection 1: Referential confabulation is just misremembering

One line of objection holds that so-called 'referential confabulation' is misclassified. On this line of objection, insofar as patient C.A. is mnemically referring to some past occasion(s) of cooking Christmas lunch for her family with her mother, she is *remembering* such occasion(s). She is simply *misremembering* when the event(s) took place. Along the lines of this objection, Robins (2019) has suggested that some everyday memory errors may be confabulations and some clinical errors may be mere misrememberings: "the case of Alzheimer's patients confusing the temporal order of events […] looks to be a case of misremembering" (2148). As I discussed earlier, this way of developing the objection treats confabulation as a particular kind of output or content error, rather than a processing error that is perhaps constitutively a malfunction. While the objection need not be developed in exactly this way, it does depend on drawing a contrast between 'genuine confabulations' and 'mere misremembering' in a way that *overlaps the contours of mnemic reference*.

However, it is unclear that this contrast is motivated by anything other than prior commitment to causalism. I have argued that mnemic confabulation is not characterised by content-level features but by malfunction of retrieval-related processes, in line with the psychological literature. In cases like C.A.'s, retrieval misfires: she retrieves task-irrelevant information, fails to monitor its inappropriateness, and constructs a contextually false

clear enough that *some* cases of mnemic confabulation (even if not C.A.'s case) will involve at least as close a 'match' as we would be willing to require of everyday instances of genuine remembering. (Note that the issue here is not what *accurate* remembering requires, since it would be enough of a problem for the causalist to accept that C.A. is genuinely remembering, albeit inaccurately, the event(s) in her distant past.)

narrative. These are not markers of misremembering but of confabulatory malfunction. Genuine (mis)remembering is characterised by proper, task-relevant retrieval, not just the activation of some memory trace or other. The *right* trace must be identified by proper search mechanisms, and its output monitored and evaluated against the cue demands. That this objection is at odds with the psychological literature matters, because we cannot be taken to have *intuitions* about whether something counts as confabulation or as misremembering.

Second, recasting such cases as mere misrememberings imposes a heavy burden on the distinction. It would imply that a vast number of clinical cases typically classified as confabulation—especially those involving temporal displacement—are, in fact, just misrememberings. Given just how often mnemic confabulation is taken to involve temporal displacement in the clinical literature, this line of objection dramatically decreases the number of genuine mnemic confabulations. What this revisionist objection needs is not only that some of the cases I have described or alluded to are mere misrememberings, the objection involves the claim that *all* cases which involve genuine mnemic reference *just are* misrememberings. I think we should be pessimistic about the prospects for this strategy, but it suffices to say that the burden is on revisionist causalists to make the case.[29]

6.2 Objection 2: Referential confabulation and misremembering can coincide

According to a more concessive line of objection, insofar as patient C.A. is mnemically referring to some past occasion(s) of cooking Christmas lunch with her mother, she is misremembering. But this does not mean that she is not also confabulating. While the misremembering aspect of the case explains mnemic reference (via appropriate causation), other aspects of the case suffice to mark out the distinction between confabulation and (mis)remembering.

The first problem with this reply is that, once again, the claim must be that what is *generally* going on across the relevant class of confabulation is that the subject is both misremembering and confabulating. In that case, we again get the conclusion that there are far fewer instances of full-fledged confabulation than are usually recognised across the memory sciences.

[29] One avenue to consider is that "investigation may reveal that the distinction between misremembering and confabulation is blurry; these errors may be better understood as ends of a continuum" (Robins 2019: 2149; see also Moscovitch (1995: 246)). How to square this with the idea that "confabulation is a malfunction" (Robins 2019: 2141) is unclear.

Second, it is true that there is some temptation to distinguish between what Moran (2022) has called (borrowing the partial-total hallucination distinction in philosophy of perception) 'partial' and 'total' confabulation. The causalist might insist on zooming in to the fine-grained characteristics of C.A.'s mnemic representation, suggesting that certain elements comprise genuine remembering while others are confabulatory, so that this is a case of partial confabulation. While many aspects of C.A.'s mnemic representation may be causally connected to a past event (explaining mnemic reference), the confabulatory aspects lie in fictitious embellishments in other aspects. Confabulation and remembering can thus co-occur in a single episode, just as hallucination and veridical perception might blend in certain visual illusions. However, this changes the subject matter, once again making confabulation a distinction at the level of content-bearing representational output, rather than a distinction in the underlying processes that produce them. Furthermore, many clinical cases traditionally classified as confabulation would now become hybrid episodes—part misremembering, part confabulating. That may be descriptively tolerable, but it burdens causalists with explaining how appropriate causation suffices for reference while not for remembering, and how this supposed mixture maps onto mnemic function and malfunction. This will be a far more complicated story than they have so far been inclined to tell.

Finally, this reply wrongly inflates the explanatory role of memory traces. Traces can be activated in ways that fall far short of proper retrieval—e.g., during dreams. What distinguishes remembering from these cases is not merely the presence of an appropriate causal link but the success of the system(s) in solving the memory task: identifying a relevant trace, retrieving it strategically, and monitoring its contextual fit. In the absence of such success, the subject is just not remembering, even if a trace is causally implicated. The objection therefore risks mistaking episodic retrieval for genuine remembering.

7 Implications for causalists

On the picture of confabulation that has emerged, the specific sense of 'ill-groundedness' in the characterisation of (mnemic) confabulation$_N$ provided in §2 is one attributable to malfunction of strategic retrieval (failure to identify a task-relevant episodic trace) and monitoring (failure to detect the first failure).

Confabulation$_N$ The honest, confident, and literal production of an ill-grounded narrative, characteristically concerning one's past and with apparent

> epistemic first-handedness of source, the ill-groundedness of which is at least partly attributable to error of some distinctively *mnemic*, retrieval-related process.

It is a sense that is compatible with referential success, perhaps in those cases where some (task-irrelevant) episodic trace is retrieved and monitoring processes fail to identify its irrelevance or contextual implausibility.

What should the causalist conclude from the phenomenon of referential confabulation? Causalists must distinguish their explanatory story for mnemic reference from their explanatory story for genuine remembering. The most promising move is to supplement their account with a condition on process integrity: strategic retrieval and monitoring must function properly for an episode to count as remembering. In other words, the conflicting pressures on the notion of appropriate causation could be alleviated by claiming that, while entertaining an apparent memory representation that is appropriately causally connected—via an episodic memory trace—to some event(s) in one's personal past suffices for mnemic reference, it does not suffice for genuine remembering. Genuine remembering requires, moreover, properly functioning strategic retrieval and monitoring processes. Patients C.A., S.B., and B.E. might be mnemically referring to their personal past, but they are not remembering.

I think that this is the right reaction for the causalist.[30] Genuine remembering involves more than entertaining mnemic representations grounded in appropriate causal links to past events, it involves proper functioning of the mechanisms that manage those links. Patients C.A., S.D., and B.E. are mnemically referring to their personal pasts, but they fail to be genuinely remembering because the mechanisms responsible for strategic retrieval and monitoring malfunction.

However, if genuine remembering requires (and differs from confabulation to the extent that it involves) properly functioning strategic retrieval and monitoring processes, over and above the having of a mnemic event representational appropriately causally connected via an episodic trace to some event(s) in the subject's personal past, then there is a whole facet of the causalist story on which even the most sophisticated causalist theories of memory (e.g., Werning (2020)) have been more or less silent. If the boundary between genuine remembering and mnemic confabulation turns on the integrity of executive processes, causalists can no longer explain memory's reliability—or its breakdowns—by appeal to appropriately caused memory traces alone.

---

[30] I am grateful to Nikola Andonovski for discussion of ideas in this section.

Moreover, it would seem that to buy into something like this characterisation of mnemic confabulation, the causalist must now think of strategic retrieval and monitoring as core *mnemic* processes. If properly functioning strategic retrieval and monitoring processes are key components of genuine remembering, this raises some big questions. It might entail rethinking the architecture of memory itself. Insofar as monitoring processes are employed not only in remembering but in the service of other capacities, does this further suggest that autobiographical remembering is an interaction effect, constitutively drawing on not only multiple declarative memory systems but executive processes too? It is not uncommon to suggest that the accuracy or success conditions of remembering are situation-dependent, so that they vary along with pragmatic features of the memory task. But the current suggestion might imply that the *genuineness* conditions of remembering are also situation-dependent. For whether one is genuinely remembering will often depend on whether strategic retrieval and monitoring processes properly function, a question to which there may be no univocal, situation-insensitive answer. Once causalists acknowledge that appropriate causation is not sufficient for remembering, the elegance and simplicity standardly claimed for such theories requires some new honest labour to maintain.[31]

8 Conclusion

If I am right that there is a robust class of cases properly describable as referential mnemic confabulations, causalists face new challenges. The root of these challenges is that causalists use appropriate causation to pull off too many feats: to be the key mechanism that underlies remembering as a natural kind; to explain what determines mnemic reference; to explain the reliability of remembering; to distinguish remembering from relearning; and, to distinguish genuine remembering from mnemic confabulation. Referential confabulation raises an immediate challenge for views on which an appropriate causal connection to the past is taken both to fix mnemic reference and to distinguish genuine remembering from confabulation. Referential confabulations shows that these two roles can come apart. Appropriately caused memory traces may underwrite reference

[31] Openshaw & Michaelian (2024) suggest that 'post-causalist' theories (which deny that 'appropriate causation' is metaphysically necessary for remembering), which may separate their account of remembering from their account of mnemic reference, can explain cases of referential mnemic confabulation. They are cases which do not involve remembering (e.g., proper functioning of the episodic system) but which *do* involve mnemic reference (e.g., retrieval of an episodic trace). It is only fair to note that the possible inclusion of executive and monitoring processes within Michaelian's (2016a) episodic construction system raises similarly interesting questions.

without underwriting genuine remembering. Causalists can respond by revising their account: they can retain a causal theory of mnemic reference, while requiring additionally that strategic retrieval and monitoring processes function properly for an episode to qualify as remembering. But this revision has implications. It blunts the principal theoretical virtue that causalists have often claimed: the simplicity and elegance of explaining remembering and confabulation alike in terms of appropriate causation. Instead, causalists must appeal to a more complex picture, one that invokes the proper functioning of executive and monitoring processes alongside causal history. I do not take this to be an implausible cost. But causalists have overlooked work to do in elucidating the implications of this on the architecture of memory.

This diagnosis also has broader implications. It cautions against a widespread tendency to define 'confabulation' purely at the level of memory outputs, as a particular kind of content-level representational error, such as reference-failure or radical inaccuracy. Insofar as this tendency to characterise confabulation as a distinctive sort of content-level error is encouraged by drawing a close analogy between confabulation and hallucination, that analogy should be approached with intellectual caution. Confabulation, in its core and empirically-anchored sense, is better understood as a malfunction of distinctively mnemic, retrieval-related processes such as strategic retrieval and monitoring. It is a kind of processing error, not a proprietary kind of representational error. Even when confabulatory memory representations succeed in referring to real past events, they are marked by the breakdown of the mechanisms responsible for proper retrieval and evaluation. Of course, the greater and more radical the error, and the greater degree of irrelevance, the better the evidence is in favour of a subject having confabulated. Mnemic reference-failure might be a particularly good diagnostic, but it is unlikely to be one we ought to include within the core characterisation of mnemic confabulation. Recognizing this complicates causalist theories in particular, challenges easy analogies with hallucination, and demands a more nuanced understanding of the epistemic frailties of remembering.[*]

[*] *Acknowledgements:* This work benefited from presentations at Shanghai Jiao Tong University, the 2024 Joint Session of the Aristotelian Society and Mind Association at Birmingham, an Umeå-Uppsala colloquium, the IPMC at National Yang Ming Chiao Tung University, the University of Cardiff, and the Unconventional Memory workshop at LMU Munich. Thanks to many for their questions and comments and, in particular, to Kirk Michaelian for collaboration on work from which some of these ideas originated.

Funding: Work on this project was supported by a Nanyang Technological University Start-Up Grant. At earlier stages it received funding from ORCHID grant number 112-2927-I-A49A-501, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101032391, and the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02).

Bibliography

Baddeley, A. and Wilson, B. 1986. 'Amnesia, autobiographical memory and confabulation', in D. C. Rubin (ed.), *Autobiographical Memory*. Cambridge: Cambridge University Press.

Barkasi, M., & Sant'Anna, A. 2022. 'Reviving the naïve realist approach to memory', *Philosophy and the Mind Sciences* 3: 14.

Barkasi, M. 2024. 'Consumer-side reference through promiscuous memory traces', *Synthese* 203(86). DOI: 10.1007/s11229-024-04509-y

Berlyne, N. 1972. 'Confabulation', *British Journal of Psychiatry* 120: 31–39.

Bernecker, S. 2017. 'A causal theory of mnemonic confabulation', *Frontiers in Psychology* 8: 1207.

Bernecker, S. 2023. 'An explanationist model of (false) memory', in *Current Controversies in Philosophy of Memory*. New York: Routledge.

Bleuler, E. 1923. *Lehrbuch der Psychiatrie*. Berlin: Julius Spinger Verlag.

Bortolotti, L. 2018. 'Stranger than fiction: Costs and benefits of everyday confabulation', *Review of Philosophy & Psychology* 9: 227–249.

Burgess, P. W. and Shallice, T. 1996. 'Confabulation and the control of recollection', *Memory* 4(4): 359–411.

Burgess, P. W. & McNeil, J. E. 1999. 'Content-specific confabulation', *Cortex* 35(2): 163–182.

Cao, S., Zhu, X., Zhang, W., & Xia, M. 2020. 'Anton's syndrome as a presentation of Trousseau syndrome involving the bilateral optic radiation', *Journal of International Medical Research* 48(11).

Copenhaver, R. 2024. 'Dreams, remembering, and remembering dreams: An intentionalist, direct realist, acquaintance account', in D. Gregory & K. Michaelian (Eds.), *Dreaming and Memory: Philosophical Issues*. Springer.

Dalla Barba, G., Cipolotti, L., and Denes, G. 1990. 'Autobiographical memory loss and confabulation in Korsakoff's syndrome: A case report', *Cortex* 26: 525–534.

Dalla Barba, G. 1993. 'Different patterns of confabulation', *Cortex* 29(4): 567–581.

Dalla Barba, G. 2002. *Memory, Consciousness and Temporality*. Boston, MA: Kluwer Academic Publishers.

Dalla Barba, G., & Boissé, M. 2010. 'Temporal consciousness and confabulation: Is the medial temporal lobe "temporal"?', *Cognitive Neuropsychiatry* 15(1–3): 95–117.

Damasio, A. R., Graff Radford, N. R., Eslinger, P. J., Damasio, H., & Kassel, N. 1985. 'Amnesia following basal forebrain lesions', *Archives of Neurology* 42: 263–271.

Devitt, M. 1981. *Designation*. New York: Columbia University Press.

Entwistle, M. 2025. 'Generic episodic memories', *Synthese* 205(33). DOI: 10.1007/s11229-024-04869-5

Evans, G. 1973. 'The causal theory of names', *Aristotelian Society Supplementary Volume* 47(1): 187–208.

Gilboa, A. 2010. 'Strategic retrieval, confabulations, and delusions: Theory and data. *Cognitive Neuropsychiatry 15*: 145–180.

Gilboa, A., & Moscovitch, M. 2002. 'The cognitive neuroscience of confabulation: A review and a model', in A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), *Handbook of Memory Disorders*, 2nd Edition. London: John Wiley & Sons.

Gilboa, A., & Verfaellie, M. 2010. 'Telling it like it isn't: The cognitive neuroscience of confabulation', *Journal of the International Neuropsychological Society* 16: 961–966.

Grice, P. 1961. 'The causal theory of perception', *Proceedings of the Aristotelian Society* 35(1): 121–152.

Hirstein, W. 2005. *Brain Fiction*. Cambridge: MIT Press.

James, S. 2014. 'Hallucinating real things', *Synthese* 191(15): 3711–3732.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. 1993. 'Source monitoring', *Psychological Bulletin* 114(1): 3–28.

Kopelman, M. D., Ng, N., & Brouke, O. V. D. 1997. 'Confabulation extending across episodic, personal, and generic semantic memory', *Cognitive Neuropsychology* 14(5): 683–712.

Kopelman, M. D. 1999. 'Varieties of false memory', *Cognitive Neuropsychology* 16(3–5): 197– 214.

Kopelman, M. D. 2010. 'Varieties of confabulation and delusion', *Cognitive Neuropsychiatry* 15(1–3): 14–37.

Korsakoff, S. S. 1891. 'Erinnerungstäuschungen (Pseudoreminiscenzen) bei polyneuritischer Psychose', *Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medicin* 47: 390–410.

Kraepelin, E. 1909. *Psychiatrie. Ein Lehrbuch für Studierende und Ärzte. I. Band, Allgemeine Psychiatrie,* 8th Edition. Leipzig: Johann Ambrosius Barth Verlag.

Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

La Corte, V., George, N., Pradat-Diehl, P., & Dalla Barba, G. 2011. 'Distorted Temporal Consciousness and preserved Knowing Consciousness in confabulation: A case study', *Behavioural Neurology* 24: 307–315.

Lande, K. J. 2023. 'Seeing and visual reference', *Philosophy and Phenomenological Research* 106: 402–433.

Lewis, D. K. 1980. 'Veridical hallucination and prosthetic vision', *Australasian Journal of Philosophy* 58(3): 239–249.

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

Macpherson, F., & Batty, C. 2016. 'Redefining illusion and hallucination in light of new cases', *Philosophical Issues* 26: 263–96.

Macpherson, F. 2024. 'Perception in dreams: A guide for dream engineers, a reflection on the role of memory in sensory states, and a new counterexample to Hume's account of the imagination', in D. Gregory & K. Michaelian (Eds.), *Dreaming and Memory: Philosophical Issues*. Springer.

Mahr, J. B., & Csibra, G. 2018. 'Why do we remember? The communicative function of episodic memory', *Behavioral and Brain Sciences* 41: e1.

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

Michaelian, K. 2016b. 'Confabulating, misremembering, relearning: the simulation theory of memory and unsuccessful remembering', *Frontiers in Psychology* 7: 1857.

Michaelian, K. 2020. 'Confabulating as unreliable imagining: In defence of the simulationist account of unsuccessful remembering', *Topoi* 39(1): 133–148.

Michaelian, K. 2023. 'Towards a virtue-theoretic account of confabulation', in *Current Controversies in Philosophy of Memory*. New York: Routledge.

Michaelian, K., & Sutton, J. 2017. 'Memory', in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Michaelian, K., & Wall, C. Forthcoming. 'When misremembering goes online: The 'Mandela Effect' as collective confabulation', in S. Goldberg & S. Wright (Eds.), *Memory and Testimony: New Essays in Epistemology*. Oxford: Oxford University Press.

Moll, J. 1915. 'The "amnestic" or "Korsakow's" syndrome with alcoholic etiology: an analysis of thirty cases', *The Journal of Mental Science* 61: 424–443.

Moran, A. 2022. 'Memory disjunctivism: A causal theory', *Review of Philosophy and Psychology* 13(4): 1097–1117.

Moscovitch, M. 1989. 'Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory', in H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Moscovitch, M. 1995. 'Confabulation', in D. L. Schacter (ed.), *Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past*. Cambridge, MA: Harvard University Press.

Nahum, L., Bouzerda-Wahlen, A., Guggisberg, A., Ptak, R., Schnider, A. 2012. 'Forms of confabulation: Dissociations and associations', *Neuropsychologia* 50(10): 2524–2534.

Nisbett, R. E., & Wilson, T. D. 1977. 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review* 84: 231–259.

Openshaw, J. 2023. '(In defence of) preservationism and the previous awareness condition: What is a theory of remembering, anyway?', *Philosophical Perspectives* 37(1): 290–307.

Openshaw, J., & Michaelian, K. 2024. 'Reference in remembering: Towards a simulationist account', *Synthese* 203(90). DOI: 10.1007/s11229-024-04508-z

Openshaw, J. Forthcoming. 'Memory and reference', in A. Sant'Anna & C. Craver (Eds.), *The Oxford Handbook of Philosophy of Memory*. Oxford: Oxford University Press.

Robins, S. K. 2019. 'Confabulation and constructive memory', *Synthese* 196: 2135–2151

Robins, S. K. 2020. 'Mnemonic confabulation', *Topoi* 39: 121–132.

Sant'Anna, A., & Michaelian, K. 2019. 'Thinking about events: A pragmatist account of the objects of episodic hypothetical thought', *Review of Philosophy and Psychology* 10(1): 187–217.

Sant'Anna, A. 2022. 'Unsuccessful remembering: A challenge for the relational view of memory', *Erkenntnis* 87: 1539–1562.

Servais, A., Hurter, C., & Barbeau, E. J. 2023. 'Attentional switch to memory: An early and critical phase of the cognitive cascade allowing autobiographical memory retrieval', *Psychonomic Bulletin & Review* 30: 1707–1721.

Schacter, D. L., & Tulving, E. 1994. 'What are the memory systems of 1994?', in D. L. Schacter & E. Tulving (Eds.), *Memory Systems 1994*. Cambridge, MA: MIT Press.

Schellenberg, S. 2013. *The Unity of Perception: Content, Consciousness, Evidence*. Oxford: Oxford University Press.

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

Schnider, A., & Ptak, R. 1999. 'Spontaneous confabulators fail to suppress currently irrelevant memory traces', *Nature Neuroscience* 2: 677–681.

Shakeel, M. K., & Docherty, N. M. 2015. 'Confabulations in schizophrenia', *Cognitive Neuropsychiatry*, 20(1): 1–13.

Soteriou, M. 2018. 'The past made present: Mental time travel in episodic recollection', in K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions of Research in the Philosophy of Memory*. New York: Routledge.

Talland, G. A. 1961. 'Confabulation in the Wernicke- Korsakoff syndrome', *The Journal of Nervous and Mental Diseases* 132: 361–381.

Talland, G. A. 1965. *Deranged Memory. A Psychonomic Study of the Amnesic Syndrome*. New York: Academic Press.

Turner, M. S., Cipolotti, L., & Shallice, T. 2010. 'Spontaneous confabulation, temporal context confusion and reality monitoring: A study of three patients with anterior communicating artery aneurysms', *Journal of the International Neuropsychological Society* 16: 984.

Van der Horst, L. 1932. 'Über die Psychologie des Korsakowsyndroms', *Monatsschrift für Psychiatrie und Neurologie* 83: 65–84.

Victor, M. & Yakovlev, P. I. 1955. 'S. S. Korsakoff's psychic disorder in conjunction with peripheral neuritis: A translation of Korsakoff's original article with brief comments on the author and his contribution to clinical medicine', *Neurology* 5: 394–406.

Wernicke, C. 1900. *Grundriss der Psychiatrie in klinischen Vorlesungen*. Leipzig: Thieme.

Werning, M. 2020. 'Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation', *Review of Philosophy and Psychology* 11(2): 301–333.

Werning, M., & Liefke, K. 2024. 'Remembering dreams: Parasitic reference by minimal traces in memories from non-veridical experiences', in D. Gregory & K. Michaelian (Eds.), *Dreaming and Memory: Philosophical Issues*. Springer.