Topoi (2020) 39:133–148
https://doi.org/10.1007/s11245-018-9591-z

Confabulating as Unreliable Imagining: In Defence of the Simulationist Account of Unsuccessful Remembering

Kourken Michaelian[1]

Published online: 15 October 2018

Abstract
This paper responds to Bernecker's (Front Psychol 8:1207, 2017) attack on Michaelian's (Front Psychol 7:1857, 2016a) simulationist account of confabulation, as well as his defence of the causalist account of confabulation (Robins, Philos Psychol 29(3):432–447, 2016a) against Michaelian's attack on it. The paper first argues that the simulationist account survives Bernecker's attack, which takes the form of arguments from the possibility of unjustified memory and justified confabulation, unscathed. It then concedes that Bernecker's defence of the causalist account against Michaelian's attack, which takes the form of arguments from the possibility of veridical confabulation and falsidical relearning, is partly successful. This concession points the way, however, to a revised simulationist account that highlights the role played by failures of metacognitive monitoring in confabulation and that provides a means of distinguishing between "epistemically innocent" (Bortolotti, Conscious Cogn 33:490–499, 2015) and "epistemically culpable" memory errors. Finally, the paper responds to discussions by Robins (Synthese 1–17, 2018) and Bernecker (Front Psychol 8:1207, 2017) of the role played by the concept of reliability in Michaelian's approach, offering further considerations in support of simulationism.

Keywords Confabulation · Episodic memory · Causal theory of memory · Simulation theory of memory · Epistemic innocence

1 The Simulation Theory Versus the Causal Theory

Bernecker's (2017) attack on the simulationist account of confabulation (Michaelian 2016a) takes place against the background of a larger project (see Bernecker 2008, 2010) devoted to developing and defending the causal theory of memory (Martin and Deutscher 1966). According to the causal theory, the difference between remembering a past event and merely imagining it is marked by the presence, in the case of remembering, of an appropriate causal connection between the subject's current representation of the event—his apparent memory—and his earlier experience of the event.[1] The simulationist account of confabulation is itself an application of the rival simulation theory of memory (Michaelian 2016b). According to the simulation theory, the presence of an appropriate causal connection of the sort singled out by the causal theory is not in fact necessary for the occurrence of remembering. Appealing to research on memory as mental time travel (see Perrin and Michaelian 2017), the simulation theorist argues that memory is merely one kind of imagination among others. Episodic memory (memory for experienced past events) is distinguished from episodic future thought (Szpunar 2010) by its temporal orientation (past vs. future) and from episodic counterfactual thought (De Brigard 2014) by its modal orientation (actual vs. counterfactual), but it is carried out by the same "episodic construction system" that enables us to imagine future and counterfactual events and shares the fundamental features of those processes (other than temporal and modal orientation). In particular, just as imagining a future or counterfactual event does not presuppose the existence of a causal connection between the current representation of the event and a future or conterfactual experience thereof (since the event has not been experienced), remembering a past event

✉ Kourken Michaelian
michaelian.kourken@gmail.com

[1] Laboratoire PPL, Université Grenoble-Alpes, Bat. ARSH, CS 40700, 38058 Grenoble Cedex, France

[1] An "appropriate" causal connection is typically understood as one underwritten by a memory trace originating in the relevant experience, but Bernecker, in particular, requires further that the current representation counterfactually depend on the experience; see Sect. 4.

does not (even when the event has been experienced) presuppose the existence of such a connection.

One of the main goals of a philosophical theory of memory is to provide a positive characterization of the nature of remembering that captures the difference (if there is one) between remembering the past and merely imagining it. One of the main goals of a philosophical account of confabulation is to describe, preferably on the basis of such a positive characterization of the nature of remembering, the difference between remembering and confabulating—as well, perhaps, as other forms of unsuccessful remembering.[2] The causal theory and the simulation theory thus each serve as the basis for an account of confabulation. The causalist account (Robins 2016a) distinguishes between confabulating and remembering in terms of the absence, in the case of confabulating, of an appropriate causal connection between the apparent memories in which the process results and the corresponding earlier experiences.[3] The simulationist account (Michaelian 2016a) is, according to Bernecker, to be grouped with epistemic accounts, such as that developed by Hirstein (2005), which distinguish between confabulating and remembering in terms of the unjustifiedness, in the case of confabulating, of the apparent memories in which the process results.

In addition to the causalist and epistemic accounts, there is what Bernecker refers to as the false belief account (e.g., Dalla Barba 2002), which distinguishes between confabulating and remembering in terms of the inaccuracy, in the case of confabulating, of the apparent memories in which the process results. The false belief account is unlike the causalist and simulationist accounts in that it is not linked to a positive characterization of the nature of remembering. It is also, despite the fact that it is suggested by a number of definitions given in the empirical literature, straightforwardly ruled out by the possibility of veridical confabulation.[4] Now, all accounts are bound to acknowledge that confabulations are, as a matter of empirical fact, inaccurate more often than not, and the false belief account may thus be "good enough" for most clinical purposes. Nevertheless, because confabulation can in principle result in accurate apparent memories, it would, no matter how rarely this occurs in practice, be a mistake to make to treat inaccuracy as a necessary condition for confabulation. The false belief account will accordingly be set aside in what follows.

2 The Causalist Attack on the Simulationist Account

The contest, then, is between causal accounts and epistemic accounts. Note that, in the course of his critique of the latter, Bernecker challenges certain details of Hirstein's account in particular. Since the focus here is specifically on the simulationist account, these details can be disregarded, and the discussion will focus on the two phenomena that, Bernecker argues, demonstrate the inadequacy of any epistemic account: unjustified memory and justified confabulation.

2.1 Unjustified Memory

As noted above, an epistemic account is one on which confabulating is distinguished from remembering in terms of the unjustifiedness of the apparent memories in which it results. This explains why Bernecker treats the simulationist account as an epistemic account, for the concept of reliability, borrowed from reliabilist epistemology (Goldman 2012), plays a central role in the theory of memory on which it is based. The simulation theory treats memory as a kind of imagination, but it does not claim that just any way of imagining the past amounts to remembering: the simulation theorist argues that properly functioning episodic construction systems are reliable and therefore treats reliability as a precondition for

[2] "Unsuccessful remembering" is here used as an umbrella term covering cases in which the memory process (or the metamemory process; see Sect. 3) either produces an inaccurate representation or is itself in some sense deficient.

[3] Since the causal theorist appeals to the absence of an appropriate causal connection both to distinguish both between remembering and confabulating and to distinguish between remembering and imagining, it is not entirely clear how he might distinguish between confabulating and imagining. This potential difficulty for the causal theory will not be discussed here.

[4] That veridical confabulation is possible is most easily seen by comparing confabulation, in the domain of memory, to hallucination, in the domain of perception. Hallucinations are typically inaccurate, but, in principle, they need not be. Whatever factor (e.g., causal connection), in addition to accuracy, one holds to make the difference between perception and hallucination, one must admit that that factor might be absent in the case of a given perceptual representation regardless of whether that representation is accurate; if it is absent and the representation is nevertheless accurate, the representation amounts to a veridical hallucination. Similarly, whatever factor (e.g., reliability), in addition to accuracy, one holds to make the difference between memory and confabulation, one must admit that the factor might be absent in the case of a given memory representation regardless of whether that representation is accurate; if it is absent and the representation is nevertheless accurate, the representation amounts to a veridical confabulation. It should be noted that, while philosophers of perception have long discussed veridical hallucination, philosophers of memory have only recently begun to discuss veridical confabulation. Hirstein (2005) acknowledges the possibility of veridical confabulation but does not use the expression. Robins uses the expression in passing (2016b) but fails to take veridical confabulation into account when developing her version of the causalist account (2016a, 2018). Michaelian (2016a) emphasizes the possibility of veridical confabulation in the course of his argument against Robins' account. Bernecker (2017), as we will see, argues that his version of the causalist account can in fact accommodate veridical confabulation.

the occurrence of genuine remembering. To a first approximation, the simulation theory can be understood as claiming that to remember the past just is to imagine it. More precisely, what the theory claims is that to remember the past is to imagine it in a *reliable* manner.[5] The simulationist account of confabulation thus distinguishes between confabulating and remembering in terms of the unreliability, in the case of confabulating, of the process in question.

It is, nevertheless, a mistake to treat the simulationist account as an epistemic account. Hirstein, for example, does not commit himself to a very definite epistemology, but his account of confabulation does make heavy use of normative epistemic vocabulary—describing confabulating as "ill-grounded" remembering, for example—and is clearly a properly epistemic account. The simulationist account, in contrast, employs no such epistemic vocabulary: rather than describing confabulating and remembering in terms of the (un)justifiedness of their outputs, the simulation theorist treats remembering as a reliable process and confabulating as an unreliable process. The concept of reliability at issue here is that familiar from reliabilism, and reliabilism is, of course, a normative epistemological theory, but, while the concept of reliability can certainly be employed in a normative theory, it is not—unlike justifiedness or well-groundedness—itself a normative notion. Just as a coffee machine might be reliable in the sense that it has a tendency (when certain background conditions, such as the presence of an unused capsule in the capsule drawer, are satisfied) to produce drinkable cups of coffee, an episodic construction system might be reliable in the sense that it has a tendency (when certain background conditions, such as the accuracy of the subject's earlier experiences, are satisfied) to produce accurate apparent memories. The fact that a cup of coffee has been produced by a reliable coffee machine does not, by itself, imply that one ought to or may drink it. Similarly, the fact that an apparent memory has been produced by a reliable episodic construction system does not, by itself, imply that one ought to or may form the corresponding belief. The simulationist account of confabulation is thus not an epistemic account.[6]

In support of this point, note that, given the link between accounts of confabulation and theories of memory, an advocate of an epistemic account will typically be (at least tacitly) committed to an epistemic theory of memory, a theory, that is, that defines remembering in terms of justification or knowledge (see Frise 2015). The idea would be, roughly, that remembering necessarily results in justified memories, while confabulating necessarily results in unjustified memories, where the nature of justification is as specified by the theorist's favoured epistemology. Thus, if the simulationist account were an epistemic account, then we ought to expect the advocate of that account to be committed to an epistemic

[5] Some might find the thought that remembering might occur despite the absence of an appropriate causal connection between the retrieved representation and the corresponding past experience to be puzzling; some might find the thought that remembering might be reliable despite such an absence to be more puzzling yet. Regarding the first puzzle, it is important to recall that an appropriate causal connection is, as noted above, typically understood as one underwritten by a memory trace originating in the relevant experience. Given the way memory traces themselves are typically understood, the simulation theorist's claim is thus that the occurrence of genuine remembering does not presuppose the transmission of information or content to the retrieved representation from the corresponding past experience. This is compatible with the possibility that there will inevitably be causal connections of other sorts between a given retrieved representation and the corresponding past experience—the causal theorist claims that memory is characterized by a causal connection of a specific sort, not simply that it involves a causal connection of some sort or other, and it is this claim, in particular, that the simulation theorist rejects. Regarding the second puzzle, it is important to note that the simulation theorist's claim that no appropriate causal connection is necessary for the occurrence of genuine remembering is compatible with the possibility that such a connection in fact obtains in most cases of genuine remembering. There is, indeed, no need for the simulation theorist to deny that remembering usually involves the transmission of information from experience of the remembered event, as such information plausibly often provides the basis for simulation of the event; the simulation theorist's claim is that such information does not always provide the basis for the simulation. Thus the fact that information is transmitted in most cases of remembering provides a partial explanation of the reliability of remembering. The full explanation of the reliability of remembering will also appeal to constraints on simulation, including constraints provided by knowledge of other specific events and constraints provided by general semantic knowledge. Such constraints will carry part of the weight of explaining accuracy even in cases where information is transmitted, since, even in such cases, remembering is not simply a matter of retrieving stored information unaltered, and they will carry the full weight of explaining accuracy in cases where no information is transmitted.

[6] An additional analogy may help to clarify the matter. According to a standard form of utilitarianism, a morally right act is one that maximizes net pleasure, but the fact that utilitarians make use of the concept of a net pleasure-maximizing act in stating a normative ethical theory does not, of course, imply that that concept is itself a normative ethical concept: one can (even if one is a utilitarian) employ the concept when making claims that are neither normative nor ethical in character. ("This act maximizes net pleasure.") Similarly, according to a standard form of reliabilism, an epistemically justified belief is one that is produced by a reliable process, but the fact that reliabilists make use of the concept of a reliably-produced belief in stating a normative epistemic theory does not imply that that concept is itself a normative epistemic concept: one can (even if one is a reliabilist) employ the concept when making claims that are neither normative nor epistemic in character. ("This memory belief was produced by a reliable process.") One might worry that the link, noted above, between reliability and proper function confers a normative character on the concept of reliability, as the latter figures in the simulation theory in particular, but the worry is unfounded, for the notion of proper function at issue here is not itself normative—the proper function of the episodic construction system is to produce accurate representations of events, but this simply means that the system is designed to produce such representations.

theory of memory. The simulation theory of memory is not, however, (even tacitly) an epistemic theory. Conceding that reliability "is not itself a normative concept", James (2017, p. 114) suggests that the inclusion of a reliability condition in the constructive causal theory of memory (Michaelian 2011a) can only be motivated by the view that remembering necessarily results in justified memories. But the inclusion of a reliability condition in the constructive causal theory is, as Michaelian makes clear, motivated not by epistemological considerations but rather by the view that it captures the feature of successful remembering that demarcates it, as a matter of empirical fact, from unsuccessful remembering. If James' suggestion were right, it would apply equally to the simulation theory. But, again, the inclusion of a reliability condition in the simulation theory is not motivated by epistemological considerations: Michaelian (2011a) argues that the causal condition needs to be supplemented by the reliability condition in order to capture the difference between successful and unsuccessful remembering, and Michaelian (2016b) argues, first, that the causal condition does not accurately reflect the difference between successful and unsuccessful remembering and, second, that that difference is accurately reflected by the reliability condition. The simulation theory of memory is thus not an epistemic theory.

At first glance, this might appear to be a merely terminological point. On closer inspection, it undermines Bernecker's attempt to demonstrate, by means of an appeal to the possibility of unjustified memory, that the simulationist account of confabulation is inadequate. His argument, in a nutshell, points out that remembering is compatible with the presence of undefeated defeaters for the retrieved memory (Lackey 2005), whereas justifiedness is not, which implies that there can be memories that are unjustified but not confabulatory. As we have seen, however, the simulationist account is a *reliability* account but not a *reliabilist* account, and the argument fails to demonstrate that the account is inadequate simply because one might endorse it without also endorsing a reliabilist epistemology. The advocate of the account is therefore not committed to denying the possibility of unjustified but nonconfabulatory apparent memories.

Of course, the simulation theory is *compatible* with reliabilism, and one might suspect that, were the simulation theorist to endorse a reliabilist epistemology in addition to his reliability account of confabulation, this would commit him to the further claim that beliefs produced by remembering are necessarily justified, while beliefs produced by confabulating are necessarily unjustified. The suspicion is, however, misplaced: even a reliabilist simulation theorist need not accept this claim. As Bernecker himself points out, sensible reliabilists do not claim that reliability by itself determines ultima facie justification; instead, they claim that "what confers justification on a belief is an externalist condition [such as reliability], but what takes justification away is an internalist no-defeater condition" (2017, p. 7). Thus the reliabilist simulation theorist might simply endorse a form of reliabilism that acknowledges that the prima facie justification conferred by the reliability of a belief-producing process fails to amount to ultima facie justification when undefeated defeaters are present. Since Bernecker's argument concerns ultima facie justification only—he does not attempt to describe a case of unreliable or prima facie unjustified remembering—we can conclude that his argument does not show that the phenomenon of unjustified memory poses a problem for the simulationist account, even if that account is combined with reliabilism.[7]

2.2 Justified Confabulation

Bernecker's appeal to the possibility of justified confabulation fares no better than does his appeal to the possibility of unjustified memory. Focusing on the phenomenon of boundary extension, in which one remembers more of a scene than one actually saw (see, e.g., Hubbard et al. 2010), he argues that.

> the phenomenon of boundary extension ... tends to be remarkably accurate, so much so that Michaelian claims that "boundary extension need not reduce the reliability of remembering". And since Michaelian endorses reliabilism about justification, it follows that, by his own lights, there are mnemonic confabulations that meet the justification condition. (2017, p. 7)

The reasoning here seems to be the following: boundary extension is reliable; so, assuming reliabilism about justification, boundary extension results in justified memories; apparent memories resulting from boundary extension are confabulations; so some confabulations are justified; the epistemic account says that confabulations are never justified; so the epistemic account is false. We have seen that the simulation theorist is not necessarily committed to reliabilism about justification, but, since this argument concerns prima facie rather than ultima facie justification, it can easily be reformulated in terms of reliability rather than justification. So reformulated, it would run as follows: boundary extension is reliable; apparent memories resulting from boundary extension are confabulations; the simulationist account says that confabulations are never produced by reliable processes; so the simulationist account is false.

[7] It is worth noting that even a properly epistemic account, such as Hirstein's, would not seem to be committed to the problematic claim, for an epistemic theorist is free to take the position that memories are necessarily prima facie justified and confabulations necessarily prima facie unjustified, while the ultima facie epistemic statuses of memories and confabulations depend on their relationships to defeaters.

Considered either as reformulated or as stated by Bernecker, the argument fails, simply because it relies on the assumption that the apparent memories resulting from boundary extension are confabulations. Boundary extension is not standardly treated as a form of confabulation. This is unsurprising, for, unlike standard forms of confabulation, it is a pervasive feature of ordinary remembering in healthy subjects: given the ordinariness of boundary extension, treating it as a form of confabulation would entail treating a significant fraction of our ordinary memories as confabulations, thus robbing the concept of confabulation of its theoretical utility. Boundary extension is, moreover, classified as nonconfabulatory not only by the simulationist account (since it is reliable) but also (in most cases) by the false belief account (since it typically results in accurate apparent memories) and even by some versions of the causal account (since there will typically be a causal connection between an apparent memory resulting from boundary extension and the subject's experience of the apparently remembered event).[8] There is thus no obvious reason to take boundary extension to be a form of confabulation. Indeed, we will see that there is positive reason to take boundary extension *not* to be a form of confabulation.

Bernecker concedes that apparent memories resulting from boundary extension may be accurate with respect to the remembered event but distinguishes between the "truth" of a memory and its "authenticity", where truth is a matter of accuracy with respect to the remembered event itself and authenticity is a matter of accuracy with respect to the subject's original experience of the event (Bernecker 2010) and appeals to the claim that remembering requires both truth and authenticity—"a mental state qualifies as a memory only if it accurately represents the objective reality and accords with the subject's initial perception of reality" (2017, p. 4)—to motivate the claim that boundary extension is a form of confabulation, "an error of commission" that "violates the authenticity condition" (2017, p. 3).

While the truth/authenticity distinction is important, however, the claim that remembering requires both truth and authenticity is implausible. Given the pervasiveness of reconstruction in remembering, it is unlikely that retrieved memories are ever wholly accurate with respect to the corresponding experiences; it is likely, in fact, that they are often highly inaccurate with respect to them. Routine forgetting, of course, means that retrieved memories frequently exclude information that was included in the corresponding experiences, resulting in errors of omission. More to the point, retrieved memories routinely *include* information that was *not* included in the corresponding experiences. Boundary extension provides one example of this sort of error of commission, but there are many others. Consider observer perspective memory: when one remembers an event, one often remembers it not from the perspective from which one originally experienced it (field perspective) but rather from the perspective of a hypothetical observer (observer perspective), often even seeing oneself in the remembered scene. On any standard view of the content of experience, observer perspective memories are bound to be inauthentic.[9] They are also part and parcel of ordinary remembering, and there is no apparent reason to classify them as confabulations or anything less than fully successful memories. In view of the pervasiveness of boundary extension, observer perspective memories, and other such "errors" of commission, the natural conclusion is that they are not in fact errors: memory may aim at truth, but it does not aim at authenticity.

Even if it were to turn out that remembering requires authenticity, in addition to truth, moreover, this would still not imply that boundary extension is a form of confabulation. There are two points that should be made here. First, confabulations can be veridical. Bernecker acknowledges this, but note that it goes for both truth and authenticity: just as an apparent memory that is causally unrelated to the corresponding earlier experience or produced by an unreliable episodic construction system might coincidentally be accurate with respect to the apparently remembered event (i.e., true), it might coincidentally be accurate with respect to the subject's experience (i.e., authentic). Confabulations need not be false memories, regardless of whether falsity is understood in terms of truth or in terms of authenticity. Second, false memories—again, regardless of whether falsity is understood in terms of truth or in terms of authenticity—need not be confabulations. There are memory errors other than confabulation, and boundary extension would, given that it results from the same sort of reconstructive processing that is responsible for the DRM effect, more plausibly be classified as a form of misremembering (Robins 2016a; see Sect. 3 below) than as a form of confabulation. Thus, from the fact that a given representation is inauthentic, we cannot infer that it is confabulatory.

[8] Whether the causal theory can acknowledge that boundary extension can amount to successful remembering depends on the version of the theory in question. Adopting Michaelian and Robins (2018) terminology, *neoclassical* causal theories (e.g., Bernecker 2010) endorse preservationism, the view that the content of a retrieved memories cannot exceed the content of the corresponding earlier experience; they are thus bound to deny that boundary extension can amount to successful remembering. (Note that this does not by itself imply that they amount to confabulation.) *Constructive* causal theories (e.g., Michaelian 2011a; Robins 2016b), in contrast, endorse generationism, the denial of preservationism; they are thus capable of acknowledging that boundary extension can amount to successful remembering.

[9] But see McCarroll (2018) for an argument for the view that observer perspective memories can be authentic.

At this stage, there are two alternatives open to us. On the one hand, we might follow Bernecker in classifying boundary extension as a form of confabulation, despite the fact that it usually results in true apparent memories. On the other hand, we might decline to classify boundary extension as a form of confabulation, despite the fact that it results in inauthentic apparent memories. Given that memory aims at truth rather than authenticity, the latter alternative is preferable. There is, admittedly, an asymmetry between the two alternatives: boundary extension *usually* results in true memories, but it *must* result in inauthentic memories. This asymmetry might initially seem to favour the former alternative, but it does not. Boundary extension is an effect, not a process, and the effect is *defined* in such a manner that it must result in inauthentic memories. The fact that it must result in inauthentic memories thus gives us no reason to classify it as a form of confabulation. What matters is whether the ordinary reconstructive processing that sometimes gives rise to the effect usually produces authentic memories, even if the apparent memories produced by it in cases where the effect occurs are, by definition, inauthentic. The upshot is that to classify boundary extension as a form of confabulation would be to draw a distinction where there is none to be found at the level of the memory process itself.

3 The Simulationist Attack on the Causalist Account

The simulationist account, in short, survives Bernecker's attack unscathed. We will see, in this section, that his defence of the causalist account against the simulationist attack on it is more successful. Ultimately, however, this will point the way to an improved simulationist account.

3.1 The Causalist Classification

The debate between the causal and the simulationist accounts of confabulation was triggered by Robins (2016a), who proposed the classification of memory errors depicted in Table 1. In line with the causal theory, the classification characterizes remembering as occurring when two conditions are met: first, the subject forms an accurate representation of a past event; second, his representation is based on retained information originating in his experience of the event—that is, there is an appropriate causal connection between the subject's representation and his experience. Taking it for granted that confabulation is falsidical, she characterized confabulation as occurring when neither of these conditions is met.

This classification acknowledges two (putative) memory errors in addition to confabulation. It characterizes *relearning* as occurring when the first condition but not the second is met, that is, when the subject forms an accurate representation of a past event despite not having retained information originating in his experience of the event. It is not entirely clear whether relearning—which, in a typical case, occurs when the subject acquires information, forgets it, and then reacquires it—should be counted as a memory error; see below. The classification characterizes *misremembering* as occurring when the second condition but not the first is met, that is, when the subject has retained information originating in his experience of the past event but nevertheless forms an inaccurate representation of it. Misremembering is typified by the DRM effect (Gallo 2013), which occurs when the subject is presented with a list of words (e.g., *hospital, sick, nurse*…) and later recalls having seen a thematically-related but nonpresented lure word (e.g., *doctor*). As Robins sees it, this effect can only be explained if we suppose that, despite the fact that he forms an inaccurate representation (the nonpresented lure word), the subject has retained information (the thematic gist) from the relevant experience.

Table 1 Robins' (2016a) causalist classification

Retention y		Retention n	
Accuracy y	Accuracy n	Accuracy y	Accuracy n
Remembering	Misremembering	Relearning	(Falsidical) confabulation

The notion of misremembering is useful—anticipating an argument given below, it would seem to be an "epistemically innocent" (Bortolloti 2015) error, as opposed to "epistemically culpable" errors such as confabulation—and an adequate classification of memory errors ought to include it. Robins' classification fails, however, to accommodate both *veridical confabulation* (which would have to be characterized, like relearning, as involving accuracy but not causal connection) and *falsidical relearning* (which would have to be characterized, like falsidical confabulation, as involving neither accuracy nor causal connection). The simulation theorist therefore proposes an alternative classification designed to accommodate both of these errors, as well as those acknowledged by Robins.

3.2 The Simulationist Classification

Since the role played by the causal condition in the causal theory of memory is taken over by the reliability condition in the simulation theory, the simulationist might initially suggest the classification depicted in Table 2 (Michaelian 2016a). This classification characterizes remembering as occurring when two conditions are met: first, the subject forms an accurate representation of a past event; second, the imaginative process that produces the representation is reliable (whether or not it involves the retention of

Table 2 Michaelian's (2016a) first simulationist classification

Reliability y		Reliability n	
Accuracy y	Accuracy n	Accuracy y	Accuracy n
Remembering	Misremembering	Veridical confabulation	Falsidical confabulation

Table 3 Michaelian's (2016a) second simulationist classification

	Reliability y		Reliability n	
	Accuracy y	Accuracy n	Accuracy y	Accuracy n
Internality y	Remembering	Misremembering	Veridical confabulation	Falsidical confabulation
Internality n	Veridical relearning	Falsidical relearning	Veridical relearning	Falsidical relearning

information originating in experience of the event). It characterizes falsidical confabulation (confabulating resulting in an inaccurate apparent memory) as occurring when neither of these conditions is met, misremembering as occurring when the second condition but not the first is met (that is, when a reliable imaginative process produces an inaccurate representation), and veridical confabulation (confabulation resulting in an accurate apparent memory) as occurring when the first condition but not the second is met (that is, when an unreliable imaginative process produces an accurate representation).

This initial classification accommodates veridical confabulation, but it does not accommodate falsidical relearning, and veridical relearning has dropped out of the picture as well. In order to accommodate both forms of relearning, the simulation theorist might introduce an "internality" condition, the idea behind which would be that.

> veridical relearning occurs in cases in which the subject seems to remember, and to remember accurately, but in which he himself contributes no content to the retrieved memory representation; falsidical relearning occurs in cases in which the subject seems to remember, though to remember inaccurately, and in which he himself contributes no content to the retrieved memory representation. (Michaelian 2016a, p. 10)

Doing so results in the revised classification depicted in Table 3 (Michaelian 2016a). On this classification, if the internality condition is satisfied, then the subject is either remembering, misremembering, or (veridically or falsidically) confabulating, as above. If the internality condition is not satisfied, then he is (veridically or falsidically) relearning. It is to this classification that Bernecker responds.

3.3 A New Causalist Classification

In the course of his response, Bernecker suggests that relearning is or at least need not be a memory error:

> Relearning is clearly different from remembering, but this does not mean that relearning is a memory error. Relearning is typically preceded by forgetting, which may or may not be regarded as a memory error. And relearning is sometimes accompanied by a source-monitoring error which is a type of memory error where the source of a memory is incorrectly attributed to some specific recollected experience. (2017, p. 11)

The suggestion is that relearning itself does not necessarily amount to an error but that it sometimes involves two (potential) errors and that this might explain the inclination to treat it as itself being an error. If this suggestion is right, the simulation theorist might simply revert to his initial classification. Similarly, if the causal theorist opts both not to treat relearning as a memory error and to take veridical confabulation into account, he might propose the classification depicted in Table 4.[10] While the simulation theorist might revert to his initial classification, however, he need not do so, and closer consideration of these potential errors will reveal that what is in fact required is further improvement of the simulationist classification.

Forgetting is not, in general, an error (Michaelian 2011b; Frise 2018). Some instances of forgetting do, of course, involve error, but the error in question is of a kind other than that with which we are concerned here: in cases of erroneous forgetting, the subject fails to retrieve a memory that he should be able to retrieve, whereas, in the cases with which we are concerned, the subject retrieves a memory that we want to classify as erroneous.[11] Forgetting will therefore be set aside in what follows.

[10] Since this revised causal classification parallels the initial simulationist classification, in the sense that both acknowledge the same set of errors, it is not immediately clear how we might go about deciding between the two; this issue will be discussed in Sect. 4.

[11] In some of the kinds of error described below, the subject rejects a retrieved apparent memory and thus does not form a memory belief; even in these cases, he does initially retrieve an apparent memory.

Table 4 A new causalist classification

Retention y		Retention n	
Accuracy y	Accuracy n	Accuracy y	Accuracy n
Remembering	Misremembering	Veridical confabulation	Falsidical confabulation

Source monitoring failures (Johnson et al. 1993) and other failures of metacognitive monitoring are another matter. Relearning may not necessarily involve error—in particular, if a subject relearns but does not take himself to be remembering, no error would seem to have occurred. The cases most often discussed in the literature, however, are ones in which the subject *does* take himself to be remembering. Martin and Deutscher (1966), for example, describe the case of a subject who undergoes an accident, tells a friend about it, undergoes another accident which causes him to lose all memory of the first accident, is told about the first accident by his friend, forgets having been told, and then takes himself to remember the first accident on the basis of experience; intuitively, the subject—due to his source monitoring error—does not successfully remember. Reflection on the difference between this case and an otherwise similar case in which the subject does not commit a source monitoring error suggests that there may be an important distinction between two kinds of relearning: if the subject makes a second-order error with respect to the source of the information contained in his (first-order) apparent memory, then relearning amounts to an error[12]; if he does not make this sort of second-order error, then relearning does not not amount to an error—indeed, he is, arguably, simply successfully remembering on the basis of his relearning. Now, confabulation is not characterized by a failure to determine the source of recollected information. It is thus likely that source monitoring failure and erroneous relearning constitute a distinct kind of error, and they will not be discussed further here. But metacognitive failure of a different sort does appear to play a role—a role captured by none of the classifications considered so far—in many instances of confabulation, and it is this that suggests a need for further improvement of the simulationist account.

3.4 A New Simulationist Classification: First Attempt

Typical cases of confabulation involve the production of (mostly) inaccurate representations by the filling-in of gaps in memory through the piecing-together of bits and pieces of memories of different events, the displacement of events in time, and so on. But they also involve a failure, on the part of the subject, to recognize, even in cases in which the resulting representations are highly implausible or incongruous with reality, that something has gone wrong with the retrieval process.[13] Schnider, for example, assigns a central role to failures of reality monitoring in his treatment of confabulation, arguing that confabulators "[fail] to suppress—or rather filter—activated memory traces and mental associations which do not refer to current reality" (2018, p. 215). He reports, for example, the case of "Mrs. B", who.

> confabulated events that had not taken place, falsely recognized people, confused the day and the place, and confabulated obligations that she did not have at the present time, although most of them referred to real events and experiences in her past. Her false ideas were not just false verbal statements: they betrayed a confusion of reality, which Mrs. B held with the same conviction as any healthy person. (2018, p. 7).

Schnider further points out that "[v]irtually all students of pathological, mnestic confabulations agree that confabulations emanate from some defect in the retrieval and reconstruction of memories. For some reason the brain produces incorrect memories *and* fails to check that they are false" (2018, p. 198; emphasis added). From a simulationist perspective, confabulation, in cases like that of Mrs. B, seems to involve both first-order unreliability (resulting, in most cases, in the production of an inaccurate representation) *and* second-order unreliability (resulting, in most cases, in a failure to detect the first-order unreliability).[14]

[12] Whether it amounts to a *memory* error is a further question: not every error about memory is a memory error. An account of what makes an error a memory error would certainly be a welcome addition to both the simulation and the causalist account of memory errors, but, since relearning will not be discussed further in this paper, no attempt to provide such an account will be made here.

[13] This presupposes that one can reject (i.e., refrain from believing) a retrieved memory. Philosophers sometimes take it for granted that retrieved memories are believed. There is, however, a significant empirical literature establishing the existence of nonbelieved memories (see Otgaar et al. 2014). The same literature, along with with literature on metacognition (see Michaelian 2012), should allay any concern that the account developed here is overly intellectualist.

[14] Note that Hirstein's account of confabulation, like the account developed in the remainder of this section, acknowledges the role of metacognitive failure in confabulation but employs normative vocabulary both when describing first-order processes and when describing second-order processes, stating that the confabulating subject's (first-order) thought is "ill-grounded" and that he "should" have (second-order) knowledge that his thought is ill-grounded. It thus bears reiterating here that reliability is not a normative concept.

Table 5 A new simulationist classification, first attempt

	Object-level			
	Reliability y		Reliability n	
	Accuracy y	Accuracy n	Accuracy y	Accuracy n
Meta-level				
Accuracy y	Remembering	Misremembering	Rejected veridical confabulation	Rejected falsidical confabulation
Accuracy n	Rejected remembering	Rejected misremembering	Veridical confabulation	Falsidical confabulation

While second-order unreliability is *characteristic* of confabulation, we may not want to take it to be strictly *necessary* for the occurrence of confabulation. A subject who is unlike Mrs. B in that his second-order metacognitive monitoring processes are reliable but like her in that his first-order memory processes themselves are unreliable would still seem to *confabulate*, despite the fact that he manages to filter out (most) of the (mostly) inaccurate representations produced by his unreliable first-order processes. Bearing this in mind, consider the first attempt at an improved simulationist classification depicted in Table 5. *Object-level* (un)reliability and (in)accuracy refer to the first-order properties with which we have been concerned so far: the reliability or unreliability of the retrieval process and the accuracy or inaccuracy of the apparent memories that it produces. *Meta-level* (in)accuracy refers to the accuracy or inaccuracy of the subject's metacognitive judgements with respect to his object-level apparent memories. For the sake of simplicity, we can assume that the subject always engages in metacognitive monitoring of the retrieval process, that his monitoring always results in a determinate judgement the result of which is that he either endorses or rejects the apparent memory produced by the process, and that these judgements are always simply accurate or inaccurate.[15]

Since metacognitive monitoring processes have no means of directly detecting the accuracy of a retrieved apparent memory but can detect features that are correlated with its reliability and hence with the probable accuracy of its products, the accuracy of metacognitive judgements is naturally understood as accuracy with respect to the reliability of the relevant retrieval processes, regardless of whether the apparent memories produced by those processes are themselves accurate. This understanding of accuracy has consequences for our understanding of confabulation. Suppose that we have object-level unreliability. Then the subject is either veridically or falsidically confabulating. Suppose that we nevertheless have meta-level accuracy. Then the subject judges that his retrieval process is unreliable and therefore rejects the apparent memory, regardless of whether it is accurate (in the case of veridical confabulation) or inaccurate (in the case of falsidical confabulation). Because the retrieval process is unreliable, this rejection is, in an important sense, correct: even when confabulating results in an accurate apparent memory, the apparent memory is accurate only due to (good) *luck*,[16] and the subject's decision to reject it is appropriate. In cases of "rejected confabulation", the subject's properly functioning metacognition compensates for his malfunctioning memory system,[17] resulting in a form of confabulation less severe than that displayed by Mrs. B. Full-blown confabulation, whether veridical or falsidical, of the sort displayed by Mrs. B occurs when we have both object-level unreliability and meta-level inaccuracy, corresponding to malfunction at both levels: again, the subject's failure to reject the apparent memory is, in an important sense, incorrect, regardless of whether the apparent memory is accurate (in the case of veridical confabulation) or inaccurate (in the case of falsidical confabulation).

The picture of remembering and misremembering provided by the classification is a mirror-image of this picture of veridical and falsidical confabulation. Suppose that we have object-level reliability. Then the subject is either remembering or misremembering. Suppose that we have meta-level

[15] The notion of metacognitive monitoring employed here is quite generic; additional work would be required to relate it to the rich empirical and philosophical literature on different forms of metacognition; see, for example, Arango-Muñoz (2011) on the relationship between metacognitive feelings and explicit metacognitive judgements.

[16] As Pritchard (2004) has emphasized, there are a number of varieties of luck, and the variety at work here is what he refers to as "veritic epistemic luck". Good veritic epistemic luck occurs when an unreliable process produces an accurate representation: the representation is luckily accurate in the sense that, in most nearby possible worlds, the process instead produces an inaccurate representation. Bad veritic epistemic luck occurs when a reliable process produces an inaccurate representation: the representation is (un)luckily inaccurate in the sense that, in most nearby possible worlds, the process instead produces an accurate representation.

[17] Strictly speaking, the simulationist will want to refer here not to the subject's memory system but rather to his episodic construction system; since our focus is on memory rather than other forms of mental time travel, the former term will be used for the sake of convenience.

accuracy. Then the subject judges that his retrieval process is reliable and therefore endorses the apparent memory, regardless of whether it is accurate (in the case of memory) or inaccurate (in the case of misremembering). Because the retrieval process is reliable, this endorsement is, in an important sense, correct: even when reliable retrieval results in an inaccurate apparent memory, the apparent memory is only unluckily inaccurate, and the subject's decision to endorse it is appropriate. In cases of remembering and misremembering, the subject's properly functioning metacognition ratifies the outputs of his properly functioning memory system. When the subject is remembering, no error occurs. When the subject is misremembering, an error occurs, but the error is entirely attributable to (bad) luck. "Rejected remembering" and "rejected misremembering" occur when we have object-level reliability and meta-level inaccuracy, corresponding to malfunction at the meta-level alone: again, the subject's decision to reject the apparent memory is, in an important sense, incorrect, regardless of whether the apparent memory is accurate (in the case of remembering) or inaccurate (in the case of misremembering).

This first attempt at an improved simulationist classification has two important virtues. First, it highlights the fact that luck plays a role in distinguishing among memory errors. Remembering and falsidical confabulation are the most intuitive of the outcomes distinguished by the classification, since neither involves luck: in remembering, a reliable retrieval process produces an accurate representation, as expected, and, in falsidical confabulation, an unreliable retrieval process produces an inaccurate representation, also as expected. Misremembering and veridical confabulation are less intuitive, since each involves luck: in misremembering, a reliable retrieval process unexpectedly produces an inaccurate representation, and, in veridical confabulation, an unreliable retrieval process unexpectedly produces an accurate representation. Second, the classification highlights the need for further work on rejected remembering and misremembering and rejected veridical and falsidical confabulation. Metacognitive monitoring is, of course, not perfectly reliable, but the former pair of errors are ones in which a properly functioning (and hence reliable) memory system is accompanied by malfunctioning (unreliable) metacognition, and it is not immediately obvious to what they correspond in clinical terms. The latter pair of errors are ones in which a malfunctioning (unreliable) memory system is accompanied by properly functioning (reliable) metacognition; again, it is not immediately obvious to what these correspond in clinical terms.

3.5 A New Simulationist Classification: Second Attempt

These virtues notwithstanding, the classification does not yet make fully clear the role of metacognitive success and failure. The discussion so far has tacitly assumed that meta-level accuracy and meta-level reliability are bound to go together, but just as reliability and accuracy can come apart at the object-level, resulting in either an unluckily inaccurate apparent memory (in misremembering) or a luckily accurate apparent memory (in veridical confabulation), they can come apart at the meta-level, resulting in either an unluckily inaccurate metacognitive judgement or a luckily accurate metacognitive judgement. Thus we need a classification that takes both object-level *and* meta-level accuracy *and* reliability into account.

Taking both object-level and meta-level accuracy and reliability into account produces the improved classification depicted in Table 6. The considerable additional complexity of this classification makes it somewhat more difficult to digest, but the difficulty is offset both by its greater precision and by the fact that it points to the existence of categories of memory error that have so far been overlooked. The errors distinguished by the classification might be grouped together in various different ways, but perhaps the most natural grouping considers the population of subjects with which each error is associated; within each group, errors can then be sorted by the kind(s) of luck they involve (if any). Proceeding in this way gives us us four groups of errors. The upper left quadrant of the table contains those associated with healthy subjects, that is, with subjects who have both properly functioning memory systems[18] and properly functioning metacognition. The lower right quadrant contains those associated with subjects we might refer to as "full confabulators": subjects with both malfunctioning memory systems and malfunctioning metacognition. The upper right quadrant contains those associated with subjects we might refer to as "partial confabulators": subjects with malfunctioning memory systems but properly functioning metacognition. And the lower left quadrant contains those associated with metacognitively impaired subjects: subjects with properly functioning memory systems but malfunctioning metacognition. See Table 7.

Focusing on the first group, *remembering*, on this classification, occurs when a reliable memory system produces an accurate apparent memory that is then endorsed because reliable metacognition produces an accurate judgement. Here, there is no luck at either level, since, at both the object-level

[18] Although it will be convenient to refer to properly functioning memory systems and reliable memory processes, it is, strictly speaking, the latter that matters. See Sect. 4.

Table 6 A new simulationist classification, second attempt

	Object-level			
	Reliability y		Reliability n	
	Accuracy y	Accuracy n	Accuracy y	Accuracy n
Meta-level				
Reliability y				
Accuracy y	Remembering	Misremembering	Rejected veridical confabulation	Rejected falsidical confabulation
Accuracy n	Innocently rejected remembering	Innocently rejected misremembering	Innocently endorsed veridical confabulation	Innocently endorsed falsidical confabulation
Reliability n				
Accuracy y	Culpably endorsed remembering	Culpably endorsed misremembering	Culpably rejected veridical confabulation	Culpably rejected falsidical confabulation
Accuracy n	Rejected remembering	Rejected misremembering	Veridical confabulation	Falsidical confabulation

Table 7 A new simulationist classification, second attempt; alternative presentation

	First group: healthy subjects (no malfunction)	Second group: full confabulators (object-level and meta-level malfunction)	Third group: partial confabulators (object-level malfunction)	Fourth group: metacognitively impaired subjects (meta-level malfunction)
No luck	Remembering	Falsidical confabulation	Rejected falsidical confabulation	Rejected remembering
Object-level luck	Misremembering	Veridical confabulation	Rejected veridical confabulation	Rejected misremembering
Meta-level luck	Innocently-rejected remembering	Culpably-rejected falsidical confabulation	Innocently-endorsed falsidical confabulation	Culpably-endorsed remembering
Object-level and meta-level luck	Innocently-rejected misremembering	Culpably-rejected veridical confabulation	Innocently-endorsed veridical confabulation	Culpably-endorsed misremembering

and the meta-level, a reliable process produces an accurate outcome. *Misremembering* occurs when a reliable memory system produces an inaccurate apparent memory that is then endorsed because reliable metacognition produces an accurate judgement. (Bear in mind throughout that meta-level accuracy is accuracy with respect to object-level reliability, not with respect to object-level accuracy.) Here, there is luck at the object-level, since a reliable retrieval process produces an inaccurate outcome, but there is no luck at the meta-level. What we might refer to as *innocently-rejected remembering* occurs when a reliable memory system produces an accurate apparent memory that is then rejected because reliable metacognition produces an inaccurate judgement. Here, there is luck at the meta-level, since a reliable monitoring process produces an inaccurate outcome, but there is no luck at the object-level. *Innocently-rejected misremembering*, finally, occurs when a reliable memory system produces an inaccurate apparent memory that is then rejected because reliable metacognition produces an inaccurate judgement. Here, there is luck at both levels.

Note that remembering and misremembering, in Tables 6, 7, correspond to remembering and misremembering in Table 5; we have not previously encountered innocently-rejected remembering or innocently-rejected misremembering. The same pattern holds for the subsequent groups: we have considered errors involving no luck and errors involve object-level luck but not errors involving meta-level luck or both object-level and meta-level luck. Space here is too limited to permit detailed discussion of discussion of errors of the latter two sorts; for now, it will have to suffice to describe them in general terms.

All three of the errors in this group—remembering, of course, is not an error—are, unlike those in the following groups, arguably "epistemically innocent" in that they represent epistemic costs that subjects must pay in order to obtain otherwise unobtainable epistemic benefits.[19] Misremembering is an inevitable byproduct of the kind of flexible,

[19] One might worry that the language of "innocence" and (below) "culpability" reintroduces a normative element to the classification, but referring to an error as innocent simply indicates that it is bound to occur if the subject is to obtain certain desirable outcomes, and referring to it error as culpable indicates that it is not merely a byproduct of this sort.

constructive processing that enables memory to play a role in episodic future thought and episodic counterfactual thought, and Puddifoot and Bortolotti (2018) argue that it might have other otherwise unobtainable epistemic benefits as well. Innocently-rejected remembering and misremembering are simply inevitable byproducts of imperfect but efficient metacognitive monitoring processes.

Turning to the second group, the errors can be sorted in the same way. No luck: *falsidical confabulation* occurs when an unreliable memory system produces an inaccurate apparent memory that is then endorsed because unreliable metacognition produces an inaccurate judgement. Object-level luck: *veridical confabulation* occurs when an unreliable memory system produces an accurate apparent memory that is then endorsed because unreliable metacognition produces an inaccurate judgement. Meta-level luck: *culpably-rejected falsidical confabulation* occurs when an unreliable memory system produces an inaccurate apparent memory that is then rejected because unreliable metacognition produces an accurate judgement. Both object-level and meta-level luck: *culpably-rejected veridical confabulation* occurs when an unreliable memory system produces an inaccurate apparent memory that is then rejected because unreliable metacognition produces an accurate judgement.

Unlike the errors in the first group, all four of the errors in this second group are "epistemically culpable", in the sense that they represent costs of deficient mnemic and metacognitive capacities. A subject who characteristically commits the errors in the first group has a memory system that functions well overall: when he gets things wrong (misremembering), fails to get things right (innocently-rejected remembering), or nearly gets things wrong (innocently-rejected misremembering), this is due to chance; most of the time, he simply gets things right (remembering). A subject who characteristically commits the errors in the second group is like Mrs. B: when he get things right (veridical confabulation), fails to get things wrong (culpably-rejected falsidical confabulation), or nearly gets things right (culpably-rejected veridical confabulation), this is due to chance; most of the time, he simply gets things wrong (falsidical confabulation). A more nuanced treatment would have to consider potential epistemic benefits of the errors in this group, since, if any such benefits turn out to be otherwise unobtainable, the errors would qualify as epistemically innocent (Bortolotti and Sullivan-Bissett 2018). But even if the errors were ultimately to qualify as epistemically innocent, they nevertheless clearly have a degree of epistemic culpability, and this is sufficient for classificatory purposes.[20]

[20] The suggestion that confabulation is epistemically culpable is meant to be restricted to mnemic confabulation and is thus compatible with Sullivan-Bissett's (2015) suggestion that certain forms of nonmnemic confabulation may be epistemically innocent.

Moving on to the third group, the errors can again be sorted in the same way. No luck: *rejected falsidical confabulation* occurs when an unreliable memory system produces an inaccurate apparent memory that is then rejected because reliable metacognition produces an accurate judgement. Object-level luck: *rejected veridical confabulation* occurs when an unreliable memory system produces an accurate apparent memory that is then rejected because reliable metacognition produces an accurate judgement. Meta-level luck: *innocently-endorsed falsidical confabulation* occurs when an unreliable memory system produces an inaccurate apparent memory that is then endorsed because reliable metacognition produces an inaccurate judgement. Both object-level and meta-level luck: *innocently-endorsed veridical confabulation* occurs when an unreliable memory system produces an accurate apparent memory that is then endorsed because reliable metacognition produces an inaccurate judgement.

The errors in the second group were epistemically culpable in the sense that they represented costs of deficient memory *and* metacognitive capacities. The errors in this third group are epistemically culpable in a weaker sense, since they represent costs of a deficient mnemic capacity, accompanied by an adequate metacognitive capacity. A subject who characteristically commits the errors in the third group thus displays a mixture of epistemic culpability and epistemic innocence: in rejected falsidical confabulation and rejected veridical confabulation, his properly functioning metacognition compensates for his malfunctioning memory system by preventing him from forming a belief on the basis of a confabulation; in innocently-endorsed falsidical confabulation and innocently-endorsed veridical confabulation, his properly functioning metacognition fails to compensate for his malfunctioning memory system, but this is failure is due to chance—on most occasions, he manages to avoiding forming a belief on the basis of a confabulation.

Considering, finally, the fourth group, the errors can again be sorted in the same way. No luck: *rejected remembering* occurs when a reliable memory system produces an accurate apparent memory that is then rejected because unreliable metacognition produces an inaccurate judgement. Object-level luck: *rejected misremembering* occurs when a reliable memory system produces an inaccurate apparent memory that is then rejected because unreliable metacognition produces an inaccurate judgement. Meta-level luck: *culpably-endorsed remembering* occurs when a reliable memory system produces an accurate apparent memory that is then endorsed because unreliable metacognition produces an accurate judgement. Both object-level and meta-level luck: *culpably-endorsed misremembering* occurs when a reliable memory system produces an inaccurate apparent memory that is then endorsed because unreliable metacognition produces an accurate judgement.

Like the errors in the third group, the errors in this fourth group are epistemically culpable in a weaker sense than the errors in the second group, but, whereas the errors in the third group are culpable in the sense that they represent costs of a deficient mnemic capacity, the errors in the fourth group are culpable in the sense that they represent costs of a deficient metacognitive capacity. A subject who characteristically commits the errors in the fourth group thus displays a different mixture of epistemic culpability and epistemic innocence: in all four errors, his apparent memory is produced by a properly functioning memory system and his endorsement or rejection of that memory is determined by a malfunctioning memory system; thus his meta-level processes are culpable even when he ends up avoiding the formation of a false belief (in culpably-rejected misremembering) or forming a true belief (in culpably-endorsed remembering).

This second version of the improved simulationist classification improves on the first version by distinguishing between meta-level reliability and meta-level accuracy and thereby making clear that not only object-level but also meta-level luck plays a role in unsuccessful remembering. There are nevertheless several worries that one might have about the classification. One might worry, first, that errors involving luck are unlikely to be of interest beyond philosophy. This is unlikely to be the case. Object-level luck is integral to misremembering, which is already studied in psychology. And meta-level luck is crucial to understanding metacognitive impairment, on which there is a large literature. Even errors resulting from both object-level and meta-level luck are likely to be of interest—an adequate empirical framework will need, for example, to distinguish between innocently-rejected misremembering, in which a subject with a properly functioning memory system and properly functioning metacognition by chance fails to form an inaccurate memory belief, and culpably-rejected veridical confabulation, in which a subject with a malfunctioning memory system and malfunctioning metacognition by chance fails to forming an accurate memory belief. This is not to say, of course, that all of the errors distinguished by the classification will be of equal empirical interest but simply to suggest that they are not mere philosophical curiosities.

One might worry, second, that it is unclear how some of the errors predicted by the account could be tested for and that no direct evidence for their occurrence has been provided. Regarding the first aspect of this worry, note that the goal of this paper is not to describe means of testing for the different errors but only to describe them in general terms. It may be possible to test for many of them, and, if it is not, this does not imply that they do not occur. Regarding the second aspect of the worry, note that the goal of the paper is not to provide evidence for the occurrence of the different errors but only to make a prima facie case for their existence. It may be that some of them do not occur in practice, but this cannot be judged in advance.

One might worry, finally, that a classification of errors as elaborate as that proposed here is unlikely to be of any clinical utility. The simulationist account is first and foremost a philosophical account of confabulation and related errors, and a lack of clinical utility would thus do little to undermine it. It is not, however, clear that the account would be particularly difficult to apply in clinical settings. What matters in such settings is whether a given subject falls into one of the second, third, or fourth groups distinguished in Table 7. In order to determine whether a subject falls into one of these groups, all that is required is to determine whether he displays object-level malfunction, meta-level malfunction, or both, and this can be determined by looking for evidence of object-level or meta-level unreliability. Such evidence will normally take the form of frequent inaccurate retrieved memories or inaccurate metacognitive judgements about retrieved memories and should not be particularly difficult to obtain. The issue of clinical utility is addressed further in Sect. 4.

4 The Role of Reliability in the Simulationist Account

The fact that the errors acknowledged by the revised causalist classification (Table 4) are the same as those acknowledged by the original simulationist classification (Table 2) raises the question of how we might go about deciding between the two classifications. With an improved simulationist classification (Tables 6, 7) in place, this question might seem to lose some of its urgency. But we can, of course, ask whether the causal theorist might not propose a classification that acknowledges the same errors as those acknowledged by the improved simulationist classification. In principle, it seems that he might. At the object-level, he can continue to invoke causal connection where the simulation theorist invokes reliability. At the meta-level, causal connection is ill-suited to replace reliability—there is presumably always a causal connection of some sort between a retrieval process and a metacognitive judgement about it, and it is not evident what it might be for a causal connection of this sort to be "appropriate"—but the causal theorist might follow the simulation theorist in invoking reliability at the meta-level. The resulting causalist classification would acknowledge the errors acknowledged by the improved simulationist classification. This hypothetical causalist classification will be less attractive than the improved simulationist classification to the extent that it fails to make the role of object-level luck (and interactions between object-level and meta-level luck) clear, but this is, perhaps, not a decisive consideration. The question of how to decide between simulationist and

causalist classifications thus regains its urgency. In principle, determining whether cases in which there is reliability without retention of information or retention of information without reliability ought to be classified as cases of confabulation might enable us to decide between the accounts on empirical grounds (see Michaelian 2016a). In practice, such cases may be difficult to identify. At least initially, then, it makes sense to look elsewhere for a means of deciding between the accounts, and both Robins and Bernecker have argued that the concept of reliability is, as it figures in the simulationist account, problematic. This final section of the paper will argue that, far from being problematic, the concept of reliability in fact confers an important advantage on the simulationist account.

4.1 Robins on Reliability

In her 2016a,[21] Robins argued that a purely constructive conception of memory—a conception, such as that offered by the simulation theory, that does not treat retention of information as a prerequisite for memory—is bound to obscure the distinction between confabulations and mismemories. This may be true of some purely constructive conceptions, but we have seen that, because it treats reliability as a prerequisite for memory, the simulation theory, in particular, is capable of acknowledging the distinction. In a subsequent article, Robins has more or less conceded the point. She continues to maintain that purely constructive conceptions that do not treat reliability as a prerequisite for memory (e.g., that defended by De Brigard 2014) are incapable of acknowledging the distinction between confabulation and mismemory but grants that the simulation theory is capable of doing so: "Michaelian's account ... allows us to say that the memory errors that occur in everyday cases [such as misremembering] are consistent with memory's function because they are outnumbered by cases where remembering is reliable. Clinical confabulations, on the other hand, are malfunctions because these errors are the more common result of attempts at remembering" (2018). Robins is, however, sceptical about whether this is the right way of distinguishing between confabulations and mismemories:

> Errors may be more common for clinical patients, or it may be only that these errors are more noticeable or that reports from patients are met with more skepticism. Determining how many attempted remembering are errors, in either everyday or clinical cases, is difficult outside of controlled experimental conditions. (Robins 2018)

The suggestion, in short, is that the simulationist account begs the empirical question of the frequency of inaccurate memories among confabulators.

Robins' sceptical argument depends on an understanding of reliability in terms of frequency of error. Given such an understanding, the simulationist account would, in effect, imply that a given apparent memory counts as a (mis)memory if the subject whose memory it is retrieves mostly accurate apparent memories and counts as a (veridical or falsidical) confabulation if that subject retrieves mostly inaccurate apparent memories. Strictly speaking, however, the simulationist account says nothing about the *frequency* with which (in)accurate apparent memories are retrieved in healthy or clinical *subjects* but rather focuses on the *tendency* of certain retrieval *processes* to produce (in)accurate apparent memories. While it was convenient, in developing the improved simulationist classification in Sect. 3, to elide the distinction between properly functioning memory systems and reliable retrieval processes, it is ultimately the latter that matters, since a properly functioning system might operate unreliably on a particular occasion, just as a malfunctioning system might operate reliably on a particular occasion.[22] Consider a coffee machine with a defect such that, when it is activated, it usually produces an undrinkable cup of coffee. One possibility is that the machine employs a single unreliable process. Another possibility is that it usually employs an unreliable process but sometimes employs a reliable process. Similarly, a subject who usually retrieves inaccurate memories might have a malfunctioning memory system, but this does not necessarily mean that his memory system always operates unreliably. If it sometimes operates reliably, then, on those occasions, the subject (mis)remembers, rather than confabulating. In other words, the simulationist account is compatible with the possibility that confabulators (subjects who have malfunctioning memory systems) do not always confabulate but sometimes remember. On the simulationist account, then, a given apparent memory counts as a (mis)memory if the process that produces it is such that it tends to produce mostly accurate apparent memories and counts as a (veridical or falsidical) confabulation if the process that produces it tends to produce mostly inaccurate

[21] In order to avoid any confusion, note that Robins (2016a) was written before Michaelian (2016b) was published, and the simulationist view to which she primarily responds there is De Brigard's (2014) version (see below); Robins (2018) responds to the simulationist view developed in Michaelian (2016b) but was written before Michaelian (2016a).

[22] One might object here that a process that unfolds on a particular occasion cannot be (un)reliable, since the concept of reliability applies to processes with repeated instances. It is indeed the case that reliability is, in the first instance, a property of process types rather than process tokens, but a process token is legitimately counted as reliable if the relevant type is reliable (i.e., such that, when tokened, it tends to produce an accurate representation).

apparent memories. Consequently, the account does not beg the question of the frequency of inaccurate memories among confabulators.

4.2 Bernecker on Reliability

Even if it is possible in principle that most retrieved apparent memories are accurate among confabulators, it can be taken for granted that most *confabulations* are inaccurate. The false belief account of confabulation may thus, as noted in Sect. 1, be good enough for most clinical purposes. Nevertheless, it is worthwhile to ask whether considerations of clinical utility might favour either the simulationist account or the causalist account.

A difference between the versions of the causal theory respectively endorsed by Robins and Bernecker is potentially important here. Whereas Robins, as we have seen, understands mnemic causation in terms of the retention of information, Bernecker understands it in terms of counterfactual dependence: a retrieved apparent memory, for him, counts as being appropriately causally connected to an earlier experience if it counterfactually depends on that experience.[23] Robins' version of the causalist account would appear to be straightforwardly inapplicable in clinical contexts, simply because there is in general no practicable means of determining whether, in a given case of apparent remembering, information retained from the relevant experience has played a role in the production of the apparent memory. Bernecker initially appears to concede that his version of the causalist account, too, is inapplicable in clinical contexts, remarking that "it does not appear to be possible to verify whether the process that gives rise to a patient's memory belief satisfies the counterfactual dependence clause" of his version of the causal theory (2017, p. 11). But he later seems to suggest that we might determine whether the counterfactual dependence clause is satisfied by checking for "manipulability". To say that one event (such as the retrieval of a certain apparent memory) counterfactually depends on another (such as the undergoing of a certain experience) is to say that the first event can be manipulated (influenced) by manipulating the second. Bernecker thus suggests that, "given the connection between counterfactual dependence and manipulability it seems to be possible to interpret experiments that test for the presence of influence and manipulability as testing for the presence of counterfactual dependence" (2017, p. 12).

The suggested interpretation of the relevant experiments, however, is not viable, for the causalist account does not claim that a subject's apparent memories, taken as a type, counterfactually depend on his experiences, taken as a type, but rather that a token apparent memory counts as a memory if it counterfactually depends on the relevant token experience. And there is no conceivable experiment that can test for the presence of counterfactual dependence between token memories and token experiences, simply because we cannot travel back in time to manipulate an experience in order to then (travelling to the present) check for changes in the apparent memory of interest. Bernecker's version of the causalist account thus appears to be, like Robins' inapplicable in clinical contexts.

Bernecker suggests that the simulationist account is in the same boat, that is, that there is no practicable means of determining whether, in a given case of apparent remembering, the process that produced the apparent memory was reliable. The thought here is that, since reliability is a modal notion, whether a given process counts as reliable depends on what happens in other possible worlds: we cannot devise an experiment "to figure out whether a subject in some possible world would acquire more true than false beliefs on the basis of some process" (2017, p. 12). That may be so, but, as long as a process is used multiple times in the actual world, we can often determine its reliability indirectly but with a high degree of confidence. I have used my coffee machine many times, and it has always produced a drinkable cup of coffee; I can therefore be confident that it is reliable, even if the fact that it has always produced a drinkable cup of coffee does not, strictly speaking, guarantee that it has the relevant modal properties. Similarly, as noted in Sect. 3, if a subject retrieves memories many times and those memories are often inaccurate, we can be confident that his memory system is unreliable, even if the fact that it has often produced inaccurate memories does not, strictly speaking, guarantee that it has the relevant modal properties. The upshot is that the simulationist account, unlike the causalist account, is, in principle, applicable in clinical contexts.

This response to Bernecker does not, of course, amount to a positive argument for the clinical superiority of the simulationist account, but there does appear to be an important sense in which the simulationist account aligns better with clinical concerns. What alerts us to the fact that a patient is a confabulator, in practice, is that he appears to have an unreliable memory system—he often retrieves apparent memories that, we know or can confidently infer, are inaccurate—not that he often retrieves apparent memories that are causally unconnected to corresponding earlier experiences. Of course, we may infer that the apparent memories in question are causally unconnected to corresponding earlier experiences: the events that they describe did not occur and, a fortiori, were not experienced. But the fact that there is no causal connection does not here seem to be doing any diagnostic work. In light of this, it begins to seem unclear whether there is any real motivation for the causalist account

[23] Strictly speaking, Bernecker requires both transmission of information via a memory trace and counterfactual dependence, but this does not affect the present argument.

beyond the causal theorist's preexisting commitment to the causal theory of memory.

Acknowledgements Thanks to two anonymous referees for exceptionally detailed comments on an earlier version of this paper and to an audience at the Mental Time (Travel): Memory and Temporal Experience workshop hosted by the Centre for Philosophy of Time at the University of Milan in 2018 for extremely helpful feedback. This paper grew out of discussions at the PERFECT Memory workshop at the University of Cambridge in 2017 and was greatly stimulated by Sarah Robins' talk on confabulation at the Philosophical Perspectives on Memory workshop at the University of Adelaide in 2017.

Compliance with Ethical Standards

Conflict of interest Michaelian declares that he has no conflict of interest.

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

References

Arango-Muñoz S (2011) Two levels of metacognition. Philosophia 39(1):71–82
Bernecker S (2008) The metaphysics of memory. Springer, New York
Bernecker S (2010) Memory: a philosophical study. Oxford University Press, Oxford
Bernecker S (2017) A causal theory of mnemonic confabulation. Front Psychol 8:1207
Bortolotti L (2015) The epistemic innocence of motivated delusions. Conscious Cogn 33:490–499
Bortolotti L, Sullivan-Bissett E (2018) The epistemic innocence of clinical memory distortions. Mind Lang 33(3):263–279
Dalla Barba G (2002) Memory, consciousness and temporality. Kluwer, Boston
De Brigard F (2014) Is memory for remembering? Recollection as a form of episodic hypothetical thinking. Synthese 191(2):155–185
Frise M (2015) Epistemology of memory. In: Fieser J, Dowden B (eds). Internet encyclopedia of philosophy. http://www.iep.utm.edu/epis-mem/
Frise M (2018) Forgetting. In: Michaelian K, Debus D, Perrin D (eds). New directions in the philosophy of memory. Routledge, Abingdon, pp 223–240
Gallo D (2013) Associative illusions of memory: false memory research in DRM and related tasks. Psychology Press, New York
Goldman AI (2012) Reliabilism and contemporary epistemology: essays. Oxford University Press, Oxford
Hirstein W (2005) Brain fiction: self-deception and the riddle of confabulation. MIT Press, Cambridge
Hubbard TL, Hutchison JL, Courtney JR (2010) Boundary extension: findings and theories. Q J Exp Psychol 63(8):1467–1494
James S (2017) Epistemic and non-epistemic theories of remembering. Pac Philos Q 98(S1):109–127
Johnson MK, Hashtroudi S, Lindsay DS (1993) Source monitoring. Psychol Bull 114(1):3
Lackey J (2005) Memory as a generative epistemic source. Philos Phenomenol Res 70(3):636–658
Martin CB, Deutscher M (1966) Remembering. Philos Rev 75(2):161–196
McCarroll C (2018) Remembering from the outside: personal memory and the perspectival mind. Oxford University Press, Oxford
Michaelian K (2011a) Generative memory. Philos Psychol 24(3):323–342
Michaelian K (2011b) The epistemology of forgetting. Erkenntnis 74(3):399–424
Michaelian K (2012) Metacognition and endorsement. Mind Lang 27(3):284–307
Michaelian K (2016a) Confabulating, misremembering, relearning: the simulation theory of memory and unsuccessful remembering. Front Psychol 7:1857
Michaelian K (2016b) Mental time travel: episodic memory and our knowledge of the personal past. MIT Press, Cambridge
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, Abingdon, pp 13–32
Otgaar H, Scoboria A, Mazzoni G (2014) On the existence and implications of nonbelieved memories. Curr Dir Psychol Sci 23(5):349–354
Perrin D, Michaelian K (2017) Memory as mental time travel. In: Michaelian K, Bernecker S (eds). The Routledge handbook of philosophy of memory. Routledge, Abingdon, pp 228–239
Pritchard D (2004) Epistemic luck. J Philos Res 29:191–220
Puddifoot K, Bortolotti L (2018) Epistemic innocence and the production of false memory beliefs. Philos Stud 1–26
Robins S (2016a) Misremembering. Philos Psychol 29(3):432–447
Robins S (2016b) Representing the past: memory traces and the causal theory of memory. Philos Stud 173(11):2993–3013
Robins S (2018) Confabulation and constructive memory. Synthese 193:1561–1583
Schnider A (2018) The confabulating mind: how the brain creates reality. Oxford University Press, Oxford
Sullivan-Bissett E (2015) Implicit bias, confabulation, and epistemic innocence. Conscious Cogn 33:548–560
Szpunar KK (2010) Episodic future thought: an emerging concept. Perspect Psychol Sci 5(2):142–162