6 Towards a virtue-theoretic account of confabulation

Kourken Michaelian

6.1 Introduction

There is an ongoing debate among philosophers of memory over the nature of confabulation and related memory errors.[1] Confabulation can—very roughly—be defined as an error in which a subject who is unable to remember instead makes up an event, either by dislocating events in time or by fabricating events to fill in gaps in memory (Goodwin 1989, p. 65).[2] The representations that result from these processes are sometimes implausible, but they are sometimes perfectly plausible—at least when considered in isolation. Consider two cases discussed by Dalla Barba (2009). First, patient MB,

> while he was hospitalized, said on one occasion that he was looking forward to the end of the testing session because he had to go to the general store to buy some new clothes, since he hadn't been able to the day before, because he had gotten lost in the center of Paris, where he had fortunately met a nurse who kindly took him back to the hospital.
>
> (p. 227)

Though implausible when considered in relation to the subject's circumstances at the time, the events described by MB are not intrinsically implausible. Second, patient SD, when asked what he had done the day before, replied: 'Yesterday I won a running race and I have been awarded with a piece of meat which was put on my right knee' (Dalla Barba 2009, p. 227). Though highly implausible, the event described by SD was composed largely of elements drawn from his personal past, though not from the previous day, illustrating both the dislocation of events in time and the fabrication of events: SD, Dalla Barba tells us, 'was actually involved in running races', and '[i]t was actually during a running race in the mountains that he fell, sustaining a severe head trauma and an open wound to his right knee' (p. 227).

These cases illustrate two characteristic features of the phenomenon of confabulation. First, confabulations are typically *false*. The event described by SD did not occur and could not easily have occurred. The event described by MB could more easily have occurred, but as a matter of fact he had not gotten lost the day

DOI: 10.4324/9781003002277-10

before, and it was highly unlikely that he would go to the store after the session. Second, confabulation occurs not only in remembering the past but also in *imagining the future*. MB, like many of the patients described by Dalla Barba (2002, 2009), confabulates both with respect to his personal past and with respect to his personal future. The first of these features has played an important role in the ongoing confabulation debate; indeed, we will see that the importance of falsity has, if anything, been overestimated. The second feature, however, while unsurprising in light of the firmly established link between episodic memory and episodic future thought (see Michaelian 2016b), has so far played little role in the debate.

In this chapter, I review the confabulation debate, giving space to all available philosophical accounts of confabulation but making a case for the superiority of an updated, virtue-theoretic version of the simulationist account—based on the simulation theory of memory (Michaelian 2016b), which views episodic memory as a form of imagination distinguished from episodic future thought merely by its temporal orientation—over the rival false belief, causalist, and epistemic accounts, with an emphasis on the causalist account. Adopting a naturalistic outlook, the chapter takes for granted that confabulation is typified by clinical cases of the sort reported by Dalla Barba and that an adequate philosophical account of confabulation will be responsive primarily to the features of such cases; in other words, it takes for granted that an adequate philosophical account of confabulation will harmonise with the relevant empirical science.[3]

I discuss the false belief account, an early version of the causalist account, an early version of the simulationist account, an updated version of the causalist account, and the epistemic account in Sections 6.2–6.6. I then formulate an updated, virtue-theoretic version of the simulationist account in Section 6.7 and respond to the explanationist model of confabulation proposed in Bernecker's chapter in this volume in Section 6.8. I conclude by summing up the case in favour of the virtue-theoretic version of the simulationist account in Section 6.9.

6.2 The false belief account

As its name suggests, the false belief account of confabulation is inspired by the fact that confabulations are typically false or inaccurate. Not every false apparent memory is a confabulation, and false belief accounts (see Berrios 1998 for a review) differ with respect to which other features they take to be necessary for confabulation. But they have in common that they take confabulations to be false memories that are such that the subject is unaware of their falsity.[4] As Dalla Barba (2002) sees it, for example,

> [c]onfabulation is a symptom which is sometimes found in amnesic patients and consists in involuntary and unconscious production of 'false memories', that is the recollection of episodes, which never actually happened, or which occurred in a different temporal-spatial context to that being referred to by the patient.
>
> (p. 28)

The false belief account may work in practice, but it does not work in theory, simply because *veridical confabulation* is possible (Hirstein 2005; Robins 2016b). Take Dalla Barba's patient SD. Suppose that, as a matter of fact, SD really had won a race the day before and been awarded with a piece of meat on his right knee, but change nothing else about the case. Suppose, in particular, that there is no connection whatsoever between SD's experience of that event and his apparent memory of it. (The event that induced his amnesia might, for example, have occurred after the race.) SD's apparent memory is then true, but it nevertheless remains a confabulation.

Because the false belief account makes sense of the fact that confabulations are *typically* false by taking them to be *necessarily* false, it fails to accommodate the possibility of veridical confabulation and can thus be ruled out. Before moving on to the causalist account, however, it is worth pausing to ask whether the false belief account can acknowledge future-oriented confabulation. Though Dalla Barba himself calls attention to the existence of future-oriented confabulation, he does not quite come out and say that future-oriented confabulations are false episodic future thoughts. This is understandable, given the existence of disagreements regarding the truth-aptness of representations of future events. On many views in the metaphysics of time, the future is open. If the future is open, representations of future events would seem either not to have determinate truth values or to be systematically false. If they lack determinate truth values, the false belief account is straightforwardly inapplicable to future-oriented confabulation. If they are systematically false, the false belief account would seem to imply that all episodic future thoughts are confabulations.

Debates in the metaphysics of time notwithstanding, we ordinarily assume that at least many representations of the future have determinate truth values and are not systematically false. If we take that assumption for granted, the false belief account can in principle acknowledge future-oriented confabulation. It nevertheless fares no better with respect to confabulatory future thinking than it does with respect to confabulatory remembering, simply because, if veridical past-oriented confabulation is possible, then, assuming that representations of the future can be true, veridical future-oriented confabulation is possible.

6.3 The causalist account

The false belief account has been and remains influential in the empirical sciences of memory, but—perhaps because philosophers are used to considering such unlikely but theoretically important possibilities as veridical hallucination[5]—it has played little role in the current debate, which has unfolded essentially between partisans of causalist and simulationist accounts. Indeed, the debate was triggered by Robins's (2016a, 2019) proposal of a causalist account.

Inspired by the causal theory of memory (Martin & Deutscher 1966), Robins proposed a classification of confabulation and other forms of unsuccessful remembering based on two conditions (see Table 6.1). The first is an *accuracy* condition, which requires that the subject form an accurate representation of the

Table 6.1 Robins's (2016) causalist classification

Appropriate causation		*~ Appropriate causation*	
accuracy	*~ accuracy*	*accuracy*	*~ accuracy*
successful remembering	misremembering	relearning	confabulation

past event. The second is an *appropriate causation* or retention condition, which requires that the subject's representation be causally linked to his original experience of the event via the retrieval of stored information deriving from that experience. If both conditions are satisfied, the subject *successfully remembers* the represented event. If neither condition is satisfied, the subject *confabulates*. This initial causalist classification recognises two errors in addition to confabulation. If the accuracy condition is satisfied but the appropriate causation is not, the subject has *relearnt* the event. If the appropriate causation condition is satisfied but the accuracy condition is not, the subject is not remembering but *misremembering*.

The notion of relearning can be illustrated by a hypothetical case described by Martin and Deutscher. Suppose that a subject experiences an event, recounts it to a friend, loses all memory of it, is told about the event by the friend to whom he recounted it, loses all memory of being told, and then comes, under the influence of what he has been told, to entertain a representation that happens to be accurate with respect to the event in question. In this case, there is a causal connection between the subject's current representation of the event and his original experience of it, but the causal connection goes via another person and is therefore inappropriate. The notion of misremembering can be illustrated by the DRM effect, in which the subject is presented with a list of thematically related words (e.g., *hospital*, *sick*, *nurse*, etc.) and later recalls having seen a thematically consistent but nonpresented lure word (e.g., *doctor*) (Gallo 2010). It can also be illustrated by the misinformation effect, in which inaccurate post-event information is incorporated into the subject's memory for an event, resulting in retrieval of an inaccurate memory (e.g., the subject receives the suggestion and ends up remembering that there was a stop sign at the scene of an accident that he witnessed, when in fact he saw a yield sign) (Loftus 1996). In both the DRM effect and the misinformation effect, there is an appropriate causal connection between the subject's current representation of the event and his experience, but his representation is nevertheless inaccurate. Note that, unlike confabulation, misremembering is not a clinical but rather an everyday error: the DRM effect and the misinformation effect can be produced in the laboratory, but even in the laboratory, the subjects who display them are perfectly healthy and have intact memory systems, and the conditions that produce them in the laboratory are not dissimilar to conditions encountered in everyday life.

Because it calls attention to the existence of these additional errors,[6] Robins's causalist classification represents an important advance over the false belief account. Because it takes confabulations to be necessarily false, it nevertheless inherits that account's main problem: it fails to accommodate the possibility of veridical confabulation. By the same token, it fails to accommodate the possibility of *falsidical relearning*. The possibility of veridical confabulation was established earlier. The possibility of falsidical relearning is equally easy to establish. Take the subject in Martin and Deutscher's friend case. Suppose that the subject's experience of the event that he later recounts to a friend is entirely hallucinatory, but change nothing else about the case. Suppose, in particular, that there is a causal connection between the subject's current representation of the event and his original experience of it but that the causal connection goes via another person and is therefore inappropriate. The subject's apparent memory is then false, but if the original case is an instance of relearning, so is this variant of it.

We will see that the causalist account can be revised so as to enable it to accommodate falsidical relearning and veridical past-oriented confabulation. It cannot, however, be revised so as to acknowledge future-oriented confabulation (whether veridical or falsidical). The notion of future-oriented confabulation makes little sense if the defining feature of confabulation is the absence of an appropriate causal connection between the represented event and the subject's experience of it: since future experiences cannot cause present representations, it is trivial that episodic future thinking can never involve appropriate causation. The causalist would seem to have a choice between two strategies. First, he might classify all episodic future thinking as confabulatory. To opt for this strategy would be to stretch the category of confabulation beyond all recognition. Second, he might treat the application of the concept of confabulation to episodic future thought as a category mistake. To opt for this strategy would be to preserve the meaningfulness of the concept of confabulation as it applies to episodic memory at the cost of parting ways with the empirical science. Neither strategy is satisfactory.

6.4 The simulationist account

Before ruling the causalist account out entirely, we will consider an updated version of the account developed in part in response to the simulationist challenge. The present section discusses the simulationist account; the revised causalist account is discussed in the following section.

Drawing on the simulation theory of memory, I proposed a classification of errors based on two conditions (Michaelian 2016a; see Table 6.2). The first is an *accuracy* condition equivalent to Robins's. The second is a *reliability* condition, which requires that the subject's representation be produced by a properly functioning and hence reliable episodic construction system (where the episodic construction system is the system responsible for carrying out episodic remembering and episodic future thinking) that aims to produce a representation of an event from the subject's personal past. If both conditions are satisfied, the subject *successfully remembers* the represented event. If neither condition is satisfied, the subject *falsidically confabulates*. If the accuracy condition is satisfied but the reliability condition is not, the subject *veridically confabulates*. If the reliability condition is satisfied but the accuracy condition is not, the subject is not remembering but *misremembering*.

Table 6.2 Michaelian's (2016a) first simulationist classification

Reliability		*~ Reliability*	
accuracy	*~ accuracy*	*accuracy*	*~ accuracy*
successful remembering	misremembering	veridical confabulation	falsidical confabulation

This initial simulationist classification is designed to accommodate veridical confabulation and can be extended so as to acknowledge future-oriented confabulation. There are two approaches to defining a *future-oriented reliability* condition. The first approach is to say that, just as episodic remembering is reliable to the extent that it tends to produce accurate representations of events that occurred in the subject's personal past, episodic future thinking is reliable to the extent that it tends to produce accurate representations of events that *will* occur in the subject's personal future. There are two difficulties with this approach. On the one hand, it presupposes both that the future is determinate and that episodic future thinking is a matter of attempting to predict the future. The future may or may not be determinate, but it is clear that episodic future thinking often does not aim at predicting events that will in fact occur in the personal future but, more modestly, at producing representations of events that are *likely* to occur in the personal future. The second approach is to say that episodic future thinking is reliable to the extent that it tends to produce accurate representations of events that are likely to occur in the personal future. The relevant class of events—what Dalla Barba (2002) refers to as 'the probable possible'—will have to be specified more carefully before the future-oriented reliability condition can be fully spelled out, but if this can be done, then the simulationist will be able to treat *successful episodic future thinking* as occurring if both the accuracy condition and the reliability condition are satisfied, *'veridical' confabulatory future thinking* as occurring if the accuracy condition is satisfied but the reliability condition is not, *'falsidical' confabulatory future thinking* as occurring if neither the reliability nor the accuracy condition is satisfied, and *the future-oriented analogue of misremembering* as occurring if the reliability condition is satisfied but the accuracy condition is not. The latter error has not so far figured in the confabulation debate but would be worth investigating.

The obvious problem for this first simulationist classification is that it does not acknowledge relearning (whether veridical or falsidical). In order to acknowledge relearning, I proposed a second simulationist classification (Michaelian 2016a; see Table 6.3). The second classification incorporates an *internality* condition requiring that the subject himself contribute content to the retrieved apparent memory. If he does, then he is either (mis)remembering or (veridically or falsidically) confabulating, as before. If he does not, then he is either veridically or falsidically relearning.

There are two problems for this second simulationist classification. First, as Bernecker (2017) points out, relearning does not always amount to an error.

Table 6.3 Michaelian's (2016a) second simulationist classification

	Reliability		*~ Reliability*	
	accuracy	*~ accuracy*	*accuracy*	*~ accuracy*
internality	successful remembering	misremembering	veridical confabulation	falsidical confabulation
~ internality	veridical relearning	falsidical relearning	veridical relearning	falsidical relearning

Take Martin and Deutscher's friend case. If the subject takes his apparent memory to originate in his experience of the apparently remembered event, then he commits an error. But he need not take his apparent memory to originate in his experience of the apparently remembered event, and, if he does not do so, then the case need not involve error. Second, the causalist should arguably want to treat relearning as an error, simply because the appropriate causation condition is not satisfied in cases of relearning. The simulationist, in contrast, arguably should not: given that he rejects the appropriate causation condition, it is unclear what motivation the simulationist might have for not treating the satisfaction of the reliability and accuracy conditions as being sufficient for remembering. The first simulationist classification thus appears to be more adequate than the second.

6.5 A revised causalist account

In order to accommodate veridical confabulation, the causalist could in principle propose a classification analogous to the first simulationist classification (see Table 6.4).[7] The difference between this classification and the simulationist classification is simply that it replaces the reliability condition with the appropriate causation condition.

Given that he wants to count relearning as an error, however, the causalist will not be satisfied with this classification, which does not acknowledge relearning. Robins (2020) proposes a revised classification that takes into account not only appropriate causation but also the sort of inappropriate causation that figures in the friend case. As proposed by Robins, the classification does not accommodate the possibility of falsidical relearning, but it can easily be modified so as to do so (see Table 6.5). On this variant of Robins's classification, the subject

Table 6.4 A potential revised causalist classification

Appropriate causation		*~ Appropriate causation*	
accuracy	*~ accuracy*	*accuracy*	*~ accuracy*
successful remembering	misremembering	veridical confabulation	falsidical confabulation

Table 6.5 A variant of Robins's (2020) revised causalist classification

Causation				*~ Causation*	
appropriate causation		*~ appropriate causation*		*~ appropriate causation*	
accuracy	*~ accuracy*	*accuracy*	*~ accuracy*	*accuracy*	*~ accuracy*
successful remembering	misremembering	veridical relearning	falsidical relearning	veridical confabulation	falsidical confabulation

veridically or falsidically confabulates if the *causation* condition (and hence, trivially, the appropriate causation condition) is not satisfied. If the causation and appropriate causation conditions are satisfied, the subject successfully remembers or misremembers. If the causation condition is satisfied but the appropriate causation condition is not, the subject has veridically or falsidically relearnt.

This revised causalist account is a clear improvement over the original causalist account, but it faces a number of problems. We have already encountered some of these. First, it is unclear whether relearning should be treated as an error. Second, and more seriously, the account cannot be made to acknowledge future-oriented confabulation. The second of these problems, in particular, may already provide sufficient reason to rule the account out, but it faces three additional problems that have not so far been discussed.

The first problem concerns the status of confabulation as a clinical error. Robins takes confabulation to be exemplified by the sort of error at issue in suggestibility studies such as that reported by Loftus and Pickrell (1995). 'These studies', Robins writes, 'show that, as a result of mildly suggestive questioning, participants can come to "remember" events they never experienced, such as being lost in a shopping mall as a small child or having been hospitalized overnight' (2016a, p. 434). In a typical 'lost in the mall' (LITM) case, the subject is given inaccurate information to the effect that he was lost in the mall as a child and, under the influence of this information, comes to seem to remember being lost in the mall as a child. It is no surprise that Robins takes LITM apparent memories to be confabulations. Indeed, since there is, in these cases, no causal connection between the subject's apparent memory and his experience of the represented event, the causalist is bound to treat them as confabulations: he must treat a typical LITM case as an instance of falsidical confabulation and an LITM case in which the subject is given accurate rather than inaccurate information—such cases do not typically figure in the empirical literature but obviously might occur both in the laboratory and in everyday life—as an instance of veridical confabulation. The problem is that the error at issue in suggestibility studies differs fundamentally from what we have been referring to as confabulation. Confabulations, we have supposed, occur in clinical subjects suffering from amnesia and other malfunctions of the memory or episodic construction system. LITM apparent memories, in contrast, occur (primarily) in ordinary, healthy subjects. Confabulations, moreover, because they occur in subjects with malfunctioning memory systems, are most often false. LITM apparent memories, in contrast, may most often be true: assuming that, outside of the laboratory, others do not typically attempt to mislead us about our personal pasts,[8] and given that LITM apparent memories occur primarily in subjects with properly functioning memory systems, veridical LITM apparent memories may well be more frequent than falsidical LITM apparent memories.[9] It is thus highly misleading to apply the term 'confabulation' to both phenomena: the causalist is free to use the term however he wants, but there are two things here, not one, and our terminology ought to reflect that fact.

The second, related problem is that, because it treats LITM apparent memories as confabulations, causalism may have difficulty explaining why confabulations should tend to be false. The fact that there is no causal connection between a subject's apparent memory of an event and his original experience of it does not, of course, guarantee that the apparent memory is false. (Again, veridical confabulation is possible.) If, as suggested earlier, veridical LITM apparent memories are more frequent than falsidical LITM apparent memories, and if LITM apparent memories are sufficiently widespread, then, if LITM apparent memories are confabulations, veridical confabulation may be more frequent than falsidical confabulation. Causalism thus may have difficulty recognising the fact that confabulations are typically false.

The third problem, which can likewise be illustrated by means of LITM apparent memories, concerns the relationship between confabulation and relearning. On the revised causalist account, the difference between confabulation and relearning boils down to the presence or absence of a causal connection between the apparent memory and the subject's original experience of the apparently remembered event: in relearning, there is such a causal connection (though it is inappropriate); in confabulation, there is not. The problem is that this way of distinguishing between confabulation and relearning makes the distinction depend on irrelevant factors. Falsidical LITM apparent memories do not pose any difficulty. Such memories do not satisfy the accuracy condition. And because the events that they depict did not occur, it is trivial that the relevant subjects did not experience them and hence that they do not satisfy the causation condition. The causalist account will thus necessarily classify them as instances of falsidical confabulation. But consider veridical LITM apparent memories. Such memories do satisfy the accuracy condition. But though it is natural to assume (as we did in the previous paragraph) that they do not satisfy the causation condition, this is not necessarily the case. We must distinguish here between two possibilities.

First, take a standard non-laboratory veridical LITM case in which the subject's parents give him accurate information to the effect that he was lost in the mall as a child and, under the influence of this information, he comes to seem to remember being lost in the mall as a child. Suppose that the information provided by the parents derives entirely from their own experience of the event. There is then no causal connection between the subject's experience of being lost in the mall and the information provided by his parents, and hence there is no causal connection between the subject's experience of being lost in the mall and his apparent memory of being lost in the mall. The causalist account will thus classify the case as an instance of veridical confabulation.

Second, take the same non-laboratory veridical LITM case. But now suppose that the information provided by the parents is based in part on what the subject told them about the event before childhood amnesia took hold and he forgot it. There is now a causal connection between the subject's experience of being lost in the mall and the information provided by his parents, and hence there is a causal connection between the subject's experience of being lost in the mall and his apparent memory of being lost in the mall. *The causalist account will thus classify the case as an instance not of veridical confabulation but rather of veridical relearning.*

The account therefore implies that an apparent memory can be turned from an instance of confabulation into an instance of relearning by altering factors that have nothing to do with the operation of the subject's own memory system. In both of the scenarios just described, the subject forgets an event and later forms a representation of it on the basis of information provided by his parents. The only difference between them is that, in the second scenario, he happened to tells his parents about the event before forgetting it. The causalist account thus makes the status of an apparent memory, as an instance of confabulation or an instance of relearning, depend not on how the subject's memory system operates in the present or even on how it operated in the past, but rather on how someone else learned about the event. Note that the point generalises: any instance of confabulation can be turned into an instance of relearning by performing an appropriate alteration to the causal chain, and vice versa. (Martin and Deutscher's friend case, for example, can be turned into a case of confabulation merely by supposing that the basis for the information provided to the subject by his friend was not the subject's experience of the event but the friend's own experience of the event.) This way of classifying memory errors is clearly at odds with the understanding of confabulation at work in the relevant empirical fields.

If the causalist must classify LITM cases as instances of confabulation or relearning, it would seem that the simulationist is bound to say that, as long as the episodic construction system operates reliably in producing an LITM apparent memory, then the subject is either misremembering or successfully remembering. Indeed, I have argued (Michaelian 2016b) that what goes wrong in a standard LITM case is not that the subject fails to remember but rather that he misremembers, drawing on information received from others to construct a representation of an event that did not actually occur. By the same token, it would seem that the simulationist should argue that nothing goes wrong in a veridical LITM case: given that, in such a case, the subject draws on information received from others to construct a representation of an event that did in fact occur, simulationism would seem to imply that he simply remembers.

6.6 The epistemic account

We will see below that the implications of simulationism for LITM cases are not quite so straightforward. Before turning to this matter, a brief discussion of the relationship between the simulationist account of confabulation and the epistemic account proposed by Hirstein (2005) is in order, as the latter is similar in certain respects to the simulationist account. Two points about the relationship between the accounts should be noted.

First, the accounts are similar in that both emphasise the role of metacognitive failure in confabulation. Hirstein defines confabulation, roughly, as ill-grounded belief that the subject ought to but does not know is ill-grounded.[10] Similarly, the revised simulationist account reviewed in the next section (Michaelian 2020) treats confabulation as involving unreliability both at the object level (the retrieval process itself) and at the meta level (metacognitive monitoring for unreliability in the retrieval process). While full-blown confabulation arguably involves some form of metacognitive failure (cf. Schnider 2018), there is insufficient space here to take this aspect of confabulation into account, and it will be set aside in what follows.

Second, the accounts are similar in that both appear to be epistemic accounts. Hirstein defines confabulation in terms of ill-groundedness, a notion closely related to that of unjustifiedness, and he notes that his account of confabulation may be compatible with reliabilist analyses of justification (Goldman 1979). I define confabulation directly in terms of unreliability. Bernecker (2017) thus groups Hirstein's and my accounts together, treating both as epistemic accounts. This is, however, a mistake. The fact that reliabilists employ the concept of reliability in their analysis of justification does not imply that reliability is itself an epistemic concept, any more than the fact that utilitarians employ the concept of happiness in their analysis of moral rightness implies that happiness is a moral concept. One is free to make use of the concept of happiness while rejecting utilitarianism, and one is free to make use of the concept of reliability while rejecting reliabilism. The simulationist account of confabulation, in other words, may be compatible with reliabilism, but it does not entail it. Thus, while Hirstein's account is an epistemic account, mine is not. Whether this represents an advantage for my account will depend on the empirical respectability of accounts that employ (epistemic, ethical, or other) normative vocabulary. Standard forms of naturalism suggest that an adequate account must not employ such vocabulary, but we will not explore this question any further.

6.7 A revised simulationist account

Misremembering and veridical confabulation would appear to involve a form of *luck*: in misremembering, a reliable retrieval process happens by chance to produce an inaccurate representation, whereas in veridical confabulation, an unreliable retrieval process happens to produce an accurate representation (see Table 6.6). The failure involved in misremembering is thus not attributable to the subject, just as the success involved in veridical confabulation is not attributable to him.

Though I previously (Michaelian 2020) treated the form of luck involved in misremembering and veridical confabulation as exhausting the extent of luck involved in attempted remembering, I have more recently (Michaelian 2021)

Table 6.6 The simulationist classification (grey cells indicate luck)

Reliability		*~ Reliability*	
accuracy	*~ accuracy*	*accuracy*	*~ accuracy*
successful remembering	misremembering	veridical confabulation	falsidical confabulation

argued that a distinct form of luck is involved in certain LITM cases. Consider, on the one hand, falsidical LITM cases. In such cases, the subject forms an apparent memory on the basis of inaccurate information received from an external source. This need not (as noted earlier) result in unreliability, and assuming that the retrieval process is reliable, the simulationist account will treat these cases as instances of misremembering. Consider, on the other hand, veridical LITM cases. In such cases, the subject forms an apparent memory on the basis of accurate information received from an external source. Assuming that the retrieval process is reliable, the simulationist might treat these cases as instances of successful remembering (Michaelian 2016b). He ought, however, to distinguish between two kinds of veridical LITM case. In *non-lucky* veridical LITM cases, the external source (e.g., a family member) intends to provide accurate information and does so. There is no luck at work in such cases, and it therefore makes sense for the simulationist to treat them as instances of successful remembering. In *lucky* veridical LITM cases, the external source (e.g., an experimenter) intends to provide inaccurate information but inadvertently provides accurate information. There is a form of luck at work in such cases, and it therefore makes sense for the simulationist to treat them as instances of unsuccessful remembering. The form of luck at work in these cases, however, differs from that at work in misremembering and veridical confabulation.

Misremembering and veridical confabulation involve a single 'layer' of luck: a reliable process happens to produce an inaccurate representation (bad luck), or an unreliable process happens to produce an accurate representation (good luck). Lucky veridical LITM cases, in contrast, involve two layers of luck: first, the subject is, for example, the victim of experimenters seeking to implant in him a false memory of being lost in the mall as a child (bad luck); second, unbeknownst to them, he happens in fact to have been lost in the mall as a child (good luck). This two-layer structure recalls the structure of the Gettier cases that demonstrate the inadequacy of the "justified true belief" analysis of knowledge (Zagzebski 1994). Suppose, for example, that a subject truly believes that it is 9:00. Suppose that the subject formed this belief by looking at a clock that has always kept good time and believing what it indicated. His truly believing that it is 9:00 may nevertheless be due to luck: the clock may have stopped at 9:00 the day before (bad luck) and the subject happened to look at it at precisely 9:00 today (good luck).

The analogy between the form of luck at work in lucky veridical LITM cases and that at work in Gettier cases suggests looking to epistemology for clues as to how to handle the latter.[11] In developing a form of virtue reliabilism designed to cope with Gettier cases, Sosa (2007) observes that what goes wrong in such cases is that, while the subject's belief is true and is formed by a reliable process, it is not true *because* it is formed by a reliable process—it is, instead, true due to luck. Taking this observation as my starting point, I propose a virtue-theoretic version of the simulation theory and a virtue-theoretic version of the simulationist classification by introducing an *accuracy-because-reliability* condition, a condition requiring that the apparent memory be accurate because it was produced by a reliable process (Michaelian 2021). This makes room for the form of 'bad luck cancelled out by good luck' involved in lucky veridical LITM cases. I similarly introduce an *inaccuracy-because-unreliability* condition, a condition requiring that the apparent memory be inaccurate because it was produced by an unreliable process, making room for an analogous form of 'good luck cancelled out by bad luck' (see Table 6.7).[12]

On the resulting classification, successful remembering occurs when the accuracy, reliability, and accuracy-because-reliability conditions are satisfied. Misremembering occurs when the reliability condition is satisfied but the accuracy condition (and hence, trivially, the accuracy-because-reliability condition) is not. Lucky LITM cases occur when the accuracy and reliability conditions are satisfied but the accuracy-because-reliability condition is not. Falsidical confabulation occurs when the accuracy and reliability conditions are not satisfied and the inaccuracy-because-unreliability condition is satisfied. Veridical confabulation occurs when the accuracy condition is satisfied and the reliability condition (and hence, trivially, the inaccuracy-because-unreliability condition) is not satisfied. The nature of the sort of error that occurs when the accuracy and reliability conditions are not satisfied and the inaccuracy-because-unreliability condition is not satisfied is not immediately obvious but merits further investigation.[13]

6.8 The explanationist model

In his contribution to this volume, Bernecker argues that an explanationist model of memory may enable us to 'overcome the impasse' between causalist and simulationist approaches to memory error.

Table 6.7 Michaelian's (2021) revised simulationist classification (grey cells indicate luck)

Reliability			*~ Reliability*		
accuracy		*~ accuracy*	*accuracy*	*~ accuracy*	
accuracy b/c reliability	*~ (accuracy b/c reliability)*			*~ (~ accuracy b/c ~ reliability)*	*~ accuracy b/c ~ reliability*
successful remembering	lucky veridical lost in the mall	misremembering	veridical confabulation		falsidical confabulation

The explanationist model is motivated by two supposed problems. The first is what Bernecker refers to as the 'bootstrapping' problem: 'the criteria used to determine whether a given case qualifies as confabulation', he writes, 'rely on the very theory of confabulation, which the case is supposed to provide evidence for'. The suggestion here is that, because confabulation is a technical concept, we cannot rely on folk intuition in order to determine whether a given case ought to be categorised as an instance of confabulation. This leads to reliance on our favoured theories and thus to circularity: causalists have causalist intuitions and therefore categorise cases in accord with the causalist approach, and these classifications are then used as evidence in favour of causalism; and likewise for simulationism. For example, Bernecker suggests, 'we already have to assume the simulationist approach for the characterisation of an unrealistic future expectation as an instance of *mnemonic* confabulation to make sense' (Bernecker, Chapter 5, this volume). This is, however, not true. The simulationist approach has played no role in the empirical literature on confabulation, but the existence of future-oriented confabulation, which arises under the same conditions as past-oriented confabulation and appears to be due to the same mechanisms as the latter, is nevertheless widely recognised in that literature. We are thus not invariably bound, when asking whether a given case is an instance of confabulation or another kind of memory error, to rely on intuition. Instead, we often have independent purchase on the kind of error in question, enabling us to categorise specific cases without relying on intuition. These categorisations, in turn, can serve as independent evidence in favour of or against causalist and simulationist classifications, allowing us to avoid circularity.

The second problem is what Bernecker refers to as the 'red herring' problem: 'the debate about confabulation', he writes, 'is a proxy battle between the two leading accounts of memory—causalism and reliabilism [i.e., simulationism]. As soon as the controversy between causalism and reliabilism is (re)solved, the dispute about the individuation of confabulation comes to an end' (Bernecker, Chapter 5, this volume). The correspondence between theories of remembering, on the one hand, and classifications of memory errors, on the other, is, however, not as close as the 'proxy war' metaphor suggests. A theory of remembering does, of course, amount to an account of successful remembering, but it does not determine an account of memory error. As we have seen, multiple causalist and multiple simulationist accounts have been proposed. In general, multiple accounts will be compatible with a given theory. A given account of memory error may, however, rule out certain versions of the corresponding theory. For example, the virtue-theoretic account described in Section 6.7 implies that the original version of the simulation theory of memory (Michaelian 2016b) must be replaced with a virtue theory (Michaelian 2021). Far from being a mere proxy war, then, the confabulation debate may provide a means of making real progress in the ongoing dispute between causalism and simulationism.

The motivation for explanationism is thus lacking. That the motivation for the view is lacking does not, of course, imply that explanationism may not nevertheless shed light on the relationship between causalist and simulationist approaches to confabulation. Ultimately, however, it fails to do so. Bernecker takes as his starting point the observation that both causalists and simulationists classify many of the same cases as instances of successful remembering and confabulation. This suggests that they may have something in common—a shared core. Explanationism, the view that 'remembering amounts to memorially representing something because it is true' (Bernecker, Chapter 5, this volume), is meant to capture this shared core, which Bernecker takes to be a matter of ruling out luck (the coincidental correspondence of a present apparent memory to a past experience). But, while causalists and simulationists may agree on the need to rule out luck, they disagree about the nature of the luck that needs to be ruled out. They thus disagree about how to classify many cases. This should come as no surprise: because reliability is possible without appropriate causation, the theories are fundamentally opposed to each other—they lack a shared core.[14] Cases of future-oriented confabulation provide one example. Cases of LITM memory provide another: standard non-lucky veridical LITM memories are treated by simulationism, but not by causalism, as instances of successful memory. The examples are harder cases than those that Bernecker has in mind, but this does not make them any less important. It is no surprise that causalism and simulationism agree about the easy cases; the causalist–simulationist dispute will undoubtedly be decided—as philosophical disputes usually are—on the terrain constituted by the hard cases.

6.9 Conclusion

Overall, the simulationist account, particularly in its virtue-theoretic version, is in better shape than its rivals. The foregoing discussion has identified problems for the false belief, causalist, and epistemic accounts. *The false belief account* can be ruled out simply because it cannot accommodate the possibility of veridical confabulation. *The causalist account* faces a number of serious problems. It treats relearning as an error, though it appears that many cases of relearning are not errors. It is unable to acknowledge future-oriented confabulation. In addition, it does not treat confabulation as a clinical error, has difficulty explaining why confabulations tend to be false, and is committed to an implausible view of the distinction between confabulation and relearning. It remains to be seen whether the account can be modified so as to handle luck, including the form of luck involved in lucky veridical LITM cases, but it is by no means obvious what a suitably modified account would look like. *The epistemic account* shares some of the virtues of the simulationist account but has the disadvantage of being of doubtful empirical respectability. The simulationist account thus appears—at least at present—to be our best bet.

Even if the simulationist account is on the right track, however, it is not, in the version developed here, fully satisfactory. To see this, note that, while it classifies both falsidical LITM memories and DRM memories as instances of misremembering, there is an important difference between these, in that, while falsidical LITM memories are wholly false, DRM memories are false in detail.

The simulationist account is not alone in failing to acknowledge this difference. Indeed, despite the fact that the distinction between misremembering and confabulation is naturally taken to be intimately related to the distinction between memories that are wholly false and memories that are false in detail, the latter distinction plays a role in none of the accounts of confabulation and related errors that have so far been proposed in the literature. Determining the appropriate role for that distinction will be an important task for partisans of the simulationist, causalist, and epistemic accounts as the literature continues to develop.

In addition to this shared worry, the simulationist account faces a worry that may not be faced by its rivals. As noted at the outset, the motivation for the simulationist account is naturalistic in character. It is natural, from a naturalistic standpoint, to suppose that the factors that determine whether an apparent memory is an instance of error and, if so, of what kind of error it is an instance pertaining to the operation of the memory or episodic construction system; 'external' factors are irrelevant. The involvement of luck—which, according to the virtue-theoretic version of the simulationist classification, makes the difference between veridical LITM memories that qualify as successful memories and veridical LITM memories that do not—would, however, appear to be precisely such an external factor. The virtue-theoretic version of the simulationist classification thus appears to be at odds with the naturalistic motivation for the simulationist account. While I grant that the introduction of the notion of luck represents a departure from previous versions of the simulationist account, I suggest that it is not incompatible with a naturalistic approach to memory error. Luck plays a role in determining success and failure in many domains other than remembering, and there is, on the face of it, nothing to prevent a naturalist from acknowledging this. If so, then the naturalist can presumably likewise acknowledge a role for luck in determining whether remembering, in particular, is successful or unsuccessful. Further work will, however, need to be done both to articulate the worry and to determine whether the suggested line of response is viable.

Acknowledgements

Thanks to the participants in the Centre for Philosophy of Memory's internal seminar and in the Current Controversies in Philosophy of Memory online conference for feedback on earlier versions of this chapter. Thanks also to two reviewers for helpful reports. This work is supported by the French National Research Agency in the framework of the 'Investissements d'avenir' program (ANR-15-IDEX-02) and by CAPES-COFECUB (grant Sh 967/20).

Notes

[1] Confabulation is sometimes defined broadly, so that it includes both mnemic and nonmnemic errors. For example, Hirstein's (2005) epistemic account of confabulation (discussed in this chapter) is meant to apply to nonmnemic as well as mnemic confabulation. This chapter is concerned exclusively with mnemic confabulation.

[2] Definitions of confabulation in the psychological literature (see Berrios 1998 for a survey) often refer both to temporal displacement and to gap-filling. These two mechanisms would, however, appear to be distinct, and confabulations resulting exclusively from the former might differ in interesting ways from confabulations resulting exclusively from the latter. This possibility has not so far been considered in the philosophical literature but would be worth investigating.

[3] Robins, whose work was responsible for launching the confabulation debate and is discussed in detail later in this chapter, explicitly rejects this clinical conception of confabulation (Robins 2020). There is insufficient space here for a response to Robins's argument against the clinical conception.

[4] The fact that confabulators are typically unaware of their confabulations is emphasised by the epistemic and revised simulationist accounts of confabulation discussed in this chapter but will not be discussed here in any detail.

[5] On the comparison of confabulation to hallucination, see Robins (2020).

[6] In fact, it is not entirely clear that relearning should be treated as an error. We come back to this point later.

[7] Bernecker (2017) may have roughly such a classification in mind.

[8] This is, of course, an empirical claim; see Michaelian (2013) for a defence.

[9] In fact, some veridical LITM apparent memories may not amount to errors at all; see Section 6.7.

[10] Hirstein's definition, which is meant to apply to nonmnemic as well as mnemic forms of confabulation, reads in full:

S confabulates if and only if:

> *S* claims that *p*;
> *S* believes that *p*;
> *S*'s thought that *p* is ill-grounded;
> *S* does not know that her thought is ill-grounded;
> *S* should know that her thought is ill-grounded;
> *S* is confident that *p*. (2015: 187)

This rich definition merits a more detailed discussion than can be provided here.

[11] Note that, though the revised simulationist account draws inspiration from epistemology, it employs no epistemic concepts and is no more an epistemic account than was the original simulationist account.

[12] Like the classification proposed in Michaelian (2020), that proposed in Michaelian (2021) acknowledges the possibility of meta-level error; again, this will be set aside here.

[13] One might suppose that, if lucky veridical LITM cases occur when the accuracy-because-reliability condition is not satisfied, lucky falsidical LITM cases occur when the inaccuracy-because-unreliability condition is not satisfied, but this is not right. An unlucky falsidical LITM case would be an LITM case in which the external source intends to provide accurate information but inadvertently provides inaccurate information. Such a case involves only one layer of (bad) luck, amounting to misremembering.

[14] For an argument for the view that reliability presupposes appropriate causation, see Werning (2020).

References

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

Berrios, G. E. (1998). Confabulations: A conceptual history. *Journal of the History of the Neurosciences*, 7(3), 225–241.

Dalla Barba, G. (2002). *Memory, Consciousness, and Temporality*. Springer.

Dalla Barba, G. (2009). Temporal consciousness and confabulation: Escape from unconscious explanatory idols. In W. Hirstein (Ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy* (pp. 223–260). Oxford University Press.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848.

Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and Knowledge* (pp. 1–23). Springer.

Goodwin, D. M. (1989). *A Dictionary of Neuropsychology*, Springer.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation*. MIT Press.

Loftus, E. F. (1996). *Eyewitness Testimony* (2nd ed.). Harvard University Press.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725.

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

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

Michaelian, K. (2016a). Confabulating, misremembering, relearning: The simulation theory of memory and unsuccessful remembering. *Frontiers in Psychology*, 7, 1857.

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

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

Michaelian, K. (2021). Imagining the past reliably and unreliably: Towards a virtue theory of memory. *Synthese*, *199*, 7477–7507.

Robins, S. K. (2016a). Misremembering. *Philosophical Psychology*, *29*(3), 432–447.

Robins, S. K. (2016b). Representing the past: Memory traces and the causal theory of memory. *Philosophical Studies*, *173*(11), 2993–3013.

Robins, S. K. (2019). Confabulation and constructive memory. *Synthese*, *196*(6), 2135–2151.

Robins, S. K. (2020). Mnemonic confabulation. *Topoi*, *39*(1), 121–132.

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

Sosa, E. (2007). *Apt Belief and Reflective Knowledge. Vol I: A Virtue Epistemology*. Oxford University Press.

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.

Zagzebski, L. (1994). The inescapability of Gettier problems. *The Philosophical Quarterly*, *44*(174), 65–73.