Topoi (2020) 39:121–132
https://doi.org/10.1007/s11245-018-9613-x

Mnemonic Confabulation

Sarah Robins[1]

Published online: 10 December 2018

Abstract
Clinical use of the term "confabulation" began as a reference to false memories in dementia patients. The term has remained in circulation since, which belies shifts in its definition and scope over time. "Confabulation" now describes a range of disorders, deficits, and anomalous behaviors. The increasingly wide and varied use of this term has prompted many to ask: *what is confabulation?* In recent years, many have offered answers to this question. As a general rule, recent accounts are accounts of *broad confabulation*: attempts to unify the seemingly disparate features of all or most confabulatory phenomena under a shared set of characteristics or mechanisms. In this paper, I approach the question differently. I focus on a particular form of confabulation—*mnemonic confabulation*—so as to understand its distinctive features and the ways in which it does (or does not) fit into accounts of broad confabulation. Understanding mnemonic confabulation is a project in the philosophy of memory; it plays an important role in guiding theories of remembering, as a form of error that must be distinguished from genuine remembering. Mnemonic confabulation, as I define it in Sect. 2, 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. This account draws on my own theory of remembering, but shares many important points of consensus with other accounts of mnemonic confabulation, which I highlight in Sect. 3. In Sect. 4, I turn to accounts of broad confabulation—identifying three features such accounts have in common—and, for each, I argue that mnemonic confabulation lacks the requisite feature. As an error, mnemonic confabulation has more in common with perceptual hallucination than with the confabulatory phenomena included in standard accounts of broad confabulation. Recognizing that, despite the shared use of the term "confabulation" mnemonic confabulation and broad forms of confabulation are unrelated, is important for continued progress in debates about each.

Keywords Memory · Confabulation · Hallucination · Remembering · Misremembering

1 Introduction

Clinical use of the term "confabulation" began as a reference to false memories in dementia patients (Korsakoff 1889; Wernicke 1906; see Berrios 1998 for discussion). The term has remained in circulation since, which belies shifts in its definition and scope over time. "Confabulation" now describes a range of disorders, deficits, and anomalous behaviors. Some cases of confabulation do not involve memory. Patients with *anosognosia* (lack of awareness of their illness or injury), for example, will often invent stories to account for the limitations brought on by their illness or injury. Ramachandran (1996) describes a patient experiencing paralysis as the result of a stroke. When asked why he wasn't using his arm, rather than citing the paralysis, he said, "These medical students have been probing me all day and I'm sick of it. I don't want to use my left arm!" (1996, p. 125). Other confabulations involve false memories that occur in everyday life, not in clinical settings (e.g., Loftus and Pickrell 1995). And still others involve neither memory nor clinical patients—as in confabulations of decision-making and moral reasoning (e.g., Nisbett and Wilson 1977; Haidt 2001). The increasingly wide and varied use of this term has prompted many to ask: *what is confabulation?*

In recent years, many have offered answers to this question (e.g., Carruthers 2005; Hirstein 2005; Turner and Coltheart 2010; Strijbos and de Bruin 2015). For the purposes of this paper, what is of interest is the strategy these accounts share, not their details and differences. As a general

✉ Sarah Robins
skrobins@ku.edu

[1] Philosophy Department, University of Kansas, 1445 Jayhawk Blvd., Lawrence, KS 66045, USA

rule, recent accounts are accounts of *broad confabulation*: attempts to unify the seemingly disparate features of all or most confabulatory phenomena under a shared set of characteristics or mechanisms.[1]

In this paper, I approach the question differently. I focus on a particular form of confabulation—*mnemonic confabulation*—so as to understand its distinctive features and the ways in which it does (or does not) fit into accounts of broad confabulation. Understanding mnemonic confabulation is a project in the philosophy of memory; it plays an important role in guiding theories of remembering, as a form of error that must be distinguished from genuine remembering. Mnemonic confabulation, as I define it in Sect. 2, 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. This account draws on my own theory of remembering, but shares many important points of consensus with other accounts of mnemonic confabulation (Michaelian 2016b; Bernecker 2017), which I highlight in Sect. 3. In Sect. 4, I turn to accounts of broad confabulation—identifying three features such accounts have in common—and, for each feature, I argue that mnemonic confabulation lacks it. Mnemonic and broad confabulation accounts differ in the ways that they allow for veridicality; broad confabulations are ill-grounded, whereas mnemonic confabulations are not, and finally, recovery from broad confabulation is possible, at least in principle, while there is no such possibility for mnemonic confabulation. As an error, mnemonic confabulation has more in common with perceptual hallucination than with the confabulatory phenomena included in standard accounts of broad confabulation. Recognizing that, despite the shared use of the term "confabulation" mnemonic confabulation and broad forms of confabulation are distinct, is important for continued progress in debates about each.

2 Remembering

In this section, I offer an account of mnemonic confabulation, situated within a more general account of remembering. Mnemonic confabulation is a particular form of memory error. An understanding of confabulation and other memory errors must begin with an account of successful remembering.

Accounts of remembering in the philosophy of memory can be divided into two general camps: *causal* and *postcausal* views.[2] Causal theorists (e.g., Bernecker 2010, 2017; Debus 2010; Robins 2016, 2017a) argue that remembering requires a causal connection between the event being remembered and the subsequent representation of it, whereas postcausal theorists deny the need for such a connection, developing simulationist, constructivist, or functionalist accounts (Michaelian 2016a, b; De Brigard 2014; Fernandez 2018, respectively) that rely instead on the creation of a plausible representation by a reliable mechanism. Bernecker (2017) and Michaelian (2016b) have offered causal and simulationist accounts of confabulation, respectively. The account I develop here derives from my previous work defending a version of the causal theory (2017b). My account is distinct from both Bernecker's and Michaelian's, but given that the focus of this paper is on the relationship between mnemonic confabulation and broad confabulation, I will focus on the similarities in our views and note the differences only in passing.

One similarity is worth noting quickly at the outset: theories of remembering focus on memory of particular past events or experiences, a form of memory most often referred to as *episodic memory*. Individual accounts may differ in how they characterize episodic memory (e.g., whether it is thought to involve mental time travel), but most accounts of remembering—including the one on offer here—focus on this paradigmatic form of memory. In what follows, all references to remembering should be understood as remembering *episodically*, even if the term "episodic" is omitted. I do not endorse any particular account of episodic remembering as mental time travel, however, I do think episodic remembering requires a mental state of *seeming to remember*. I start with a characterization of seeming to remember below and then identify three requirements that must be met in order for a state of seeming to remember to qualify as remembering.

Developing an account of episodic remembering requires evaluating the set of mental states that involve *seeming to remember*, sorting the cases of successful remembering from memory error. Seeming to remember, as I define it here, occurs when a person has an occurrent mental representation, the content of which targets a representation in her personal past. By framing an account of remembering in terms of seeming to remember, I am making a number of commitments and assumptions. First, I am stipulating that seeming to remember is necessary for remembering.[3] It is not sufficient;

[1] My use of the label "broad confabulation" comes from Bortolotti and Cox's (2009) helpful taxonomic review of definitions of confabulation.

[2] For a more thorough overview of the debates between causal and post-causal approaches to memory, see Michaelian and Robins (2018).

[3] With this commitment, I'm taking a stand on Martin and Deutscher's (1966) "painter case" where they claim a person can remember without seeming to do so. Despite general agreement with Martin and Deutscher on the causal theory, I reject their interpretation of this case. For an argument in favor of including a seeming to remember requirement, see Debus (2010).

many cases of seeming to remember fail to be instances of remembering. This reveals a second commitment: remembering is factive—remembering just is successful remembering. Third, many memory errors involve ways of seeming to remember without actually remembering. These are the errors I focus on here. There are, however, other memory errors that do not involve this seeming—most notably, forgetting.

Further elaboration on the occurrent mental representations involved in seeming to remember helps to make clear what remembering requires and how various memory errors fall short of these requirements. I focus on two structural features: target and content. My appeal to these features is drawn from Cummins' (1996) use of them in his general account of mental representation. The *target* is the aim of the mental representation: what the person doing the representing intends to represent, or takes themselves to be representing.[4] The *content* is the meaning of the mental representation—what is actually represented. If the mental representation were a game of darts, the target would be some segment of the dartboard and the content would be the position on the board where the dart lands. In cases of successful representing (and darts), the target and the content align. But the two can diverge—the target can fail to exist or the content can differ from the target in a number of ways. Cummins uses this potential gap between target and content to explain misrepresentation and other errors.

I am not offering a wholesale defense of Cummins' (1996) view of mental representation. Instead, I want to make use of an important insight embedded in his distinction between targets and contents: for at least some mental states, our evaluation of the mental representation's content requires consideration of the mental representation's target. Cummins introduces the target-content distinction with the example of a chess-playing machine, which determines what move to make by considering, for each possible move, the resulting board position and subsequent moves available (ibid, pp. 5–7). Suppose that, in this process, the machine makes an error: when considering one possible move it represents the board position incorrectly. The machine tokens a representation of the knight in position (c5), but the machine was targeting a representation of the knight after a move that would result in the knight being in position (c7). Understanding whether the content of the board position that has been tokened is an error requires consideration of the target of the representation.

I take this point to be illustrative for remembering (a mental state that Cummins does not discuss). Targets are critical for theorizing about remembering. Determining whether an occurrent mental representation is one of seeming to remember—and in turn, whether that instance of seeming to remember is successful—requires consideration of the target of that mental representation. Targets play two key roles in my account of remembering. First, they unify instances of seeming to remember under a shared target type. Second, they guide the evaluation of whether any token instance of seeming to remember is one of successful remembering. I discuss these in turn.

Seeming to remember episodically picks out a set of mental representations that share a target type, which I term a *3P event*: the target is a *particular* event or experience in the representer's *personal past*.[5] The event is targeted as being in the past, rather than the present or the future (or some hypothetical or counterfactual). Moreover, the target must be situated in the representer's *personal* past—as something that has occurred in his or her lifetime. Finally, the target is a particular event or experience, rather than a general fact (e.g., Buenos Aires is the capital of Argentina) or a broad period of time (e.g., one's childhood). Of course, one can seem to remember that Buenos Aires is the capital of Argentina and seem to remember one's childhood. Restricting the account to particular events or experiences is a way of focusing on the particular form of memory of most interest for philosophers of memory (i.e., episodic memory), through which errors like confabulation are understood. Together, these features of a target make a tokening of a mental representation an instance of seeming to remember episodically.

To determine whether any particular case of seeming to remember qualifies as an instance of successful remembering requires a more fine-grained evaluation of the target and its relation to the content of the occurrent representation. Let's consider an example. Jamari is in a state of seeming to remember—he has a mental representation of opening a gift at his college graduation party; inside the box is a beautiful watch. This counts as an attempt at remembering because he is targeting a particular event in his personal past. The target is the graduation party.[6] The content is the information

[4] Cummins argues that "the notion of a representational target is essentially a functional notion" (1996, p. 7).

[5] Cummins does not discuss elements of the target or differentiate between target types. This elaboration is my own development. It seems possible that distinguishing amongst target types could provide a fruitful way to classify distinct types of mental states. I will not, however, have time to pursue that line of thought any further here.

[6] The event targeted here could be specified more finely: perhaps it's the opening of a gift at this party rather than the entire party. In further development of this view, I will have to say more about how targets are individuated—and what I have to say will be informed by empirical work on event processing (most especially Event Segmentation Theory, Kurby and Zacks 2008). For now, I think that it is enough to say that the scope of an event can, and likely will, vary across contexts. Getting the target right is important, of course, because how the target is understood will influence our determinations of when a memory is in error. Thanks to an anonymous reviewer for encouraging me to make this feature of the account more explicit.

Table 1 Remembering, memory errors, and the requirements that distinguish them

	(1) Target exists	(2) Accurate content (content = target)	(3) Causal history (content produced by target memory trace)
Remembering	Yes	Yes	Yes
Misremembering	Yes	No	Yes
Relearning	Yes	Yes	No, causal history is deviant
Confabulation	Yes or no	Yes or no	No, there is no causal history

contained in the occurrent mental representation he forms.[7] Jamari's representation would likely include details about the gift—that it was a small box, wrapped in blue and silver paper, containing a watch with a dark leather strap. These contents might be represented as propositions or as images. The representation might also include physical sensations, like the feel of the wrapping paper and the weight of the watch, and other phenomenological features or emotional qualities.

There are three conditions that must be met for this, or any, instance of seeming to remember to be actual remembering:

(1) *Target*

The first condition is a stipulation about the targeted event: it must exist. Jamari must have had a graduation party. The existence of the target is critical not only because memory is factive, but because the targeted event frames the evaluation of the next two conditions on remembering.

(2) *Accuracy*

The content of the mental representation must be accurate. Accuracy is not simply a matter of the representation being true, but of its being true *of the event targeted*.[8] Jamari may have received such a watch, but if he received it for a birthday rather than graduation, it would fail to be an instance of successful remembering.

(3) *Causal history*

The third requirement is distinctive of the causal theory: the content of the representation must have been produced in the right way, where the 'right way' involves a causal connection between the original event and the current representation, and more specifically, a causal connection maintained by a memory trace (a mental and/or neural mechanism for retaining memories). Often causal theorists articulate this requirement as the need for an unbroken causal path between the event and the subsequent remembering. I prefer instead to think of the condition as a requirement on the causal history of the mental process by which the representation is produced. Jamari's representation of the watch now must have been produced by a memory trace he formed as a result of the graduation party.

Memory errors occur when one or more of these conditions are violated. I discuss three memory errors below: misremembering, relearning, and confabulation. These errors can involve violations of multiple conditions, but each error is characterized most essentially by a failure to meet one of the three conditions. Table 1 displays the requirements on remembering, illustrating how they are involved in successful remembering, misremembering, relearning, and confabulation.

2.1 Misremembering

Misremembering is an error that concerns the accuracy of a memory's content. It arises because of a mismatch between the content and the target. Robins (2016) defines misremembering as follows:

> Misremembering is a memory error that relies on successful retention of the targeted event. When a person misremembers, her report is inaccurate and yet the error is explicable only on the assumption that she has retained information from the event her representation mischaracterizes (p. 434).

This form of misremembering can be illustrated by tweaking the Jamari example from above. Suppose Jamari is in a state of seeming to remember his college graduate party. His representation targets that event in his personal past, but the

[7] I avoid endorsing any particular view about the nature of mental content in general or the contents of episodic memory in particular. My aim is to remain ecumenical, as what matters for my purposes is the relation of the target and the content, however that content is understood.

[8] Much more needs to be said about how to determine the accuracy of memory. Bernecker (2010) makes a helpful distinction between truth and authenticity, where truth refers to the objective veridicality of the memory state and authenticity refers to its relation to the person's earlier representation of the event. Bernecker claims that both are required for veridicality. An alternative, made possible by the framework I have introduced here, would be to allow that features of how the event is being targeted (e.g., the event as it occurred or the event as it was experienced) can be used to assess the accuracy of the representation.

content of his representation is not entirely accurate. Maybe his representation depicts the wrapping paper on the gift box as gold, rather than blue and silver, or depicts the gift as a fountain pen rather than a watch. Any of these would be errors, and would keep this from being an instance of successful remembering. Still, the representations are mostly accurate; the inaccuracies involved are distortions of the actual event. This form of misremembering is illustrated most clearly in the errors generated by the DRM paradigm in cognitive psychology (Roediger and McDermott 1995; see also Robins 2016).

There may also be other forms of misremembering, which stretch the original definition above. Misremembering could occur when there is a mismatch between the target and the content, as when the target of Jamari's representation is the graduation party, but the content is drawn from the opening of a gift at a different celebration, like a birthday. In such a case, accuracy is still the primary concern. The content generated is an accurate representation of some event in Jamari's past, but not the event he takes himself to be representing and remembering.

2.2 Relearning

Relearning errors occur when there is problem with the causal relationship between the content generated and the event being targeted. In cases of relearning, a person has an experience, forgets it, and then learns of it from somewhere else, and this later relearning is at some point confused for remembering. Suppose Jamari received a watch at his college graduate party, but forgot all about it, either through natural means or because of some kind of neurological trauma that produced amnesia. He later discovers a video of the party, including his opening of the watch. Over time, he forgets how he acquired this information about receiving a watch at his graduation and when he thinks about this party, he takes the mental representations he forms to be a memory of the party (rather than a memory of watching a video of the party). In such a case, the target of Jamari's representation is the graduation party, and the content of his mental representation accurately depicts that event.[9] The error is in the relationship between the target and the content. It seems to Jamari that he remembers his graduation party, but he does not. Relearning can be understood as a form of memorial hearsay, where information *about* an event or experience is misinterpreted as being *from* an experience.

2.3 Confabulation

Confabulation occurs when there is no relation between a person's feeling as if they remember a particular event/experience and any event or experience from their past—either because there is no such event in their past or because any correspondence to such an event is entirely coincidental. Where cases of relearning and misremembering involve some form of mismatch between the target and content, cases of confabulation involve no connection at all. In this way, confabulation is a memory error that is concerned most directly with the requirement that the targeted event must exist. In most cases of confabulation, it does not. Suppose Jamari never went to college, and has in fact spent most of his adult years living in a psychiatric facility. He could enter a state of seeming to remember receiving a watch at his college graduation party. This instance of seeming to remember would be a confabulation because there is no event in his personal past corresponding to the one he is targeting. This is the form of confabulation that has generated the most clinical and theoretical attention and interest. Many definitions of confabulation have even made the falsity of the memory an essential feature of confabulation. But, as others have noted, confabulations could be veridical. In these cases, the targeted event exists, but there is no connection between the target and the content. To see the point, we can return to the form of the example where Jamari graduates from college, and has a party afterward where he receives a watch. He then forgets about this experience, perhaps as the result of some amnesia-producing trauma. At some later point in life, saddened by the memories he has lost to amnesia, he begins inventing stories about his past. After frequent retelling, he begins to consider many of these stories genuine memories. One of these stories, as it happens, matches exactly his experience of receiving a watch at his college graduation party. The representation is an accurate depiction of an event from his personal past, but this does not come about because of any information Jamari has about the graduation party (either from his memory or a video), as Jamari has lost this information and has not reencountered it in any other context. There is no connection between the event and his representation of it. It is just an instance of serendipitous confabulation.

3 Mnemonic Confabulation

In the previous section, I offered an account of mnemonic confabulation, situated within a more general theory of remembering. Mnemonic confabulation occurs when there

[9] In standard cases of relearning, as originally introduced by Martin and Deutscher (1966), what is relearned is true or accurate of the event. The case is meant to illustrate the importance of causal history for remembering, and so relearning is meant to be a case where all other requirements (including accuracy) are met. Michaelian (2016a, b) has argued that taxonomies of memory errors should also accommodate falsidical relearning. Although I have not built it in explicitly, falsidical relearning is consistent with the account of remembering and its errors that I am developing here.

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. Some of the details in my presentation of mnemonic confabulation are particular to my own account of remembering, but the general description of mnemonic confabulation is shared by other philosophers of memory who have written about this phenomenon (Bernecker 2017; Michaelian 2016a, b). In what follows, I draw out three key points of consensus amongst philosophical account of mnemonic confabulation, setting the stage for a comparison of mnemonic and broad accounts of confabulation in Sect. 4.

First, philosophers of memory endorse process-based views of confabulation. We agree that confabulation is an error in the *process* by which a memory is generated. Although most confabulations are inaccurate, accuracy is not the defining feature. When the process of remembering has gone wrong, the result will often be an error—but the error is only a symptom, not the underlying issue. There are differences in how the process of remembering is understood amongst philosophers of memory, and so there are distinct accounts of how that process has failed in cases of confabulation. Causal theorists, like myself (Robins 2017a) and Bernecker (2017) argue that the process is causal, and confabulations are errors because they lack a causal connection between the event and its representation. For a simulation theorist like Michaelian (2016a, b), the process of remembering is one that derives from a reliable mechanism, and so confabulations are representations that come from a difference process—i.e., an unreliable mechanism.

Second, we agree that confabulation is a memory error that must be considered in theorizing about memory. Debates over how to taxonomize memory errors are relatively new to the philosophy of memory, and there are ongoing discussions about which errors to include. There are disagreements, for example, over whether to include relearning: Michaelian (2016b) includes it, but Bernecker (2017) argues that it should be left out. No similar disagreements have occurred over confabulation, and they are unlikely to arise. This is because everyone shares a commitment to the same, similar reason for including confabulation: *it's possible*. Developing a theory of memory requires an articulation of the process of remembering. This, in turn, requires recognition that the representations produced in remembering could come about in other ways (i.e., through a process that is not remembering). Consideration of this possibility thus serves as an important constraint on theorizing about memory. This is, in fact, how confabulation entered the discussion. In their landmark paper introducing the causal theory of memory, Martin and Deutscher (1966) began their account by considering a case of veridical confabulation. The causal theory was motivated by the need to distinguish such cases from genuine remembering. Even simulationists like Michaelian (2016a, b), who reject the causal theory, consider confabulation an important test case, as "the possibility of veridical confabulation will never vanish entirely" (Michaelian 2016b, fn 4).

This quotation highlights a third point of consensus about mnemonic confabulation amongst philosophers of memory: confabulations can be veridical. Veridical confabulations may be unlikely, as the highly contrived example involving Jamari's post-amnesia stories illustrates, and in fact there may never have been a veridical confabulation. But veridical confabulation is possible—it could happen—and the definition of mnemonic confabulation should therefore not rule it out. In previous work on confabulation, where I was focused on distinguishing this error from misremembering, I characterized confabulations as false, in comparison to misrememberings that are merely distorted (reference withheld). This was a mistake, as Michaelian (2016b) and Bernecker (2017) have helpfully pointed out. The account sketched in Sect. 2 corrects for this.[10]

Pausing to note these points of consensus about mnemonic confabulation is important because it reveals where interest in this phenomenon arose for philosophers of memory. As a mental/cognitive error, mnemonic confabulation is most analogous to hallucination in perception. This comparison can be found in all three of the recent accounts of confabulation in the philosophy of memory literature: Michaelian (2016b), Robins (2017b), and Bernecker (2017). To see the point, compare the above discussion of mnemonic confabulation to these remarks about hallucination in the *Stanford Encyclopedia of Philosophy* entry on the problem of perception:

> A hallucination is an experience which seems exactly like a veridical perception of an ordinary object but where there is no such object there to be perceived. Like illusions, hallucinations in this sense do not necessarily involve deception. And nor need they be like the real hallucinations suffered by the mentally ill, drug-users or alcoholics. They are rather supposed to be merely possible events: experiences which are indistinguishable for the subject from a genuine perception of an object. For example, suppose one is now having a veridical perception of a snow covered

[10] Michaelian (2016b) argues further that by allowing confabulation to be veridical, the causal account becomes unable to distinguish between confabulation and relearning. Bernecker (2017), for this reason, drops relearning from his taxonomy of memory errors. There isn't space in this paper for me to respond fully to either of these challenges. The account offered in Sect. 2 should make clear that my account can accommodate veridical confabulation and do so in a way that is distinct from relearning. A more complete defense of this position and the continued inclusion of relearning will have to wait for another paper.

> churchyard. The assumption that hallucinations are possible means that one could have an experience which is subjectively indistinguishable—that is, indistinguishable by the subject, "from the inside"—from a veridical perception of a snow covered churchyard, but where there is in fact no churchyard there to be perceived (Crane and French 2015).

Perceptual hallucination and mnemonic confabulation are both possible events, cases whose consideration prompts the inclusion of something beyond the felt experience of a mental state in an account of perception or memory. 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. The initiation of confabulation into clinical discussions of patients' symptomology also began with memory, in Korsakoff's (1889) and Wernicke's (1906) descriptions of the bizarre stories told by patients suffering from amnesia.[11] Contemporary interest in accounts of broad confabulation derives from this work, with an interest in expanding the original definition to accommodate a range of other cases. In spite of the shared focus on memory, mnemonic confabulation's initial motivations were distinct from those that initiated work on broad confabulation.

Of course, the lack of a shared history is not on its own enough to establish that mnemonic confabulation and broad confabulation are different. They could be two routes to identifying the same phenomenon. And some philosophers of memory have treated them as such: Bernecker (2017) defends his causal account of confabulation as not only a component of his theory of memory, but also as the account with the most clinical utility. But the difference in origin is enough to prompt the question of how mnemonic confabulation relates to broad confabulation, and I turn to this question in the next section.

4 Mnemonic Confabulation Versus Broad Confabulation

In the previous two sections, I developed an account of mnemonic confabulation. Here the focus turns to broad confabulation. As stated in the introduction, accounts of broad confabulation are attempts to unify the seemingly disparate features of all or most confabulatory phenomena under a shared set of characteristics or mechanisms. There are a number of views I consider to be accounts of broad confabulation: Hirstein's (2005) is the most extensive and one of the most prominent, but broad accounts can also be found in Carruthers (2005), Coltheart and colleagues (Coltheart 2017; Colheart and; Turner 2009; Turner and Coltheart 2010; Coltheart et al. 2010), as well as Strijbos and de Bruin (2015). I intend this list to be illustrative, not exhaustive. These accounts differ in many substantive ways, and my neglect of those differences here should not be taken as an indication of their perceived importance. For my purposes, what matters is the shared assumption from which these accounts derive—namely, that a range of delusional behaviors and false reports are similar enough to warrant a shared, systematic treatment. Hirstein (2005) illustrates this well in the opening of his book-length treatment of confabulation: "the apparent diversity of confabulation syndromes invites a search for something they have in common" (p. 3).

There is no established list of confabulatory phenomena, but standard accounts include the false memories that arise from neurocognitive damage, including both the often bizarre and fantastical "memories" reported by persons with Korsakoff's and schizophrenia, but also the more mundane (but still false) memory reports given by patients following an *aneurysm of the anterior communicating artery* (ACoA). Most accounts of broad confabulation also include the false memories produced by participants in certain psychology experiments, who outside of those experimental conditions have no recorded memory deficit or disorders (e.g., Loftus and Pickrell 1995). A range of non-memory cases are often included as well: both clinical delusions as arise from Anton's Syndrome and Capgras', as well as the rationalizing explanations of decision-making (Nisbett and Wilson 1977) and moral judgment (Haidt 2001) given by non-clinical participants in experimental settings.

It is an open question whether mnemonic confabulation, as I have characterized it in Sects. 2 and 3, is one of the phenomena that should be included in the set accommodated by broad confabulation. Despite differences between accounts of broad confabulation, the debates and discussions out of which such accounts have arisen have produced a set of general features critical to most portrayals of broad confabulation. In what follows, I identify three general features of broad confabulation and argue that, for each, mnemonic confabulation lacks this feature.

4.1 Role of Veridicality

In the initial clinical use of "confabulation," this term picked out memories and other behavioral reports that were *false*. Korsakoff, for example, relates the story of a patient who claimed that he had ridden a bike into town the day before, when in fact he had been hospitalized for months (1889/1955, p. 399). Many subsequent definitions

[11] An example from Korsakoff (1889/1955): "When asked to tell how he has been spending his time, the patient would very frequently relate a story altogether different from that which actually occurred, for example, he would tell that yesterday he took a bike ride into town, whereas in fact he has been in bed for two months, or he would tell of conversations which never occurred" (p. 399).

of confabulation have retained the emphasis on falsity (e.g., Talland 1961; Berlyne 1972), but more recent accounts of broad confabulation have lessened this emphasis, allowing for some limited role of veridicality in confabulation.

These initial remarks might lead the reader to suspect that I have identified a similarity between mnemonic and broad confabulation, rather than a difference. The acknowledgment that veridicality can play a role in confabulation is similar across accounts of mnemonic and broad confabulation, but the reasons *why* veridicality is included are importantly different.

Broad confabulation theorists offer a range of reasons for de-emphasizing falsity. Moscovitch and Melo (1997) move away from the insistence on falsity because they want to include as confabulations a range of cases where the reports are distorted, but not entirely false. For Hirstein (2005), the concern is about the role of luck. A patient with a malfunctioning memory system could, on occasion, produce a true report. As he explains, "A patient who gets a question right after supplying wrong answers to the previous six has not miraculously stopped confabulating. Confabulator's claims have a low probability of being true because of brain malfunction" (ibid, p. 199). In much the same way that a broken clock can tell the correct time if one happens to look at it at a particular moment, a confabulator's report can inadvertently be veridical.

These accounts allow veridicality by situating instances of confabulation within confabulators. That is to say, they are *system-based* accounts of confabulation. Confabulation is the result of a disordered or malfunctioning cognitive/neural mechanism. The determination of whether a particular response, judgment, or action is a confabulation is determined by the system from which it emanates. Confabulations come from malfunctioning or disordered systems. Given the condition of the system, most of its products will be false; but some could be true or only distorted.

Mnemonic confabulations can be the result of a malfunctioning system. A system-level malfunction would disrupt the memory-forming process and so there could be representations of seeming to remember that are produced by other processes (and so are confabulations). But importantly, mnemonic confabulations are not restricted to malfunctioning systems; they can also occur in otherwise healthy and well-functioning systems.[12] It's about the process of producing individual instances of seeming to remember. Even 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.

4.2 Ill-Groundedness

In place of falsity, most accounts of broad confabulation emphasize the ill-groundedness of confabulations, arguing that this is the critical feature shared by these phenomena. As Hirstein puts it, "If the confabulator's brain were functioning properly, she would know that the claim is ill-grounded and not make it" (Hirstein 2009, p. 652). Similarly, Turner and Coltheart (2010) focus on confabulations as *unsubstantiated reports*. This characterization is similar to Nisbett and Wilson's (1977) explanation of the results of their decision-making studies: participants in these studies lack access to the processes that guide their behavior, and as a result, generate "reasons" for acting that sound plausible, but have no evidential basis. Even though there are differences between particular accounts, recognition that something like ill-groundedness unifies all instances of confabulation has been an important and energizing claim for discussions of broad confabulation.

The shift of emphasis from falsity to ill-groundedness highlights an interesting feature of many if not all of the phenomena included in accounts of broad confabulation: they are, by and large, *explanations*. They are a person's attempt to explain why they have made a particular decision, judgment, or action. Consider the patient with Anton's Syndrome, who confabulates when asked why they keep running into things. Such patients are experiencing cortical blindness, and so often bump into objects and persons around them. When these accidents occur, these patients standardly attribute them to environmental conditions rather than their limited vision (Swartz and Brust 1984). They confabulate about the reason for their poor navigation. Something similar happens with the participants in the Nisbett and Wilson (1977) studies, who confabulate the reasons for choosing a particular item from an array. All of the items are exactly similar, but participants routinely select the rightmost item. When asked *why* they've selected that item, position is not one of the reasons that they give. Instead, participants discuss the quality or distinctiveness of the item—a confabulation, since all items share these allegedly distinctive features. Even the cases of memory confabulation that are standardly included are considered noteworthy not (only) because of the false memory report, but because of the additional rationalizations that are piled on to support the initial report. An often-noted example of this is a patient, who begins by reporting that he has been married for 4 months.

[12] Some of the phenomena standardly included in accounts of broad confabulation also occur in healthy and well-functioning systems—e.g., confabulations about decision-making (Nisbett and Wilson 1977). But these confabulations are still system level issues. On a standard interpretation of these results, they indicate the ways in which our cognitive systems are structured so as to limit introspective access to our reasons for acting and deciding. Thanks to an anonymous reviewer for raising this issue.

When pushed to explain how this is compatible with having four adult children, the patient claims they were adopted (Moscovitch 1989).

Cases of mnemonic confabulation are different. There is no explanation or need for justification. A person who seems to remember an event that never in fact occurred (a confabulation), may later try to justify this alleged memory or, recognizing its lack of justification, may reject it. This often happens when two people who experienced an event together are later reunited and reminiscing. If the contents of what each seems to remember are in conflict, then they may each try to defend the accuracy of how it seems to them. In non-clinical cases, they may recognize inconsistencies in these seemings (e.g., that the representation involves interactions between persons who were not alive at the same time) and downgrade or dismiss the representation. But mnemonic confabulation concerns the initial state of seeming to remember, not these later processes of evaluation and justification.

We can push the point of distinction between mnemonic confabulation and the ill-groundedness of broad confabulation further: cases of mnemonic confabulation are *well*-grounded. Philosophers of memory have not devoted substantial attention to what grounds or justifies the claim to remember some past event or experience, but it seems clear that the most essential feature is just that it seems to the person that they remember the event or experience. Debus (2010) claims that remembering requires treating the representation as *epistemically relevant*: making use of the remembered information in applicable contexts. Others claim that seeming to remember carries with it immunity to error through misidentification (Hamilton 2009; Fernandez 2017). Seeming to remember is then justified because one cannot misidentify themselves in such an instance. One needn't endorse either of these claims to accept the more minimal point being made here: insofar as it seems to you that you remember a particular event in your personal past, you have at least some grounds for thinking that you do in fact remember this event from your personal past (even if it turns out that you are mistaken).

It is worth noting that many of the phenomena that are standardly included in accounts of broad confabulation involve memory: patients with Korsakoff's, those experiencing amnesia following ACoA, and even non-clinical participants in Loftus' and colleagues' experiments. Accounts of broad confabulation framed around ill-groundedness may have difficulty accommodating such cases, if what I have said here is correct. Some, e.g., Hirstein (2005) get around this issue by focusing on the secondary confabulations in these cases: not the original false memory, but the continued endorsement and justification and elaboration on that memory, in light of overwhelming evidence to the contrary.

4.3 Possibility of Recovery

The benefits of developing an account of broad confabulation go beyond the intellectual merits of creating a rich, unified theory. A deeper understanding of what confabulation is, and how its various forms are related, also affords insight into how they might be treated, limited, or even eliminated. For many forms of confabulation, there is a route to recovery: even if the relevant information cannot be accessed in the standard way, there is often an indirect alternative available. As discussed in Sect. 4.2, confabulations are explanations, requests for the reasons and/or causes of one's action, decision, or behavior. The confabulations often come in response to why questions. For the patient with Anton's Syndrome, *why do you keep running in to the furniture?* For the participant in Nisbett and Wilson's (1977) studies, *why did you select that pair of stockings?* And so on. The confabulations occur, in large part,[13] because the confabulator lacks introspective access to the causes of her action. In many clinical cases of confabulation, the lack of access is the result of neurological trauma, whereby the patient loses access to the kind of information that is standardly available to persons who have not experienced such trauma. In the case of Anton's Syndrome patients, for example, their difficulties include both cortical blindness and awareness of this deficit. Outside of such a Syndrome, persons are able to recognize when they have trouble with their vision, and appeal to this information in explaining any subsequent changes in their behavior.

In cases of everyday confabulation (i.e., confabulation in psychological experiments amongst non-clinical participants), the lack of knowledge or awareness is thought to be a more general feature of human psychology. Consider the Nisbett and Wilson (1977) cases of decision-making, where participants select the rightmost item amongst a set of exactly similar items, but claim to do so for non-positional reasons. Nisbett and Wilson, and many others since, have argued that the processes by which these decisions are made are unavailable to introspection, and that as a result we are *strangers to ourselves* (the title of Wilson's 2002 book). Haidt (2001) makes a similar argument about the nature of our moral decision making.

Whether the barrier to self-knowledge is the result of neurological trauma or is a more permanent feature of human psychology, there are alternative routes through which one could acquire the relevant information and so avoid confabulation. Perhaps patients with Anton's Syndrome and similar

[13] Of course, this cannot be the entirety of the explanation for why confabulation occurs, because a lack of knowledge or awareness could also lead one to say "I don't know" rather than generate a false report.

disorders could be presented with frequent reminders of their deficit—a bracelet inscribed with the information, a sign posted in their bedroom, etc. For the more general limitations, one could gain the information by becoming a student of cognitive and social psychology (formally or informally). Learning about the nonconscious influences on our decisions and judgments would provide an alternative source of explanations. I am not arguing that any of these approaches to recovering from confabulation would be successful. The point I want to make is a more limited, conditional one: *if* the patient or participant could acquire the information about the true causes of their action/decision/judgment by some alternative means, then their responses would no longer be confabulations and there would be nothing delinquent about the answer given.

This is an important feature of broad confabulation to draw out because it presents yet another way in which mnemonic confabulation is distinct. In cases of mnemonic confabulation, the route by which the information is acquired matters. Successful remembering requires not only an accurate representation of the past event, but a representation that was produced in the right way. Suppose someone asks Jamari, from our examples in Sect. 2, what he was doing a year ago (and this happens to be when his graduation and the subsequent watch-receiving party occurred). In order for Jamari's answer to be an instance of genuine remembering, he has to not only produce an accurate representation, but the representation must be produced by a memory trace that he acquired from the graduation party and has retained ever since. If this memory is lost, either by processes of everyday forgetting or because of some more serious amnestic trauma, then it is no longer possible for him to remember this event. Once the connection characteristic of remembering is gone, there is no getting it back. He can reacquire the information from some alternative source—from a friend who attended the party, or from a video recording, or from notes he kept in a journal. But acquiring the information from an alternative source is not remembering; it's relearning. Mnemonic 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.

This difference between mnemonic and broad confabulation is especially important to note because, in recognition of the fact that some confabulations may be avoidable or eliminable, some have begun to asking about the psychological consequences of intervening into confabulations. As Bortolotti (2017) has argued, many everyday confabulations may help a person more (or at least as much as) than they hurt: even if they lack justification, confabulations could be valuable because of the ways in which they maintain the confabulator's sense of self and agency. This is an exciting and insightful new line of inquiry, and an important conversation to be having alongside the development of accounts of broad confabulation. It also makes clear why it's important to note the differences between mnemonic confabulation and broad confabulation and keep them distinct. As these conversations about the nature of broad confabulation, its costs, possible treatments, and even possible benefits, continues, it is helpful to keep the set of phenomena under discussion as streamlined as possible. Including mnemonic confabulation, which differs in many critical respects, muddies the waters unnecessarily.

4.4 Leaving Mnemonic Confabulation Out

The development of accounts of broad confabulation, and refinements to these accounts that have occurred as further views have been introduced and debated, has generated a loose consensus about the general features of confabulatory phenomena. As Hirstein (2005) puts it, all cases of confabulation involve "a sort of pathological certainty about ill-grounded thoughts and utterances" (p. 4). These confabulations are illustrative of brain deficits and cognitive malfunction, or a misunderstanding of our everyday introspective capacities. These disorders and limitations often produce errors, but they remain disordered (and confabulatory) even when they do not. They are unified by the way in which they make certain forms of self-knowledge and awareness unavailable, and so constrain the ability to offer adequate explanations of one's actions, judgments, and decisions. There are many details left to work out (and differences amongst accounts in terms of how these features are defined and implemented), but progress in these discussions continues at a steady clip.

In this section, I have argued that—for each of these key features of broad confabulation—mnemonic confabulation does not fit. Mnemonic confabulation can be veridical, as broad confabulation can, but the ways in which this is a possibility for mnemonic confabulation are different from other confabulatory phenomena because mnemonic confabulation is a *process*, rather than *system* error. Broad confabulatory phenomena are ill-grounded; mnemonic confabulations are not. For many broad confabulations, where there is a systemic lack of access to the relevant information, there is an alternative route by which the information can be acquired and further confabulation avoided. For mnemonic confabulation, there is no such route. Given these key differences, it seems progress in theorizing about broad confabulation and about mnemonic confabulation will best continue if the two are recognized as distinct.

These features highlight important differences between mnemonic confabulation and broad confabulation, differences that have been overlooked in efforts devoted to developing accounts of broad confabulation. Interest in each form of confabulation derives from memory failures, but

the feature of confabulation that is of primary interest is distinct. There may be some instances of confabulation that that would qualify as both mnemonic and broad. Consider, for example, a patient with Korsakoff's who "remembers" going on a weekend bike ride, when in fact he has been hospitalized for several months.[14] The initial report of the bicycle ride, because of its lack of connection to a past experience, would qualify as an instance of mnemonic confabulation. But the patient is generating such reports because of system-level deficits, and may continue to insist on this account of his weekend activities even after being reminded of his hospitalization, etc., and these latter elaborations look like key instances of broad confabulation. The interrelation of these two forms of confabulation is perhaps unsurprising, as the cognitive roles of remembering and explaining are integrated. But recognizing their relations and distinctness is important for future theorizing about confabulation.

5 Conclusion

Interest in confabulation, as a symptom of clinical disorders and as a facet of everyday cognition, began with Korsakoff (1889) and Wernicke's (1906) observations of bizarre false memory reports in patients with amnesia and dementia. Accounts of broad confabulation have grown over time to include other, non-memory phenomena, but confabulations in memory have always been central to the discussion. For this reason, it is easy to suppose that debates about the nature of mnemonic confabulation amongst philosophers of memory, are conversations about the same thing. The shared interest in memory and the shared use of the term "confabulation" are, however, nothing more than surface similarities between two substantively distinct debates. Accounts of broad confabulation are attempts to unify the ways in which our cognitive and neural systems can fail to give us access to information about ourselves, either as the result of disorder/malfunction or as the result of how are cognitive systems are structured more generally. Accounts of mnemonic confabulation are offered in acknowledgment of a possibility that constrains how the nature of remembering is characterized. The latter form of confabulation has more in common with perceptual hallucination than it does with the former. Mnemonic confabulation and perceptual hallucination are both merely possible events, the recognition of which shapes how theorizing about memory and perception, respectively. The use of "confabulation" in discussions of broad and mnemonic confabulation are both apt, and the ensuing discussions of each have been interesting and illuminating. The best way to ensure that this continues, on both fronts, is to recognize that the two phenomena are distinct.

Acknowledgements Many thanks to Sophie Stammers and Lisa Bortolotti for organizing the workshop on confabulation at St. Anne's College, Oxford, which inspired this paper—and to my fellow presenters and attendees for their helpful questions and comments. Earlier versions of this paper benefited greatly from feedback I received at Jordi Fernandez's *Philosophical Perspectives on Memory* workshop in Adelaide and from a workshop with Jonathan Schwenkler's Philosophy of Psychology seminar at Florida State University.

Compliance with Ethical Standards

Conflict of interest The author declares that she has no conflicts of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Berlyne N (1972) Confabulation. Br J Psychiatry 120:31–39
Bernecker S (2010) Memory: a philosophical study. Oxford University Press, Oxford
Bernecker S (2017) A causal theory of mnemonic confabulation. Front Psychol. https://doi.org/10.3389/fpsyg.2017.01207
Berrios GE (1998) Confabulations: a conceptual history. J Hist Neurosci 7:225–241
Bortolotti L (2017) Stranger than fiction: costs and benefits of everyday confabulation. Rev Philos Psychol 9(2):227–249
Bortolotti L, Cox R (2009) Faultless ignorance: strengths and limitations of epistemic definitions of confabulation. Conscious Cogn 18:952–965
Carruthers P (2005) Consciousness: essays from a higher-order perspective. Clarendon Press, Oxford
Coltheart M (2017) Confabulation and conversation. Cortex 87:62–68
Coltheart M, Turner M (2009) Confabulation and delusion. In: Hirstein W (ed) Confabulation: views from neuroscience, psychiatry, psychology, and philosophy. Oxford University Press, Oxford, pp 173–188
Coltheart M, Menzies P, Sutton J (2010) Abductive inference and delusional belief. Cogn Neuropsychiatr 19:261–287
Crane T, French C (2015) The problem of perception. In Zalta EN (ed) The Stanford encyclopedia of philosophy, Spring 2017 edn. https://plato.stanford.edu/archives/spr2017/entries/perception-problem/
Cummins R (1996) Representations, targets, and attitudes. MIT, Cambridge
De Brigard F (2014) Is memory for remembering? Recollection as a form of episodic hypothetical thinking. Synthese 191:155–185
Debus D (2010) Accounting for epistemic relevance: a new problem for the causal theory of memory. Am Philos Q 47:17–29

[14] This case was referenced earlier in Sect. 4.1.

Fernandez J (2017) The intentional objects of memory. In: Bernecker S, Michaelian K (eds) Routledge handbook of philosophy of memory. Routledge, London, pp 88–99
Fernandez J (2018) The functional character of memory. In: Michaelian K, Debus D, Perrin D (eds) New directions in the philosophy of memory. Routledge, London, pp 52–72
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834
Hamilton A (2009) Memory and self-consciousness: immunity to error through misidentification. Synthese 171:409–417
Hirstein W (2005) Brain fiction. MIT, Cambridge
Hirstein W (2009) The name and nature of confabulation. In: Symons J, Calvo P (eds) The Routledge companion to philosophy of psychology. Taylor & Francis, New York, pp 647–658
Korsakoff S (1889/1955) Psychic disturbance in conjunction with peripheral neuritis. Trans. M. Victor and P.I. Yakovlev. Neurology 5:394–406
Kurby CA, Zacks JM (2008) Segmentation in the perception and memory of events. Trends Cogn Sci 12:72–79
Loftus EF, Pickrell JE (1995) The formation of false memories. Psychiatr Ann 25:720–725
Martin CB, Deutscher M (1966) Remembering. Philos Rev 75:161–196
Michaelian K (2016a) Mental time travel: episodic memory and our knowledge of the personal past. MIT, Cambridge
Michaelian K (2016b) Confabulating, misremembering, relearning: the simulation theory of memory and unsuccessful remembering. Front Psychol 7:1857
Michaelian K, Robins S (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. Routledge, London, pp 13–32
Moscovitch M (1989) Confabulation and the frontal systems: strategic vs. associative retrieval in neuropsychological theories of memory. In: Roediger H, Craik FIM (eds) Varieties of memory and consciousness: essays in honor of Endel Tulving. Erlbaum Associates, Hillside, pp 133–160
Moscovitch M, Melo B (1997) Strategic retrieval and the frontal lobes: evidence from confabulation in amnesia. Neuropsychologia 35:1017–1034
Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231–259
Ramachandran VS (1996) What neurological syndromes can tell us about human nature: Some lessons from phantom limbs, Capgras syndrome, and anosognosia. Cold Spring Harb Symp Quant Biol 65:115–134
Robins SK (2016) Misremembering. Philos Psychol 29:432–447
Robins SK (2017a) Confabulation and constructive memory. Synthese. https://doi.org/10.1007/s11229-017-1315-1
Robins SK (2017b) Contiguity and the causal theory of memory. Can J Philos 47:1–19
Roediger HL, McDermott KB (1995) Creating false memories: remembering words that were not presented in lists. J Exp Psychol Learn Memory Cogn 21:803–814
Strijbos D, de Bruin L (2015) Self-interpretation as first-person mindshaping: implications for confabulation research. Ethical Theory Moral Pract 18:297–307
Swartz BE, Brust JC (1984) Anton's syndrome accompanying withdrawal hallucinations in a blind alcoholic. Neurology 34:969–973
Talland GA (1961) Confabulation in the Wernicke–Korsakoff syndrome. J Nerv Ment Dis 132:361–381
Turner M, Coltheart M (2010) Confabulation and delusion: a common monitoring framework. Cogn Neuropsychiatr 15:346–376
Wernicke C (1906) Grundriss der Psychiatrie, 2nd edn. Thieme, Leipzig