When misremembering goes online: The "Mandela Effect" as collective confabulation[1]

Abstract: In recent years, popular fora have seen lively discussion of the "Mandela Effect". So called in reference to the paradigm case of a widely-shared apparent memory of Nelson Mandela's death in prison in the 1980s, the effect occurs, roughly speaking, when individuals who have never met develop highly similar memories of events that never took place. Popular explanations of this phenomenon—e.g., that the seemingly inaccurate memories in question are in fact accurate memories of events that took place in parallel universes—are, to put it mildly, fanciful, and the academic literature so far contains little discussion of the effect or of the mechanisms that might be responsible for its occurrence. The goal of this chapter is to make a case for the existence of the Mandela Effect as a novel collective memory error worthy of serious scholarly scrutiny and to sketch a general account of the mechanisms that give rise to it. We argue, in particular, that the effect is an instance of *collective confabulation*, maintaining that this error occurs when individual misremembering goes

[1] Thanks for feedback to audiences at the 2017 New Zealand Association of Philosophers conference at the University of Otago, the 2017 Philosophical Perspectives on Memory workshop at the University of Adelaide, the 2018 Mental Time Travel: Origins and Function workshop at the University of Otago, the 2018 Australasian Association of Philosophy/New Zealand Association of Philosophers joint conference at the Victoria University of Wellington, a meeting of the Otago postgraduate philosophy seminar, and the Ecole Normale Supérieure Lyon. Thanks especially to Stephen Wright for detailed written feedback. Work on this chapter was supported by grant 16-UOO-016 from the Marsden Fund (administered by the Royal Society of New Zealand) and by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02).

online: whereas, in typical offline environments, subjects who give voice to mismemories about publicly-accessible events of the sort at issue in the Mandela Effect will usually encounter contradictory testimony, subjects who do so in the unusual environments constituted by certain online discussion fora may instead encounter confirmatory testimony, resulting in the reinforcement and stabilization of their mismemories and leading to convergence on shared but inaccurate representations of the past.

1 The Mandela Effect

Growing numbers of users have been flocking to online discussion fora such as Reddit (reddit.com) to discuss *Shazaam*, a 1990s film starring the American comedian Sinbad as an incompetent genie who grants wishes to two children. Some Redditors reminisce about the appearance of the cover of the VHS tape and recall a scene in which candy rains from the sky (u/DonnaGail 2017).[2] Others report fond memories[3] of quoting lines from the film with their siblings (Tait 2016). One recalls watching the film repeatedly while inspecting the tape for defects at his job at a video rental store (u/EpicJourneyMan 2016). None of this is particularly unusual for a beloved children's film, but *Shazaam* is unusual in one crucial respect: the film does not exist.

Plausible explanations for the prevalence of beliefs about this nonexistent film—e.g., that those who claim to remember it are in fact thinking of the 1996 film *Kazaam*, which starred basketball player Shaquille O'Neal as a genie—are often met with resistance and

[2] The prefix "u/" followed by a username and a year will indicate that the reference is to a specific post by that user; the prefix "r/" followed by the name of a subreddit (i.e., a discussion thread) indicates that the reference is to that subreddit.

[3] We will, where no confusion will result, let "memory" refer to any apparent memory.

denial. Many Redditors claim to remember both films: one recalls deciding not to see *Kazaam* because it looked like an imitation of *Shazaam* (Tait 2016), while another recalls ordering two copies of *Shazaam* but only one copy of *Kazaam* for his video store (Tait 2016). Certain Redditors, indeed, are so confident that the film is real that they have—despite the fact that Sinbad himself has repeatedly denied having starred in it—offered rewards for proof of its existence (Tait 2016). (All searches for such proof have, needless to say, come up empty-handed.)

As odd as it may seem that hundreds of people who have never met might remember the same nonexistent film, this case is anything but isolated. In another illustration of the phenomenon in which we will be interested here, some Redditors claim to remember Nelson Mandela's death in prison in the 1980s. The memories in question are often highly detailed, with subjects recalling having watched news reports of the event on television or having discussed it with family members or colleagues (Tait 2016). The surprisingly widespread memory of Mandela's death in prison serves as the paradigm case of the phenomenon known as the "Mandela Effect",[4] which occurs, roughly speaking, when individuals who have never met develop highly similar memories of events that never occurred. There are numerous other instances of the effect, with distinct groups remembering evangelist Billy Graham's funeral being televised (Broome 2013), well before his actual death in 2018, *Monopoly* mascot Rich Uncle Pennybags wearing a monocle (u/TimmehTheShpee 2018), Mother Teresa being canonized before her death (u/ThadeusOfNazereth 2016), Leonardo DiCaprio giving an acceptance speech for the Academy Award for Best Actor for his role in *Titanic* (Broome 2016), and so on. In most cases, the memories at issue are not inherently implausible:

[4] The term was, as far as we can tell, coined by self-described "paranormal consultant" Fiona Broome (2009).

Sinbad's starring in a movie called *Shazaam*, Mandela's dying in prison, or DiCaprio's winning Best Actor for *Titanic* are not intrinsically unlikely occurrences. In all cases, subjects remain highly confident in the accuracy of their memories, despite being unable to provide any nonmnemic evidence for their veridicality.

Online discussions of the Mandela Effect have spawned explanations ranging from the plausible but insubstantial ("our memory isn't as good as we thought"; u/Jhoobie 2017) to the highly fanciful (we are "sliding" between parallel universes or alternate realities (u/AscendedMinds 2017). Perhaps owing to the novelty of the phenomenon, the academic literature so far contains little discussion of the effect or of the mechanisms that might be responsible for its occurrence.[5] The primary goal of this chapter is, accordingly, to make a case for the existence of the effect as a novel collective memory error worthy of serious scholarly scrutiny and to sketch a general account of the mechanisms that give rise to it. A secondary goal is to initiate a discussion of collective memory error more broadly. Both the concept of *collective memory* (see Barash 2017; Michaelian & Sutton 2018) and that of *memory error* (see Robins 2020) have recently been subjected to sustained attention in philosophy, but there has so far been no real discussion of the concept of *collective memory error* or of its relationship to individual memory error.[6] While our focus here is specifically

[5] French 2018 provides a discussion from the perspective of conspiracy theory studies. Maswood & Rajaram 2019 includes a brief discussion from the perspective of the psychology of the transmission of false memory.

[6] One exception is Tanesini's (2018) treatment of collective forgetting, but collective forgetting, if it is in fact an error, is an error of omission, whereas confabulation and the other errors with which we will be concerned here are errors of commission.

on the Mandela Effect, our hope is that this will encourage a wider discussion of erroneous collective memory.

The thesis of the chapter will be that, given an account of confabulation based either on the dominant causal theory of memory (Martin & Deutscher 1966; Bernecker 2010) or on the rival simulation theory (Michaelian 2016b),[7] the Mandela Effect is an instance of *collective confabulation*. The effect, we argue, occurs when individual misremembering—the notion of misremembering will be made precise below—goes online: whereas, in typical offline environments, subjects who give voice to mismemories about publicly-accessible events of the sort at issue in the Mandela Effect will usually encounter contradictory testimony, subjects who do so in the unusual environments constituted by certain online discussion fora may instead encounter confirmatory testimony, resulting in the reinforcement and stabilization of their mismemories and leading to convergence on shared but inaccurate representations of the past.

If we are to show that the Mandela Effect amounts to collective confabulation, we must show, first, that it is collective in character and, second, that it is a form of confabulation. We take these tasks in order: section 2 of the chapter provides background on the concept of collective memory, and section 3 argues that the Mandela Effect is collective in character; section 4 provides background on the concept of confabulation, and section 5 argues that the Mandela Effect is a form of confabulation. Section 6 brings the chapter to a close by outlining some promising directions for future research.

[7] The dominance of the causal theory has recently been challenged not only by the simulation theory but also by the functionalist theory (Fernández 2018, 2019). Because the functionalist theory has not yet been developed in detail, we have elected not to consider its implications for the possibility of collective memory error here.

2 The nature of collective memory

While the literature on confabulation is, as we will see in due course, hardly neat, the literature on collective memory is perhaps even messier. As we have no hope of providing a representative survey of this literature here,[8] we will content ourselves with briefly outlining one plausible approach (based on that developed in detail in Michaelian & Sutton 2017, forthcoming; Michaelian & Arango-Muñoz 2018; Arango-Muñoz & Michaelian 2020), acknowledging that other approaches might have other implications for the possibility of collective memory error.

Our approach has two components. The first is inspired by the distinct but complementary accounts of emergent group minds and collective mentality—including collective memory—proposed by Theiner and Huebner. Theiner (2013, 2018), on the one hand, applies the mechanistic account of emergence developed by Wimsatt (1986) to the case of collective memory, arguing that a group displays an emergent property of remembering to the extent that it satisfied several criteria bearing on the relationships among and roles of its members. Theiner's discussion is somewhat technical, and we will not review his criteria in detail here, but the key criterion, for present purposes, is that the property is affected by cooperative or inhibitory interactions among the members. Huebner (2014, 2016), on the other hand, develops an information-processing approach to collective mentality, including collective memory, arguing that memory may be attributed to a group as a collective mental state to the extent that it satisfies several criteria. We discuss Huebner's other criteria in section 3, but the key criterion, for present purposes, is that (where the members of the group have a mental capacity of the same kind as that attributed to the group, which they of course do in putative cases of collective memory) the computations performed by the group are more

[8] See Michaelian & Sutton 2017 for a brief survey.

complex than those performed by its members. As Theiner and Huebner themselves make clear, what the highlighted criteria in effect require is that the performance of the group be shaped by *interaction* among its members. We will thus take it, in what follows, that collective memory can be said to emerge in a group only to the extent that group members' convergence on a shared representation of the past—we take the necessity of such convergence for granted, but we do not assume that perfect convergence is necessary, that is, we will suppose that the contents of group members' memories must be similar but not that they must be identical—is due to interaction in virtue of which their post-interaction memories have content different from that of their pre-interaction memories.

The shape that interaction takes will, of course, vary from group to group. In the specific case of groups that remember together, we can—and this is the second component of our approach—identify two key processes during which interaction might occur: *encoding* (the transition from experience to stored memory) and *retrieval* (the transition from stored memory to occurrent representation).[9] Each of these processes can be either *parallel* (in the

[9] Note, first, that, by referring to encoding and retrieval, we do not mean to suggest that the memory process can be reduced to these operations. Research on the reconstructive character of remembering (see Michaelian 2011 for a review) has demonstrated that it cannot. We came back to reconstruction below, but note that even reconstructive approaches make use of the concepts of encoding and retrieval, the differences between these approaches and approaches on which remembering can be reduced to encoding and retrieval being that reconstructive approaches recognize that what is retrieved often differs—in many cases dramatically—from what is encoded. Note, second, we simplify by referring only to encoding and retrieval: following encoding, understood as the transition from experience to short-term stored memory, the memory process includes consolidation, the transition from short-term

sense that group members perform the same task, either encoding or retrieving, but do not interact while doing so) or *interactive* (in the sense that group members do interact while performing the task), yielding four possible combinations: parallel encoding followed by parallel retrieval; interactive encoding followed by parallel retrieval; parallel encoding followed by interactive retrieval; and interactive encoding followed by interactive retrieval. When combined with the first component of our approach, this second component suggests that we should expect to see the emergence of collective memory in cases involving interactive encoding, interactive retrieval, or both.

An illustration of an especially robustly collective form of collective memory, one characterized by interactive encoding followed by interactive retrieval, is provided by research on transactive memory (Wegner 1987; Wegner, Erber & Raymond 1991; see Ren & Argote 2011 for an overview). Transactive memory systems are formed when, for example, the members of a family establish what might be termed a "division of mnemic labour", with different members having responsibility for remembering different kinds of information or for managing different stages of the memory process. (For example, a husband and wife might contribute information about distinct aspects of a jointly-experienced event.) In a transactive memory system, as Theiner emphasizes, cooperative and inhibitory interactions among the members of the system, underwritten by their knowledge of who is responsible for carrying out what tasks in the division of mnemic labour, are critical to the successful functioning of the system. As Huebner emphasizes, moreover, the computations performed by a transactive memory system are, due to these interactions, more complex than those

---

memory to long-term memory; following retrieval, it includes reconsolidation, a stage similar to consolidation. A more fully developed framework would take the possibility of interaction during (re)consolidation into account.

performed by its members, in the sense that what the system remembers differs from what is members are capable of remembering on their own.[10] If the other relevant criteria are satisfied (see below), we are thus, given the approach adopted here, entitled to attribute memory to the transactive memory system as a whole.

We offer the case of transactive memory merely as an illustration: the shared memories of the groups in which the Mandela Effect occurs are not as robustly collective as those possessed by transactive memory systems, simply because they result from parallel encoding followed by interactive retrieval rather than interactive encoding followed by interactive retrieval. We will argue in the following section that they meet the criteria for collectivity nonetheless.

3 The Mandela Effect as collective in character

On the approach outlined above, we are entitled to attribute memory to a group when interaction among its members leads them to converge on a memory with content distinct from that of their individual pre-interaction memories. It is clear upon reading the discussions of participants in the fora in which the Mandela Effect occurs that their interaction, in which each member offers testimony about his experience of the relevant event, often has precisely this result. One Redditor, for example, writes:

[10] See the discussion of collaborative inhibition and facilitation in section 3. In addition to the quantitative forms of emergence represented by collaborative inhibition and facilitation, Harris et al. (2014, 2017) identify, in their research on long-married couples as transactive memory systems, distinct qualitative forms of emergence, including the emergence of greater emotional richness and of new understandings of the significance of remembered events. We take Harris et al.'s work into account in what follows, but we will not argue that the Mandela Effect involves qualitative emergence.

> *I read your synopsis and it's very close to what I remember.* I actually owned a copy of this movie that my mom bought from a video store because it was only like $1. One additional scene that I do remember involved a car and the kids wanting the genie to come with them somewhere but he couldn't sit in the car so he was riding on top of it like it was a flying carpet and they were like "No, you can't do that either! That's dangerous/someone will see you!" Cause he kept like almost hitting trees and things and sliding around (which never made sense cause if he's a genie couldn't he use magic? Idk [I don't know]. Lots of plot holes in this masterpiece) and people weren't supposed to see him or whatever. So then he disappeared and they were like "Where'd he go?" And they couldn't find him again until they got out at their destination and he was in the trunk and his body was all like twisted around weird and the kids thought it was so funny. *I was just curious to see if you remembered anything like that?* (u/manafrmhvn 2018, emphasis added)

Another Redditor replies:

> I don't want to inadvertently add to the "Mythos" surrounding this film by adding things that I am not 100% sure about, which is why I have never referred to the movie as "Shazzam" or any variation thereof for eaxample [*sic*] (it was a one word Title and the genie may have used it as a magic word but I can't say for sure that was the name of the movie).
>
> *I can say that yes, there was a whole segment of the film that involved Sinbad hiding and trying not to be seen, and the car scene sounds familiar* and I think had to do with the dad accidentally taking the bottle to work with him but I can't elaborate much more than that other than I think the dad nailed a presentation or meeting because the genie helped him without him knowing.

> I really think the movie was never finished being edited in post production and it was hurriedly released when the Rights to it changed hands to take advantage of Sinbad's popularity at the time.
>
> I actually wouldn't be surprised at all to find out the movie was originally filmed in 1989-90 before he was a big star...and another thing, might be nothing - but I could have swore [*sic*] the kids actually called him "Sinbad" in the movie...though I guess if it was "Shazam" it's pretty close phonetically. (u/EpicJourneyMan 2018, emphasis added)

Exchanges such as these, in which individuals report familiarity with information offered by others and contribute related information of their own, lead the group to converge on a shared representation of the past, as the testimony of multiple individuals is woven together in such a way that, ultimately, the resulting shared memory is a composite of the individual memories with which they began. Redditor u/shazaamthemovie, for example, compiled a list of "known" information about *Shazaam* based on "stuff that multiple people from various sources remember" (u/shazaamthemovie 2017), including the release date, starring actors, a description of the VHS cover, and details about particular scenes. As long as the other relevant criteria are satisfied—and we suggest below that they are—we are thus, given the approach adopted here, entitled to infer that the shared memories at issue are genuinely collective in character.[11]

[11] More cautiously: we are entitled to infer that *at least some* of these are collective in character. In section 6, we will in fact suggest that certain instances of the Mandela Effect are not genuinely collective. The cases considered so far, however, do appear to be collective in character.

One might object to this inference on the ground that nothing *new* emerges through the interactions that take place among the individual rememberers. The posts quoted above illustrate that these borrow testimonial information from each other, but such borrowing is compatible with the possibility that the total content of their post-interaction memories includes nothing not included in the total content of their pre-interaction memories. Consider a simple scenario in which each pre-interaction memory includes some content not included in any other pre-interaction memory and in which borrowing of testimonial information results in a set of post-interaction memories with highly similar contents; the content of each group member's post-interaction memory might differ from that of his pre-interaction memory, but it might nevertheless be the case that no post-interaction memory includes any content not included in one or another pre-interaction memory. The thought behind the objection is that, in a scenario such as this, it is unclear whether we are in fact entitled to attribute memory to the group.

In response to this objection, we point out that, given our approach to collectivity, this thought is simply mistaken. It is useful, in this connection, to consider a related debate over the implications for collaborative inhibition and facilitation for collective memory. Collaborative *inhibition* refers to the finding (Betts & Hinsz 2010; Weldon 2000) that, while "real" (interacting) groups often remember more than their members, in the sense that the quantity of information recalled by the members of a group, when their memories are totalled up, is (due to the fact that they recalled nonoverlapping sets of items) often greater than that recalled by any of them individually, the quantity of information recalled by a real group is (apparently due to disruption stemming from its members' use of incompatible retrieval strategies; see Rajaram & Pereira-Pasarin 2010) typically less than that recalled by the members of a nominal (noninteracting) group, when their memories are totalled up. Observed much less frequently than collaborative inhibition, collaborative *facilitation* (Meade et al.

2009; Harris et al. 2014) occurs when the quantity of information recalled by a real group is greater than that recalled by a nominal group. Attempts to identify cases of collaborative facilitation have no doubt been driven in part by the thought that only if the group remembers more than its members would remember on their own can memory legitimately be attributed to the group. But what matters, given our approach, is not that the real group recalls *more* than the nominal group but rather that what the real group recalls *differs* from what the nominal group recalls; collaborative inhibition, in other words, is just as much evidence of genuine collectivity as is collaborative facilitation.[12] Whereas research on collaborative remembering is concerned with the *quantity* of items of information recalled by group members, we are concerned here with the *content* of what they remember, but a similar point holds: what matters, given our approach, is not that content not included in any individual pre-interaction memory emerges through interaction but rather that interaction leads members to form (sufficiently similar) memories with contents different from those of their pre-interaction memories. And this requirement does appear to be satisfied by case like those illustrated by the quoted posts.

One might object to this response on the ground that, if we can legitimately refer to collective memory in any case in which each of two or more interacting subjects remembers something other than what he would remember absent their interaction, we will bound to recognize *too many* cases of collective memory—even very casual, one-off interactions among subjects might, after all, have an impact on what each of them remembers.

In response to this further objection, we point out, first, that the requirement that subjects remember something other than what they would remember absent their interaction is meant to be necessary for collectivity, not sufficient. Huebner proposes two additional

[12] See Harris et al. 2014, Theiner & Sutton 2014 for a similar line of argument.

criteria: that the behaviour of the group not result from top-down transmission of the intentions of certain group members and that the behaviour of the group not result from simple rules governing individual behaviour. These criteria—and in particular the second—do, as section 5's review of the way in which the Mandela Effect emerges through principles specific to the interactions characteristic of the relevant groups, appear to be met in the cases of interest here, but they may not be met in casual, one-off interactions. The same thing goes for Theiner's additional criteria, though we do not have space here to consider these in detail. If these further criteria (or other suitable criteria—we take no stand here on the correctness of the details of Theiner's or Hubner's approach) are not met by a group, then it will not count as manifesting collective memory. We point out, second, that the criteria proposed by Huebner and Theiner can be satisfied to a greater or lesser extent, and that this enable us to treat collectivity itself as being a matter of degree. If we are prepared so to treat it, then it begins to seem much less implausible to treat even casual, one-off interactions that have an impact on what subjects remember as being cases of collective memory—they will simply be (much) *less* collective than the cases of interest here.[13]

4 The nature of confabulation

Just as the purpose of section 2 was to provide background on the concept of collective memory, the purpose of the present section is to provide background on the concept of confabulation. Confabulation is sometimes defined extremely broadly, so as to include a wide range of both mnemic and nonmnemic errors (see Bortolotti & Cox 2009; Hirstein 2005; Robins 2020; Schnider 2018). *Nonmnemic* confabulation occurs, for example, when a subject, unaware of the real reasons for his choice, "makes up" an explanation (see Nisbett & Wilson 1977 for one well-known study). *Mnemic* confabulation occurs when a subject,

[13] The same considerations can be cited in response to the concern that the cases in question are not instances of collective memory because, while each individual's recall is caused in part by interaction with group members, the memories in question are held by the inviduals rather than the groups.

unable to remember a past event, makes one up. We remain agnostic on the relationship between nonmnemic and mnemic confabulation but will for the sake of convenience adopt a narrow definition, confining our attention to mnemic confabulation. While it has recently received a great deal of attention (Baysan 2018; Bernecker 2017; Bortolotti & Sullivan-Bisset 2018; Fernández 2015; Michaelian 2016a, 2020; Puddifoot & Bortolotti 2019; Robin 2010; Robins 2016, 2019, 2020), confabulation is not the only way for a subject to make up a past event, and, for reasons that will be made clear in section 5, the error that Robins (2016) refers to as "misremembering" is of equal importance in the present context.

The relationship between these two errors can most clearly seen against the background of the *reconstructive* character of memory. Remembering is not, as we sometimes naïvely suppose, simply a matter of storing and retrieving a trace the content of which derives from one's experience of the remembered event. Remembering may involve storing and retrieving content,[14] but, if it does, it typically also involves the integration of content available in the context of retrieval, generated by the subject at the time of retrieval, or originating in his experience of other events, including—and this will turn out to be particularly important here—the reception of testimony. A number of distinct accounts of reconstruction in remembering have been proposed—Robins (2016) distinguishes among connectionist (Sutton 1998), gist-based (Michaelian 2011), and episodic hypothetical reasoning-based (De Brigard 2014) accounts—but all are in agreement on the central point that remembering is not a matter of retrieving a preserved representation of an event but

[14] The view that remembering involves storing and retrieving content has recently been challenged (see Hutto and Peeters 2018; Perrin 2018; Michaelian & Sant'Anna forthcoming), but we will, in line with most of the literature on memory and memory error, take it for granted here.

rather of generating a novel representation via the recombination of information deriving from multiple sources.

It is the reconstructive character of remembering that makes both confabulation and misremembering possible. *Misremembering*, on the one hand, an ordinary error characteristic of everyday remembering in healthy subjects, occurs when reconstruction introduces inaccurate details into an otherwise accurate memory representation. In one study cited by Robins (2016), for example, Brewer and Treyens (1981) conducted an experiment in which subjects observed a typical office scene. The scene did not include a stapler. Nevertheless, when they were later asked to remember the scene, many subjects reported remembering a stapler. This error is typical of misremembering in that the best explanation of the fact that certain aspects of the retrieved representation are inaccurate is that the subject successfully retrieved an accurate representation of other aspects of the scene (Robins 2016).

Confabulation, on the other hand, a much more severe error characteristic of remembering in clinical (e.g., amnesic) subjects, occurs when reconstruction goes wrong in such a way that it may[15] generate a wholly inaccurate representation. Robins (2016) cites, as her key illustration of confabulation, work in Loftus's (1997) influential "lost in the mall" paradigm. In work employing that paradigm, a subject is asked to imagine an event (e.g., being lost in a shopping mall as a child) that he did not experience and, when he is later asked to remember the imagined event, may erroneously judge his memory to have originated in experience rather than imagination. The error at issue here, however, appears to differ fundamentally from that characteristic of clinical confabulators. Consider the report given by

[15] The qualifier is important: confabulation typically results in an inaccurate representation, but it might (in principle) sometimes result in an accurate representation. See the discussion of veridical confabulation below.

Dalla Barba's confabulatory patient SD, who had suffered severe head trauma, when asked what he had done the previous day: "Yesterday, I won a running race and I was rewarded with a piece of meat which was put on my right knee" (Dalla Barba 1993). Initially puzzling, this report becomes more intelligible when we are informed that SD had been a runner and had once injured his knee: the generation of an inaccurate memory, in clinical cases of this sort, results not from an inaccurate source judgement, as in "lost in the mall" cases, but rather from the inappropriate recombination of preserved elements of distinct experiences, that is, from a failure at the level of the reconstructive memory process itself.

Several accounts of confabulation, understood in the clinical sense, are available in the literature.[16] Of these, two of the most natural—the false belief account and the epistemic account—can be ruled out for fairly straightforward reasons. Two others—the causal account and the simulationist account—will provide the framework for the argument of this chapter.

4.1 The false belief account

The false belief account has received little uptake in philosophy but has been important in psychology. A number of different formulations of the account have been proposed (e.g. Feinberg 2001; Talland 1961, 1965; Berlyne 1972; see Hirstein 2005 for a summary), but these have in common that they define confabulation as sincere but *false* memory belief.

While it is natural to define confabulation in terms of falsity, the false belief account is simultaneously too broad, in that it implies that any false memory qualifies (as long as it is believed) as a confabulation, and too narrow, in that it implies that any confabulation that happens to be true is not in fact a confabulation. The narrowness of the account is of particular concern here. As Hirstein points out, "[a] patient who gets a question right after supplying wrong answers to the previous six has not miraculously stopped confabulating"

[16] See Berrios 1998 for a detailed historical review of the concept of confabulation.

(2005, 199). The possibility of veridical confabulation has been a prominent theme in recent debates (Michaelian 2016a; Bernecker 2017; Robins 2019), and we intend to allow for this possibility here. The false belief account can therefore be ruled out. We note, however, that, if the account were to turn out to be right, memories of the sort at issue in the Mandela Effect would still (in most cases) qualify as confabulations simply because they are (usually) false.

4.2 The epistemic account

In contrast to the false belief account, the epistemic account focuses not on the accuracy of the memory but rather on its *justificatory status*. On the standard version of the account, due to Hirstein (2005), if an apparent memory is a confabulation, then it is "ill-grounded", and the subject is not but should be aware that it is ill-grounded.

While the epistemic account is considerably more attractive than the false belief account, it remains problematic. That a subject *should* know that his belief is ill-grounded implies (given a plausible ought-implies-can principle) that he *can* know that it is ill-grounded, but subjects are often simply not in a position to determine whether their beliefs are ill-grounded. This suggests that confabulation cannot be adequately defined along the lines proposed by Hirstein: as Bortolotti and Cox (2009) point out, a confabulation is a confabulation even if the subject is not in a position to determine that it is ill-grounded. It may be possible to avoid this difficulty (e.g., by eliminating the requirement that the subject is not but should be aware that the memory is ill-grounded, thus analyzing confabulation in terms simply of the ill-groundedness of the memory itself), but we will not consider the epistemic account any further here. We note, however, that, if the account were to turn out to be right, memories of the sort at issue in the Mandela Effect would still (in most cases) qualify as confabulations because the subjects in question (usually) have access to defeaters for their memory beliefs—most obviously, the fact that the people who remember the event

in the same way constitute a minuscule of the population of people who remember the event—and are thus indeed in a position to know that those beliefs are ill-grounded.

4.3 The causal account

According to the causal theory of memory (Martin & Deutscher 1966; Bernecker 2010), a subject successfully remembers an event just in case her present representation of that event bears an appropriate causal connection to her earlier experience of it. Building on the causal theory, Robins proposes a taxonomy of memory errors in terms of two conditions: "retention of information from a particular past event" and "construction of an accurate representation of that event at the time of retrieval" (2016, 445). Since the notion of appropriate causation is standardly understood in terms of storage and retrieval of information, the first of these conditions in effect requires *appropriate causation*.[17] The second condition simply requires *accuracy*. Apparent memories that fail to satisfy one or both of these conditions are, according to Robins, erroneous in one or another sense: confabulation is characterized by failure to satisfy both the accuracy condition and the appropriate causation condition, misremembering by failure to satisfy the accuracy condition but not the appropriate causation condition, and the error that Robins refers to as "relearning"—which occurs when a subject learns something, forgets it, relearns it, and forgets relearning it—by failure to satisfy the appropriate causation condition but not the accuracy condition.

[17] We consider an alternative definition of appropriate causation in section 5.2 below. In more recent work, Robins (2020) has proposed a modified version of the causal account capable of distinguishing more effectively between confabulation and relearning. We do not have space here to discuss Robins' new account, but see Michaelian forthcoming.

Though elegant, Robins' proposal takes into account neither the possibility of veridical confabulation nor that of falsidical relearning.[18] It is, moreover, unable to distinguish between veridical confabulation and veridical relearning (since both of these errors would have to be characterized by satisfaction of the accuracy condition but not of the appropriate causation condition) and between falsidical confabulation and falsidical relearning (since both errors would have to be characterized by satisfaction neither of the accuracy condition nor of the appropriate causation condition) (Michaelian 2016a). Motivated in part by the need to avoid these difficulties, Bernecker (2017) proposes a modified version of the causal account. This version of the account denies that relearning is a memory error—the plausible thought is that the error apparent in cases of relearning is a matter of erroneous source judgement, not, strictly speaking, erroneous remembering—and is therefore able to acknowledge the possibility of veridical confabulation. It thus suggests (though Bernecker himself is not explicit about this) a taxonomy on which falsidical confabulation is characterized by failure to satisfy both the accuracy condition and the appropriate causation condition, misremembering by failure to satisfy the accuracy condition but not the appropriate causation condition, and veridical confabulation by failure to satisfy the appropriate causation condition but not the accuracy condition. In the remainder of the chapter, we rely on this modified version of the causal account.

4.4 The simulationist account

Whereas the causal theory maintains that appropriate causation is necessary for memory, the simulation theory (Michaelian 2016b) denies this, drawing on psychological research on

[18] We noted above that one might confabulate either an inaccurate or an accurate representation. Along the same lines, we note that, if one can relearn accurate information, one can also relearn inaccurate information.

mental time travel (see Perrin & Michaelian 2017; Michaelian, Perrin, & Sant'Anna 2020), which has revealed an intimate relationship between our ability to remember the past and our ability to imagine the future, to characterize memory as a form of simulation or imagination: just as successfully imagining a future event does not presuppose a causal link between the subject's current representation and the represented future event, the simulation theory maintains, successfully remembering a past event does not presuppose a causal link between the subject's current representation and the represented past event. Rather than appropriate causation, what is characteristic of remembering, according to the theory, is its *reliability:* a subject remembers an event just in case his present representation of that event is produced by a reliable memory system. Building on this simulation or reliability theory and inspired by Robins (2016), Michaelian (2016a, 2020) proposes a taxonomy of memory errors in terms of two conditions: reliability and accuracy.[19] Apparent memories that fail to satisfy one or both of these conditions are, according to Michaelian, erroneous in one or another sense: falsidical confabulation is characterized by failure to satisfy both the accuracy condition and the reliability condition, misremembering by failure to satisfy the accuracy condition but not the reliability condition, and veridical confabulation by failure to satisfy the reliability condition but not the accuracy condition.

[19] Bernecker (2017) has claimed that the simulationist account of confabulation is a (reliabilist) variant of the epistemic account. Michaelian (forthcoming) challenges this claim, but all that matters here is that the simulationist account, which does not require that the subject be aware of the fact that his representation was produced by an (un)reliable process, is not subject to difficulties analogous to those on the basis of which we set the epistemic account aside above.

The simulation theory defines the function of the memory system as being the production of accurate representations of events from the subject's personal past, which implies that a reliable system is a properly functioning system. The core claim of the simulationist account is thus that confabulation is the product of a *malfunctioning* memory system. In a nutshell, then, the question at issue in the debate between Robins and Bernecker, on the one hand, and Michaelian, on the other hand, is whether (falsidical or veridical) confabulations are distinguished from (mis)memories by their lack of appropriate causal connection with the relevant past event or, instead, by the malfunctioning of the systems that produce them. The two accounts are on equal footing insofar as both acknowledge the existence of the same types of error, and the debate will thus have to be decided on other grounds. Fortunately, it does not matter, for present purposes, how the debate will eventually be decided, for, as we will show in the following section, both accounts imply that the Mandela Effect amounts to a form of confabulation.

5 The Mandela Effect as a form of confabulation

Before making our positive case for the claim that the Mandela Effect amounts to a form of confabulation, we pause to deal with a preliminary worry. The claim that the shared representations at issue here—call these "cME-memories", for "collective Mandela Effect memories"— are confabulatory presupposes the claim that they are mnemic in character. Observing a striking resemblance between accounts given by participants in the relevant online fora and the narratives proposed by conspiracy theorists (French 2018), one might wonder whether the latter claim is correct. Both Mandela Effect-rememberers—"ME-rememberers"—and conspiracy theorists, in particular, posit highly implausible differences between appearance and reality and offer evidence—albeit weak evidence—in support of their counterintuitive views. The organization Architects & Engineers for 9/11 Truth (AE911Truth), for example, disputes the conclusion that the impacts of the aircraft, combined

with the resulting fires, were responsible for the collapse of the Twin Towers, appealing to (weak) evidence that the collapse was instead caused by a controlled explosive demolition (McDowell & AE911Truth Staff 2015). Similarly, some ME-rememberers dispute the belief that Mandela was released from prison and did not die until 2013, appealing to their memories of having read newspaper articles about his death in prison in the 1980s. The worry, then, is that the parallel between conspiracy theorizing and ME-remembering is sufficiently close to suggest that the latter is merely a special case of the former, distinguished only by the fact that the relevant evidence is drawn uniquely from memory.

This line of thought, however, obscures an important difference between conspiracy theories and cME-memories, namely, that, while conspiracy theorists offer alternative *explanations* for the occurrence of events the occurrence of which is disputed neither by them nor by others (or at least assume that such explanations are available), ME-rememberers maintain that events that are otherwise universally taken *not* to have occurred *did* in fact occur. ME-remembering thus appears not to be merely a special case of conspiracy theorizing. This does not, however, imply that there is no interesting relationship between the two categories, for ME-rememberers often resort to conspiracy theorizing in order to explain the lack of nonmnemic evidence for the events they claim, on the basis of purely mnemic evidence, to have occurred. When, for example, searches for the newspaper articles reporting Mandela's death in the 1980s that the ME-rememberer remembers reading prove fruitless, he might offer explanations of the failure to locate the articles that themselves take the form of conspiracy theories, claiming, say, that the "false" memories in question are in fact accurate memories of events that occurred in parallel universes or that we are living in a simulation of some sort (Holt 2018). Along the same lines, if somewhat more modestly, some have explained the fact that so many people remember *Shazaam* by claiming that the film really did exist but that its poor reception by audiences harmed Sinbad's career so severely that

Sinbad erased all trace of its existence (u/Destielluh 2017). Our suggestion, in short, is that the Mandela Effect unfolds in two stages, the first consisting of the production of cME-memories, while the second consists of the (optional) production of downstream conspiracy theories meant to explain the discrepancy between the mnemic and the nonmnemic evidence available to ME-rememberers.

With this preliminary worry out of the way, we turn to our positive case for the claim that cME-memories amount to confabulations. We begin, in section 5.1, by considering the implications of the simulationist account of confabulation. We then turn, in section 5.2, to the implications of the causal account.

5.1 Implications of the simulationist account

If the simulationist account is right, cME-memories qualify as confabulations just in case they are the products of malfunctioning memory systems. We will argue that cME-memories do indeed result from malfunction; it will turn out, however, that the malfunction in question is located at the group rather than the individual level.

5.1.1 Individual-level proper function

Collective *memory* is often described as being a matter of individuals *remembering* together, and it is thus natural to assume that collective *confabulation* must be a matter of individuals *confabulating* together. Natural the assumption may be, but collective confabulation, in this naïve sense, cannot, given the simulationist account, explain the occurrence of the Mandela Effect, simply because it is extremely unlikely to occur. Given that account, confabulation is the product of a malfunctioning memory system, where a malfunctioning system is one that is unreliable and so routinely fails to generate accurate representations. While it is possible in principle for multiple unreliable memory systems to generate highly similar representations of a particular event, the chance of this occurring is, given the wide variety of representations

that such systems might produce, vanishingly small.[20] Collective confabulation, in the naïve sense, is thus unlikely to give rise to shared representations of the sort observed in the Mandela Effect.

This does not, however, mean that the effect does not amount to collective confabulation in another, less obvious sense. Noting that, while ME-rememberers clearly commit an error of some sort, the error that they commit bears little resemblance to clinical confabulation, the idea that we want to explore here is that the Mandela Effect is a form of collective confabulation but that collective confabulation, in the relevant sense, occurs not when individuals *confabulate* together but rather when they commit a much more mundane error together, namely, that of *misremembering*.

Misremembering, in Robins' sense, is typified by the Deese-Roediger-McDermott (DRM) Effect (see Gallo 2010). In the DRM paradigm, a subject is presented with a list of thematically-related words (e.g., *hospital*, *nurse*, *medication*, and *gurney*) and, when he is later asked to remember the words on the list, erroneously reports words (e.g., *doctor*) that were not included on the list but that are consistent with its theme. The causal theorist can provide a particularly natural explanation of the occurrence of this error by supposing that a subject who falls prey to the DRM Effect has retained information about some of the words that appeared on the list and that his memory system predicts other words on the basis of the

[20] One might object here that, if memory systems tend to break down in similar ways, it is not particularly improbable for multiple malfunctioning memory systems to produce similar representations of a particular event. But this remains improbable even if systems tend to break down in similar ways, since the apparent memories generated by a given subject's memory system are constructed on the basis of information retained from that subject's experiences, and since different subjects will normally have had very different experiences.

retained information (Robins 2016). In many cases, these predictions will be right. In cases in which the DRM Effect is observed, they happen to be wrong. This explanation may come more naturally to the causal theorist than it does to the simulation theorist, but it is important to recognize that the simulation theorist can adopt precisely the same explanation (Michaelian 2016a), for, while the simulation theory does deny that information *must* be retained in order for remembering to occur, it does not maintain that information is *never* retained.

If this mechanism—erroneous prediction of past events based on successful retention of information—is responsible for the occurrence of the DRM Effect, it is likely that it is at work in the Mandela Effect as well. Sinbad did have a film career, a film called *Kazaam* appeared at about the time that *Shazaam* is thought to have appeared, and the titles of the two films are not very different. Similarly, Nelson Mandela did eventually die, and many news reports about his imprisonment were published in the 1980s. And Leonardo DiCaprio has been nominated for several Academy Awards and performed in *Titanic*, which was also highly nominated. The most natural way of seeing memories of Sinbad's starring in *Shazaam*, Mandela's dying in prison, and DiCaprio's winning an Academy Award for *Titanic* is as inaccurate predictions of past events on the basis of successful retention of information: just as, in the DRM Effect, retained information about the occurrence of (some of) *hospital*, *nurse*, *medication*, and *gurney* on the list leads the memory system to incorrectly infer that *doctor* was likewise on the list, in the Mandela Effect, retained information about Mandela's imprisonment and his death leads the system to incorrectly infer that he died in prison. The individual-level apparent memories at issue in the Mandela Effect–"iME-memories"— thus appear to be mismemories.

One might object that is an important difference between the Mandela Effect and the DRM Effect: in the Mandela Effect, the falsely remembered event (Mandela's death in prison) "overlaps" with the events from which information is retained (his imprisonment, his

death), whereas, in the DRM Effect, the falsely remembered event (the occurrence of *doctor* on the list) is merely thematically related to the events from which information is retained (the occurrence of *hospital*, *nurse*, *medication*, and *gurney*). In response, we point out that our claim is not that the iME-memories are DRM memories but rather that they are, like DRM memories, mismemories. This claim implies that the Mandela Effect is produced in part by the same general mechanism that underwrites the DRM Effect, but it is perfectly compatible with the existence of important differences between the two effects.

In support of the view that the two effects are produced in part by the same mechanism, and in line with the simulationist taxonomy of memory errors reviewed in section 4 above, we note that there is no evidence of malfunction in either case.[21] In the case of the DRM Effect, if the systems in question were malfunctioning, we would expect the reported but nonpresented words to be thematically inconsistent with the presented words,

[21] It is worth observing that different versions of the simulation theory may have somewhat different implications with respect to the DRM Effect and misremembering more generally. De Brigard (2014), whose view is sometimes treated as a version of the simulation theory, sees the function of the memory system as being *episodic hypothetical thinking:* the function of the system is to predict what *might have* happened, not what *did* in fact happen. Michaelian (2016b), in contrast, on whose version of the simulation theory we rely here, sees the function of system as being precisely the prediction of what did happen. It is uncertain whether a version of the simulation theory in line with De Brigard's view can acknowledge that DRM cases are cases of *error*, since there is a clear sense in which, for example, *doctor* "might have" appeared on a list including *hospital*, *nurse*, *medication*, and *gurney*. This version of the simulation theory can, however, like Michaelian's version, acknowledge that the memory system *functions properly* in DRM cases.

whereas they are, of course, thematically consistent with them. Crucially, because the systems in question are not malfunctioning, they tend, when they err, to err in similar ways: attempting to recall a list of words that included *hospital*, *nurse*, *medication*, and *gurney*, it is not unusual for a subject to erroneously report that it also included *doctor*. Along the same lines, iME-memories appear to result from erroneous but similar predictions by properly functioning memory systems. *Kazaam*, for example, might easily be misremembered as *Shazaam*, and multiple rememberers might thus find themselves entertaining the very same mismemory.

If the Mandela Effect arises from individual-level misremembering, it does not involve malfunction at the individual level; the malfunction involved in the effect, if a malfunction is indeed involved, must thus be located at the group level, that is, in the interactions that take place among ME-rememberers. In the next section, we argue that the effect does in fact involve malfunction.

5.1.2 Group-level malfunction

The claim that groups of ME-rememberers—"ME-groups"—are *malfunctional* presupposes, of course, the claim that they *have* a function. We readily grant that not every group has a function. Merely nominal groups—groups of noninteracting individuals who share some distinguishing feature, such as spectators watching a football match on television—presumably do not have functions. What we might think of as "accidental" groups—groups of interacting individuals who have been brought together without any unifying purpose, such as spectators at a stadium jostling each other as they attempt to gain their respective seats—likewise do not seem to have functions. But "intentional" groups—groups that have been brought together by a shared goal, such as the team striving to win the game—plausibly do have functions. In at least some cases, such groups behave as *systems:* their actions are not merely aggregations of their members' individual actions and not merely the spontaneous

results of their members' interactions but rather the outcomes of their members' interaction in pursuit of a shared goal by means of various forms of coordination and cooperation (Tollefsen 2015).[22] It is, for example, by no means unnatural to say that the function of the football team is to win the match, and the groups of ME-rememberers that come together online appear to be like the football team in this respect. Their members interact with each other and thus do not constitute merely nominal groups. And they share a goal—the only reason that the members of a subreddit dedicated to discussing Mandela's supposed death in prison are members of that subreddit is that they are interested in knowing whether he really did die in prison—and thus do not constitute merely accidental groups. The view that ME-groups have functions is therefore plausible.

The shared goals in virtue of which groups constitutes systems are of various sorts. The football team aims at winning the match, a political party might aim at gaining control of the legislature, and a scientific lab might aim at discovering truths about its area of inquiry. The shared goals characteristic of ME-groups are, we suggest, similar to that of the scientific lab in that they are *alethic* in character. Whereas the scientific lab aims (and, when all goes well, aims successfully) at discovering truths about its area of inquiry, the ME-group aims (however unsuccessfully) at discovering the truth about the event that its members misremember. The shared goals characteristic of ME-groups are thus more specifically *mnemic* in character. Our suggestion, in short, is that there is an analogy between the function of the *individual memory system* and that of the ME-group viewed as a *collective memory system:* both aim at producing accurate representations of past events. If this analogy obtains, then, because ME-groups systematically fail to produce accurate representations, we can

[22] Tollefsen defends the thesis that groups that interact in this way qualify as *agents;* our argument here does not presuppose this stronger thesis.

infer, given the simulationist account of confabulation, that they are malfunctional and thus that cME-memories are confabulations.

The suggested analogy, we acknowledge, can be challenged from both directions. On the one hand, some[23] have argued that *individual* memory systems do not in fact aim at producing accurate representations of past events. We have defended the view that memory aims at truth in detail elsewhere (Michaelian 2016b) and will therefore not respond to this challenge here. On the other hand, some have argued that *collective* memory systems do not aim at producing accurate representations of past events. Harris et al. (2014), in particular, have made a thorough case for the view that, though collective remembering may sometimes result in the production of accurate representations of the past, it primarily serves other purposes, such as the promotion of social bonds and the reinforcement of group membership and collective identity, purposes which are, moreover, often orthogonal to or outright incompatible with the production of accurate representations. Consider, for example, the way in which a married couple might revise their shared representation of a past conflict for the sake of harmony in their marriage. In a case such as this, Harris et al. emphasize, the truth will often be a consideration of minor importance. That the married couple fails to attain the truth is thus irrelevant to whether the couple, viewed as a collective memory system, functions properly. If ME-groups are like the married couple—aiming chiefly at the reinforcement of group membership or at another similar outcome—then it would be a

[23] See De Brigard 2014, whose view we described above. See also Mahr & Csibra 2018, who see the function of memory as being *argumentative* in character: truth, on their view, is secondary; the primary role of the memory system is to enable the subject to persuade his interlocutors.

mistake to characterize them as aiming at the truth, and the fact that they systematically fail to produce accurate representations would then not be evidence of malfunction.

ME-groups appear, however, to be more like individual memory systems than they are—assuming that Harris et al. are on the right track—like typical collective memory systems. These groups exist, as noted above, only because their members seek each other out, and their members seek each other out only because they are interested in knowing whether, for example, Mandela really did die in prison: in a typical case, the members of the group have come to the relevant online forum for the express purpose of finding out whether others remember the same event that they do, and the main aim of their discussion is very explicitly to figure out whether the event in question occurred. There is thus no pre-existing group membership in which remembering together might function to reinforce. And there are no bonds among group members that do not stem directly from their shared alethic/mnemic goals. It therefore seems safe to conclude that, whether or not collective memory system in general aim at the truth, the collective memory systems constituted by groups of ME-rememberers, in particular, do so.[24] And this implies, as we have seen, that ME-groups are malfunctional and hence that cME-memories are confabulations.

Before turning to the implications of the causal account, we respond to two objections to our argument regarding those of the simulationist account. The first concerns the applicability of the notion of reliability to ME-groups. Because groups of ME-rememberers typically come together for the purpose of discussing a single putative past event, they typically exist for only a short time before dissolving. The fleeting existence of the group

[24] This does not mean that the promotion of social bonds and the reinforcement of group membership and collective identity plays no role in the workings of ME-groups; we come back to this point below.

means that it forms only a single inaccurate representation rather than a series of inaccurate representations. And this, in turn, renders it difficult to see how the notion of reliability (and hence the notions of malfunction and confabulation) might be applied to it. In response to this objection, we point out that even a system that comes into existence, generates a single representation, and then goes out of existence can legitimately be classified as reliable or unreliable, simply because (un)reliability is a matter not of the observed *frequency* with which a system generates (in)accurate representations but rather of its *disposition* to generate (in)accurate representations (Goldman 2012). Considering ME-groups as a class, we can observe that they do in fact systematically generate inaccurate representations,[25] and this strongly suggests that the members of that class are unreliable, in the sense that they are disposed to generate inaccurate representations. Considering (as we do below) the forms of testimonial interaction that propel the generation of cME-memories, moreover, likewise suggests that individual ME-groups are disposed to generate inaccurate representations, as these drive the group to converge on a shared representation regardless of its accuracy.

The second objection is that, given that the Mandela Effect arises from interactions among individual misrememberers, it would be more appropriate to characterize it as a form of *collective misremembering* than as a form of *collective confabulation*. Like the individual-level iME-memories on which they are based, after all, group-level cME-memories are

[25] Lest it be thought that this is an artifact of the way in which we have defined ME-groups—by definition, they are composed of misrememberers—we point out that our definition is the natural one. ME-groups come together specifically because their memories differ from prevailing (accurate) representations of past events. Other groups may come together online to discuss events that they accurately remember, but their motivations and their operations will inevitably be different. ME-groups thus seem to constitute a natural kind.

inaccurate but close to the truth; in this, they would seem to have little in common with individual-level confabulations, which are often not only inaccurate but very far from the truth. In response to this objection, we point out that it overlooks the fact that, given the simulationist taxonomy, accuracy does not determine whether a system is confabulating. What determines whether a system is confabulating is, instead, its reliability: if the reliability condition is satisfied, then the system is either remembering or misremembering; if the reliability condition is not satisfied, then the system is either veridically or falsidically confabulating.[26] ME-groups systematically fail to attain the truth (even if they systematically come close to it); they therefore do not satisfy the reliability condition, and cME-memories therefore amount to confabulations, rather than mismemories.[27]

This response raises two additional worries. First, one might worry that there is something unsatisfying about a classification that appeals to the fact that ME-groups systematically fail to attain the truth but disregards the fact that they systematically come close to it. In response, we note that our position need not disregard the latter fact. ME-groups are, it turns out, interestingly unlike the ME-rememberers that compose them. ME-rememberers are, we reiterate, not clinical confabulators but rather ordinary rememberers who happen to misremember: they systematically attain the truth but happen, in the relevant cases, to fail to do so. Nevertheless, due to their overall reliability, they come close to the

[26] An analogous point can be made with respect to the revised causal taxonomy of memory errors: if the appropriate causation condition is satisfied, then the system is either remembering or misremembering; if the appropriate causation condition is not satisfied, then the system is either veridically or falsidically confabulating.

[27] This does not mean that the notion of collective misremembering is without interest; we come back to this below.

truth even in those cases. ME-groups, in contrast, despite being composed of ordinary misrememberers, are more like clinical confabulators, in that they systematically fail to attain the truth. They are not, however, exactly like clinical confabulators: due to the fact that cME-memories inherit content from the corresponding iME-memories, they tend to come close to the truth. A full explanation of the tendency of ME-groups to come close to the truth even while not attaining it will require a description of the mechanisms that take us from individual-level mismemories to group-level confabulations; we offer an initial description of these mechanisms below.

Second, one might worry that, if these mechanisms can take us from individual *mismemories* to *falsidical* collective confabulations, they might also be able to take us from (accurate) individual *memories* to *veridical* collective confabulations. In the scenarios with which we are concerned, the members of the group start off with inaccurate apparent memories, and the group ends up with a falsidical confabulation. It is thus natural to suppose that, in a scenario in which the group members instead start off with accurate apparent memories, the group will end up with a veridical confabulation. In response, we note that, while this possibility is compatible with our overall argument, there is an important asymmetry between the two scenarios, due to the fact that the mechanisms that take us from individual mismemories to falsidical confabulations are ineffective at filtering out inaccurate apparent memories but nevertheless effective at preserving accurate apparent memories. These mechanisms, again, will be described in below, but the basic idea is that group members start off with highly similar apparent memories which are then reinforced and stabilized as they exchange testimony about the past. This reinforcement and stabilization process is conditionally reliable (Goldman 2012): while it tends to output inaccurate representations when given inaccurate representations as input, it also tends to produce accurate representations when given accurate representations as input. If confabulation is

defined in terms of (conditional) reliability, the accurate counterparts of cME-memories will thus not amount to veridical collective confabulations.[28]

As an aside, we note that the accurate counterparts of cME-memories might be epistemically defective even though they are not veridical confabulations. They may satisfy the requirements of a process reliabilist epistemology (Goldman 2012), but they are unlikely to satisfy the more demanding requirements of a virtue reliabilist epistemology. Consider the security condition proposed by Sosa (2007). A belief is secure, in Sosa's sense, to the extent that the process that produces it produces true beliefs in nearby possible worlds. Given that the collective memories in question come about due to the stabilization and reinforcement of group members' apparent memories, regardless of the accuracy of the latter, the relevant process will produce a false belief in any nearby world in which group members start off with inaccurate rather than accurate apparent memories. It is unclear to what extent failure to satisfy the security condition means that these collective memories are epistemically defective, just as it is unclear to what extent failure to satisfy that condition would mean that individual memories are epistemically defective: Shanton (2011), for example, argues that

[28] There is another way in which veridical collective confabulation might come about. Consider a typical case of falsidical collective confabulation: subjects *accurately* experience an event and later *misremember*, in the sense that their apparent memories do not correspond to their experiences; their malfunctional interaction then gives rise to a falsidical collective confabulation. Suppose that were to *inaccurately* experience the event and then later misremember; their malfunctional interaction might then give rise to a *veridical* collective confabulation. What we have in mind here is, for example, a case in which it turns out that Mandela died in prison but was, unbeknownst to everyone, replaced by a doppelgänger. Such cases are unlikely to occur in practice, and we will not consider them any further here.

individual memories are unlikely to satisfy the security condition but acknowledges that this may simply signal a problem with Sosa's approach. Our goals in this chapter do not require us to resolve this issue, and we will not pursue it any further here.

Setting the issue of security aside, we turn to the mechanisms that take us from individual mismemories to collective confabulations. Our basic claim is that the Mandela Effect arises due to the effects that testimony has on memory in the unusual environments constituted by certain online discussion fora. The core idea is that these fora bring together groups of subjects who share similar mismemories, allowing them to exchange testimony about their putative past experiences, thus leading, through mechanisms that are likewise operative in offline environments but that do not, in offline environments, typically have the opportunity to lead to similar outcomes, simply because there is nothing that brings appropriate groups of subjects together, to the stabilization and reinforcement of their mismemories. Consider an ordinary subject with a properly functioning memory system. Suppose that, due to the reconstructive character of the memory process, he misremembers an event. In many ordinary cases, the subject's misremembering will have no consequences worth speaking of at the collective level, simply because he does not communicate his mismemory to anyone else. In other ordinary cases, it will have unsurprising consequences at the collective level, because, when the subject communicates his mismemory to others, he receives contradictory testimony in response: assuming that the subject displays a normal level of epistemic humility, the reception of contradictory testimony will lead him to reject his apparent memory or at least to suspend judgement with respect to it—his mismemory will in effect be corrected by the successful memories of his interlocutors.[29] In cases in which the Mandela Effect arises, in contrast, individual misremembering has surprising consequences at

[29] Epistemic humility may sometimes be outweighed by the vividness of the apparent memory, leading the subject to endorse the latter despite having received contradictory testimony. Even in such cases, however, his memory is unlikely to have interesting consequences at the collective level, because anyone to whom he communicates it is likely to reject it out of hand.

the collective level: when the subject communicates his mismemory to other participants in one of the relevant fora, he receives not contradictory but rather confirmatory testimony in response, because these fora put him in contact with others who share similar mismemories—his mismemory is thus reinforced and stabilized by the similar mismemories of his interlocutors.

In principle, this sort of reinforcement and stabilization—which amounts, if the simulationist account is right, to collective confabulation—might occur in offline environments. In practice, it is unlikely to occur in such environments—or, if it does by chance occur, to endure for very long—simply because the odds of one subject being in contact with others who share similar mismemories are low. Thus, even if the subject displays an abnormally low level of epistemic humility, his mismemory will remain isolated. Online environments, however, notoriously enable subjects to identify and interact preferentially with likeminded others. And this means that even an epistemically humble subject may end up persisting in accepting a representation of the past that he would otherwise reject as an obvious mismemory. One mechanism that can in principle be operative offline but is far more effective online is the formation of echo chambers, in which "most available information conforms to pre-existing attitudes and biases" (Lewandowsky, Ecker, & Cook 2017, 359).[30] Even in threads that do not explicitly forbid the expression of sceptical views, members who do express such views are often ostracized, as demonstrated by the responses garnered by a particular sceptical comment (since deleted!) on a thread entitled

[30] It has been argued that the problem of online echo chambers is overstated (Dubois and Blank 2018; Garrett 2017). This may be the case, but our claim is not that echo chambers by themselves explain the Mandela Effect but rather that they are one feature among others of the online environment that make it particularly likely to give rise to the effect.

"254 Confirmed Mandela Effects: List" (u/ezydown 2017). U/melossinglets, for example, challenges the comment with:

> firstly,why dont you go to your precious google and look up the meaning of the word "skeptic"...then,once youve let that set in and marinate a little,see if you reckon that that definition correlates nicely with "a bunch of people coming in and simply telling everyone they disagree with that they are wrong and trotting out the exact same "cover-all" excuse for hundreds and hundreds of folk theyve never met in their life,basically making one huge assumption about all of their various experiences and painting them all with the same brush".....im not entirely sure it will.....but cool,whatever. (u/melossinglets 2017)

Many threads, moreover, do explicitly enable the suppression of sceptical views. The subreddit r/MandelaEffect, for example, allows users to filter posts by "skeptic" or "no skeptic" tags (u/Denominax 2017), presumably facilitating the segregation of "believers" from sceptics.

The formation of echo chambers itself results from a number of more general factors. In an ME-group, as in any group of ideologically-aligned subjects, members may be "unconsciously motivated to resist empirical assertions […] if those assertions run contrary to the dominant belief within their groups" as a form of "identity self-defense" (Kahan 2013, 408). Groups are, in general, sensitive not only to truth but also to agreement, and the members of a group may "accept" certain propositions for the sake of agreement within the group even when they do not in fact believe them.[31] This is so even when the group's aims are explicitly veritic. Discussing this phenomenon in the context of scientific publication

[31] In extreme cases, the group may thus believe a proposition that no individual member believes; see, e.g., Hakli 2006, 2007; Tuomela 1992; Wray 2001.

practices, Rehg and Staley refer to it as "*heterogeneous consensus*, in which a collaborator agrees to the publication of an evidence claim, while disagreeing on the premises offered in that publication as support for the claim" (2008, 10). For instance, the four hundred and fifty (!) individual researchers involved in the Collider Detector at Fermilab Collaboration who endorsed the findings of that collaboration overwhelmingly reported that the conclusion was "basically correct," but disagreed on other points pertaining to the conclusion (Staley 2007). Similarly, ME-group members may be willing to endorse the gist of the emergent cME-memory even if certain aspects of it conflict with their own individual memories. Once a cohesive group has coalesced around a given memory, moreover, its members may tend to continue to endorse their apparent memories for the sake of group stability and continuity even when they would otherwise doubt them. Again, this may be the case even if the aims of the group are explicitly veritic (as we have argued that they are in the case of ME-groups). All of these factors, like echo chambers themselves, may be operative in offline groups; the point is, first, that they are in many cases amplified by the online environment (there is no option to filter one's interlocutors by "skeptic" and "no skeptic" tags offline) and, second, that that environment is what enables groups of subjects to congregate on the basis of similar mismemories in the first place.

In addition to the formation of echo chambers, which is a general feature of online discourse, there may be mechanisms that are specific to ME-groups. First, forum participants are often encouraged to identify new instances of the Mandela Effect and to offer new evidence for existing instances. Consider, for example, the subreddit r/MandelaEffect, which has a permanent post with the following instructions.

> Do you believe you've discovered a new Mandela Effect? Post it in the comments below to see if anyone else has experienced it too! Make sure you include why you think it could be a Mandela Effect and *as many details as possible so people can*

> *respond and discuss with what they remember*. (u/AutoModerator 2018, emphasis added)

Or consider the following report from one Redditor.

> "Three of my coworkers and I were talking about a product we have, and the name sounded similar to Sinbad. We ended up discussing the actual movie called Sinbad, and then it led to discussing the comedian. I listened to this exact conversation go down between the three of them:
>
> 1: "Yeah, he was in a movie! Umm ... he was like a genie or something." 2: "Oh yeah, I remember that. It was called .... Shazam? Oh wait, that was Shaq." 3: "No, no. That was Kazam. Different movie." 2: "Oh okay. Man, I haven't seen Shazam in so long."
>
> I had no influence on the discussion. One sort of knows about the Mandela Effect, but I confirmed after this conversation that he had no idea that there was anything about the existence of Shazam." (u/Fae_Leaf 2018)

Second, contributed details often cue other members to contribute further details, leading to the gradual production and refinement of a shared representation. The Mandela Effect may thus arise in part due to the Misinformation Effect (Loftus 2005). We discuss the Misinformation Effect in more detail below, but the basic idea is that, when participants present information about their own mismemories, this information—even if false—may be incorporated into the memories of others.

There is, of course, much more to be said about the details of the mechanisms that give rise to the Mandela Effect, but the foregoing should suffice to make their basic contours clear. As research on the Misinformation Effect and related phenomena has made abundantly clear (see Michaelian 2013), testimony regularly has an impact on memory. What is distinctive of the Mandela Effect is that it occurs when online discussion fora bring together

groups of subjects who share similar mismemories, allowing them to exchange testimony about their supposed experiences, thus leading, through mechanisms that are likewise operative in offline environments, to the stabilization and reinforcement of their mismemories. This process amounts, if the simulationist account is right, to collective confabulation.

5.2 Implications of the causal account

In the previous section, we considered the implications of the simulationist account of confabulation for the status of cME-memories, arguing that, because ME-groups aim at but systematically fail to attain the truth, these apparent memories are, given that account, confabulatory in character. In the present section, we consider the implications of the causal account. If that account is right, cME-memories qualify as confabulations just in case they fail to satisfy the appropriate causation condition. We will argue that—given the most plausible available version of the causal theory—cME-memories do indeed fail to satisfy that condition.

5.2.1 The classical causal theory

Before going any further, we need to determine whether and how the appropriate causation condition might apply to group-level representations such as cME-memories. The application of the condition to individual-level representations is straightforward: an individual-level representation satisfies the condition just in case it is causally connected to the subject's original experience of the represented event and the causal connection in question is sustained by the transmission of information originating in that experience. Group-level representations, at least in the cases of interest here, cannot satisfy the condition in the same way, simply because, in those cases, the group did not experience the event that it represents: even if we grant that a group might in principle experience an event (and this is not entailed by the claim that a group might remember an event), the subjects that compose an ME-group

did not, at the time at which the relevant event occurred, constitute a group in anything more than the nominal sense, as they did not then interact and, a fortiori, were not then united by a shared goal, meaning that, while the *subjects* may have experienced the event, the *group* could not have done so.[32] This observation suggests two possible views that we might adopt with respect to the application of the appropriate causation condition to group-level representations: first, we might claim that it is simply a mistake to suppose that the appropriate causation condition might apply to group-level representations; second, we might claim that a group-level representation satisfies the appropriate causation condition if the corresponding individual-level representations do so.

If the causal theorist is to have any hope of applying his theory to collective memory in general and to the Mandela Effect in particular, he must adopt something like the latter view.[33] This view, of course, presupposes that the notion of group-level (apparent) memory is to be taken seriously. Given the approach to collective memory outlined in sections 2 and 3, we are entitled to treat a group as (apparently) remembering if new content emerges through the interactions among its members. ME-groups, in particular, can be said to (apparently) remember because, when ME-rememberers encounter each other online, new content indeed

[32] Of course, there is a further reason for which the group could not have experienced the event: the event did not occur. Our point here is that the group could not have experienced the event even if the event did occur.

[33] He must do so, that is, as long as he is wedded to the *classical* causal theory (Martin & Deutscher 1966). As we will see below, if the classical causal theory is replaced by a *constructive* causal theory (Michaelian 2011), it is no longer the case that the group-level representation satisfies the appropriate causation condition if the corresponding individual-level representations do so.

emerges through their interactions, as they borrow details from each other's testimony, with the result that the content of their post-interaction iME-memories differs from that of their pre-interaction iME-memories. The cME-memory that corresponds to their post-interaction iME-memories satisfies the appropriate causation condition, on the view in question, if the latter memories themselves satisfy it, that is, if they are appropriately causally connected to the relevant ME-rememberers experiences of the represented event: appropriate causation is inherited by the group level from the individual level.[34]

The question for the causal theorist is, then, whether post-interaction iME-memories satisfy the appropriate causation condition. We saw in section 5.1.1 that pre-interaction iME-memories are produced by a mechanism similar to that at work in the DRM Effect: transmission of information from experience of the represented event. They therefore appear, given the causal account, to be mismemories. Presumably, if pre-interaction iME-memories are mismemories, then so are post-interaction iME-memories: content is modified via the incorporation of testimonial information during the course of interaction with other group members, but this is a matter of modification, not of wholesale replacement. Since appropriate causation is defined in terms of storage and retrieval of information, post-interaction iME-memories appear, like pre-interaction iME-memories, to be mismemories. If this argument is on the right track, then, because cME-memories satisfy the appropriate causation condition if the corresponding post-interaction iME-memories do so, it would seem

[34] This formulation assumes that all of the post-interaction iME-memories satisfy the appropriate causation condition. It is not clear what we should say about cases in which some but not all of the post-interaction iME-memories satisfy the condition, but we need not deal with this matter here, as there is no reason to suppose that typical cases of the Mandela Effect have this feature.

that the causal theorist must conclude that the Mandela Effect is a form not of collective *confabulation* but rather of collective *misremembering*.

This argument is, however, too quick. The potential problem is not with the claim that post-interaction iME-memories are mismemories if pre-interaction iME-memories are mismemories but rather with the claim that the best explanation of the occurrence of pre-interaction iME-memories is that information is transmitted from experience of the represented event. Information is certainly transmitted; the question is whether it is transmitted from experience of the *represented* event. Whether the content of a subject's present representation of a past event includes content originating in his experience of that event depends in part on how the event is individuated. Consider the *Shazaam* case. If the event is individuated *broadly*—e.g., as there being a children's movie about a genie—then the event occurred. If the event is individuated *narrowly*—as Sinbad's starring in a children's movie about a genie—then the event did not occur. If the event did occur, then, if the subject experienced it, content may well have been transmitted from his experience. But if the event did not occur, then, trivially, the subject did not experience it, and content cannot have been transmitted from his experience. In other words, if the event is individuated sufficiently broadly, then pre-interaction (and hence post-interaction) iME-memories may well satisfy the appropriate causation condition and hence be mismemories; if the event is individuated sufficiently narrowly, then pre-interaction (and hence post-interaction) iME-memories do not satisfy the appropriate causation condition and hence cannot be mismemories. The question for the causal theorist, then, is how broadly or narrowly the relevant events should be individuated.[35]

[35] This is a special case of a more general problem for the causal account of confabulation, one that causal theorists have yet to deal with in any detail. Because the simulation theory

It might, at first glance, seem to be obvious that they should be individuated quite narrowly: what unites an ME-group is, after all, that its members disagree with everyone else about (for example) whether *Sinbad* starred in a children's movie about a genie, not whether there *was* a children's movie about a genie. On closer inspection, however, this option is unattractive. If the events are individuated this narrowly, then iME-memories themselves can no longer be classified as mismemories; instead, they will have to be classified as confabulations (due to the absence of a causal connection). This would, of course, prevent the causal theorist from giving the argument described above. But it would also result in a highly implausible classification of iME-memories, as it would require us to ignore the difference between the relatively minor error characteristic of misremembering (as misremembering is standardly understood) and the major error of confabulation. Consistency, moreover, would presumably require individuating the events at issue in the DRM Effect narrowly as well, in which case not only iME-memories but also DRM-memories would turn out to be confabulations. Indeed, opting for a sufficiently narrow individuation policy might prevent the causal theorist from acknowledging the existence of *any* instances of misremembering at all. The causal theorist ought, then, to opt for broad individuation.

If the relevant events are individuated more broadly, then the causal account implies that iME-memories are mismemories. And, as we have seen, if iME-memories are mismemories, so are cME-memoires: the Mandela Effect is an instance of collective misremembering, not of collective confabulation. The upshot is that the simulationist account and the causal account *disagree* about the group level if they *agree* about the individual level:

---

does not require, in order for successful remember to occur, the existence of an appropriate (content-transmitting) causal connection between the subject's present representation and his past experience, simulation theorists do not face the same problem.

if iME-memories are mismemories, then the causal account implies that cME-memories are mismemories, whereas the simulationist account implies, as we saw in the previous section, that cME-memories are confabulations.

One might object at this point to the claim that post-interaction iME-memories are mismemories if pre-interaction iME-memories are mismemories, that is, that post-interaction iME-memories satisfy the appropriate causation condition if pre-interaction iME-memories do so. There is a natural line of thought according to which, because a post-interaction iME memory depends heavily on the testimony of others about their experiences of the (apparently) remembered event—perhaps, in some cases, more heavily than it does on the subject's own experience of the event—the causal connection between the post-interaction iME-memory and the original experience may be inappropriate even if the causal connection between the corresponding pre-interaction iME-memory and that experience is appropriate. If this is line of thought is right, post-interaction iME-memories—and hence cME-memories—are instances of confabulation rather than misremembering.

The objection assumes that a causal connection between a given present representation and a given past experience is appropriate only if all or at least most of the content of the present representation derives from that of the past experience. This assumption is intuitively plausible, in large part because it aligns with the phenomenology of remembering: one's memories do, after all, present themselves to one as *coming from* one's own past experience.[36] There are nevertheless several persuasive reasons in favour of rejecting the objection.

[36] There is no consensus in the literature on how to describe this aspect of the phenomenology of memory. Tulving (1985) and Klein (2015), for example, describe it in terms of autonoetic (self-knowing) consciousness. Dokic (2014) describes it in terms of an episodic feeling of

First, while it may be true in some cases that *most* of the content of the post-interaction iME memory originates in testimony rather than in the subject's original experience, this need not be true in all cases. Nor is it necessarily true that *much* of that content originates in testimony. In the Sinbad case, for example, it might well be that the subject is the source of the most of the content of his mismemory, with only minor details deriving from the testimony of the members of his ME-group.

Second, even if we restrict our focus to cases in which most of the content of the post-interaction iME memories originates in testimony rather than in the subject's original experience, and even if we assume that this means that the appropriate causation condition is not satisfied at the individual level, this does not necessarily imply that the appropriate causation is not satisfied at the *collective* level.[37] The iME-memories fail to satisfy the condition (if the objection to which we are responding is right) because too much of their content comes in via testimony. But the testimony in question is received from other members of the group. It will thus normally be the case that most of the content of the cME-memory traces back to the experiences of the members of the group. And it is thus not obvious that the appropriate causation condition is not satisfied at the group level.

---

knowing (in contrast to the more familiar semantic feeling of knowing; Koriat 2000), while Perrin (2018) provides an alternative feeling-based account. Mahr and Csibra (2018), and Fernández (2019) provide metarepresentational accounts. Despite this lack of consensus on how to *describe* this feature, it is widely accepted that the phenomenology of memory *has* the feature. Thanks to André Sant'Anna for discussion of this point.

[37] Note that we have been assuming that the cME-memory satisfies the condition if the corresponding iME-memories do so, not that the iME-memories satisfy the condition if the cME-memory does so.

Finally, and most importantly, it is by no means obvious that the fact that most of the content of the post-interaction iME-memories originates in testimony rather than experience means that the appropriate causation condition is not satisfied at the *individual* level. As noted in section 4, the incorporation into memory of information not deriving from the subject's experience of the remembered event, including information available in the context of retrieval, generated by the subject at the time of retrieval, or originating in his experience of other events, including the reception of testimony, is an ordinary and frequent occurrence. Loftus's (2005) work on the Misinformation Effect provides a standard example. In the misinformation paradigm, subjects experience an event and subsequently receive misleading testimonial misinformation about that event, often in the form of questions with misleading presuppositions. (For example, the subject might observe a car accident involving a yield sign and later be asked how fast the car was going when it passed the stop sign.) When they later recall the event, they often incorporate the misinformation into their representation of the event. (So the subject might now recall a stop sign at the scene of the accident.) One question raised by work on the Misinformation Effect is whether the mechanisms responsible for it inevitably lead to decreased accuracy. There is an argument to be made for the view that they do not (Michaelian 2013), but this is not the question that concerns us here. The question that does concern is us whether the apparent memories at issue in the Misinformation Effect are *merely* apparent—that is (taking the causal theory for granted), whether the appropriate causation condition is satisfied. One might be prepared to concede that the condition is satisfied in Misinformation Effect cases, since, in these cases, only minor details of the retrieved representation derive from testimony, without being willing to concede that the condition may be satisfied even in cases in which a large fraction (or even the majority) of the content of the retrieved representation derives from testimony. Given what we know about the way reconstruction in memory works (see Michaelian 2016b), however, cases in

which much (or even most) of the content of the retrieved representation derives from sources other than experience of the represented event are bound to be widespread. There is, moreover, no reason to suppose that the memory system differentiates, when incorporating information deriving from sources other than experience of the represented event, between information originating in testimony and information originating in other sources. It is thus likely that cases in which much (or most) of the content of the retrieved representation derives from testimony are widespread. Insisting that the appropriate causation condition is not satisfied by such representations would therefore require us to concede that merely apparent memories are far more widespread than we ordinarily take them to be.

We can thus set the objection aside: we ought to grant that post-interaction iME-memories satisfy the appropriate causation condition if pre-interaction iME-memories do so. But it would be premature to conclude that, given the causal account, cME-memories are mismemories, for the final point made in response to the objection raises an additional difficulty for that view.

5.2.2 The constructive causal theory

Our discussion so far has taken for granted a fairly *classical* version of the causal theory (i.e., a version close to that developed by Martin and Deutscher 1966), in the sense that it has assumed that, according to the causal theory, what makes the difference between appropriate and inappropriate causation is the storage and retrieval of information originating in the subject's experience of the apparently remembered event. Though the assumption is rarely articulated, partisans of the classical causal theory tend to assume that all or at least most of the content of a genuine memory originates in experience of the remembered event. Once we grant that genuine remembering may occur even when most of the content originates in other sources, however, it becomes considerably less plausible to see the difference between appropriate and inappropriate causation as being determined exclusively by the storage and

retrieval of information originating in the subject's experience of the apparently remembered event. As the case of confabulation itself illustrates (recall Dalla Barba's patient SD), there are many ways in which information originating in other sources might be incorporated into a retrieved representation. Advocates of *constructive* versions of the causal theory have thus argued that the requirement that the retrieved representation be causally connected to the original experience via storage and retrieval of content should be supplemented with an additional requirement, namely, that the representation be produced by a properly functioning and hence reliable memory system (Michaelian 2011). Appropriate causation would thus imply both transmission of content and (where new content is generated) reliable generation of content, transforming the classical causal theory into a *causal reliability* theory.

The alert reader will see where this is going and will wonder whether the suggestion that the classical causal theory be replaced with a causal-reliability theory is not an ad hoc means of ensuring agreement, with respect to the question whether the Mandela Effect is a form of confabulation, between partisans of the simulation theory and partisans of the causal theory. But the suggestion is by no means ad hoc: staunch partisans of the causal theory have themselves acknowledged that it may be necessary, in order to enable the theory to distinguish between memory errors of different types, to add a reliability condition to the theory. Robins (2019), for example, emphasizes the role of malfunction in distinguishing between confabulation and misremembering. And Bernecker (2017), responding to Michaelian's (2016a) claim that, in some cases, confabulation itself may involve the transmission of content, which would render the revised causal taxonomy unable to distinguish between falsidical confabulation and misremembering (since both would satisfy the "appropriate" causation condition and fail to satisfy the accuracy condition) and between veridical confabulation and successful remembering (since both would satisfy both the "appropriate" causation condition and the accuracy condition), points out that the causal

theorist is free to invoke a reliability condition in order to distinguish between the errors in each of these pairs. The suggestion is, moreover, plausible in its own right: if one takes an unbiased look at the literature on confabulation, one cannot, regardless of whether one is a causal theorist, fail to be struck by the fact that confabulators' memory systems appear, in contrast to those of healthy subjects, to be unreliable.

If the classical causal theory is replaced by a causal reliability theory, the revised causal taxonomy will treat remembering and misremembering as being characterized by both transmission and reliability and veridical and falsidical confabulation as being characterized by either nontransmission or unreliability.[38] And if this version of the causal taxonomy is right, then it is no longer the case that the group-level representation satisfies the appropriate causation condition if the corresponding individual-level representations do so. Let us take for granted the claims about individual-level reliability and group-level unreliability developed in section 5.1. If the causal account is right, then, because the appropriate causation condition (understood in part in terms of reliability) is satisfied by both pre- and post-interaction iME-memories, the Mandela Effect amounts to misremembering at the individual level; because the appropriate causation condition (again, understood in part in terms of reliability) is not satisfied by cME-memories, the Mandela Effect amounts to confabulation at the group level. The causal account and the simulationist account thus imply a common classification of cME-memories: given either account, we can conclude that the Mandela Effect is a form of collective confabulation.

[38] The simulation theorist will object, at this point, that, once the reliability condition is added to the account, there is no longer any need for the causal condition. As our aim here is not to settle the debate between the simulation theory and the causal theory, we will not develop this objection in detail or consider responses to it.

6 Conclusions

The conclusion that the Mandela Effect is a form of collective confabulation needs to be qualified, for the "effect" does not appear to be a unified phenomenon. Cases in which, for example, subjects misremember the title of the children's book series, *The Berenstain Bears*, as *"The Berenstein Bears"* or in which they remember that the fictional character Carmen Sandiego wore a yellow trench coat rather than a red one seem to be importantly different from the *Shazaam* case, in which subjects fabricate an entire movie that never existed. There are only so many errors that one might make about the spelling of "Berenstain", making it relatively likely that multiple individuals could independently arrive at the same mismemory. Since their pre-interaction and post-interaction memories are the same, there is no need to appeal to interaction among the misrememberers to explain the fact that they converge on a shared representation of the name. Such cases thus appear to be neither collective nor confabulatory. Indeed, they appear to be instances of merely shared misremembering—they are shared, in the sense that subjects entertain similar representations, but not properly collective, since the fact that subjects enterain similar representations is not due to their interaction.

The introduction of the category of shared misremembering alongside that of collective confabulation suggests the need for a taxonomy of group-level memory errors. From the revised causal taxonomy and the simulationist taxonomy, we have the notions of misremembering and confabulation. From the collective memory literature, we have the distinction between merely shared and genuinely collective memory.[39] Putting these two distinctions together, we might expect to be able to identify four broad types of group-level memory error: in addition to *collective confabulation* (e.g., the core Mandela Effect cases)

[39] See, e.g., Olick 1999 on collect*ive* vs. collect*ed* memory.

and *shared misremembering* (e.g., the *Berenstain/Berenstein Bears* case), there may be instances of *shared confabulation* and *collective misremembering*.[40]

Shared confabulation is unlikely to be an interesting category: since the representations in question are merely shared, they do not arise due to interaction among the subjects in question; given that they are confabulations (and hence, if either the simulationist account or the causal reliability account is right, the product of unreliable memory systems) their similarity would thus have to be due to chance.[41] Collective misremembering may be a more interesting category: as we saw in section 5, the classical causal account can acknowledge the possibility of collective misremembering, and it remains to be determined whether the either the simulationist account or the causal reliability account can likewise do so. These brief remarks are merely provisional. As stated at the outset, in addition to our specific goal of providing an account of the Mandela Effect as a form of collective confabulation, our secondary goal in this paper is to encourage discussion of the concept of collective (or group-level) memory error more broadly. We are therefore content simply to flag shared confabulation and collective misremembering, alongside collective confabulation and shared misremembering, as promising areas for future research.

References

[40] In fact, there are six errors that would need to be taken into account here, since both shared and collective confabulation might be either falsidical or veridical. We would also need to take the distinction between shared and collective (successful) remembering into account, giving us a taxonomy including eight states in total.

[41] Recall the "naïve sense" of collective confabulation introduced above.

Arango-Muñoz, S., & Michaelian, K. (2020). From collective memory … to collective metamemory? In A. Fiebich (Ed.), *Minimal Cooperation and Shared Agency* (pp. 195–217). Springer.

Barash, J. A. (2017). Collective memory. In S. Bernecker & K. Michaelian (Eds.), *The Routledge Handbook of Philosophy of Memory* (pp. 255–267). Routledge.

Baysan, U. (2018). Memory, confabulation, and epistemic failure. *Logos & Episteme*, 9(4), 369–378.

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

Bernecker, S. (2008). *The Metaphysics of Memory*. Springer.

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

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.

Betts, K. R., & Hinsz, V. B. (2010). Collaborative group memory: Processes, performance, and techniques for improvement. *Social and Personality Psychology Compass*, 4, 119–30.

Bortolotti, L., & Cox, R. E. (2009). "Faultless" ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, 18(4), 952–965.

Bortolotti, L., & Sullivan-Bissett, E. (2018). The epistemic innocence of clinical memory distortions. *Mind & Language*, 33(3), 263–279.

Brewer, W. F., & James C. Treyens. (1981). Role of schemata in memory for places. *Cognitive Psychology* 13(2), 207–230.

Broome, F. (2009). Mandela effect - alternate realities. https://mandelaeffect.com/. Accessed: 19/01/2019.

Broome, F. (2013). Billy Graham's funeral on TV. https://mandelaeffect.com/billy-grahams-funeral-on-tv/. Accessed: 25/05/2019.

Broome, F. (2016). DiCaprio wins…Again? https://mandelaeffect.com/dicaprio-wins-again/ Accessed 19/01/2019.

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

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

Dokic, J. (2014). Feeling the past: A two-tiered account of episodic memory. *Review of Philosophy and Psychology*, *5*(3), 413–426.

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5), 729–745.

Feinberg, T. E. 2001. *Altered Egos: How the Brain Creates the Self*. Oxford: Oxford University Press.

Fernández, J. (2015). What are the benefits of memory distortion? *Consciousness and Cognition*, 33, 536–547.

Fernández, J. (2018). The functional character of memory. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions in the Philosophy of Memory* (pp. 52–71). New York: Routledge.

Fernández, J. (2019). *Memory: A Self-Referential Account*. Oxford University Press.

French, A. (2018). The Mandela effect and new memory. *Correspondences: Journal for the Study of Esotericism*, 6(2), 201-233.

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

Garrett, R. K. (2017). The “echo chamber” distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, 6(4), 370–376.

Goldman, A. I. (2012). *Reliabilism and Contemporary Epistemology: Essays*. Oxford University Press.

Hakli, R. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7(2–3), 286–927.

Hakli, R. (2007). On the possibility of group knowledge without belief. *Social Epistemology*, 21(3), 249–266.

Harris, C. B., Barnier, A. J., Sutton, J., & Keil, P. G. (2014). Couples as socially distributed cognitive systems: Remembering in everyday social and material contexts. *Memory Studies*, 7(3), 285–297.

Harris, C. B., Barnier, A. J., Sutton, J., Keil, P. G., & Dixon, R.A. (2017). “Going episodic: Collaborative inhibition and facilitation when long-married couples remember together. *Memory*, 25(8), 1148–1159.

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

Holt, D. (2018). Are we in a simulation & is the Mandela Effect real? *Mr. Futurist*. https://mrfuturist.com/are-we-in-a-simulation-is-the-mandela-effect-real/. Accessed 19/01/2019.

Huebner, B. (2014). *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford University Press.

Huebner, B. (2016). Transactive memory reconstructed: Rethinking Wegner’s research program. *Southern Journal of Philosophy*, 54(1), 48–69.

Hutto, D. D., & Peeters, A. (2018). The roots of remembering: Radically enactive recollecting. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions in the Philosophy of Memory* (pp. 97–118). Routledge.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4), 407-424.

Klein, S. B. (2015). What memory is. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(1), 1–38.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*(2), 149-171.

Lewandowsky, S., Ullrich K. H. E., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era." *Journal of Applied Research in Memory and Cognition*, 6(4), 353–69.

Loftus, E. F. (1997). Memory for a past that never was. *Current Directions in Psychological Science*, 6(3), 60–65.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4), 361–66.

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

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review*, 75(2), 161–96.

Maswood, R., & Rajaram, S. (2019). Social transmission of false memory in small groups and large networks. *Topics in Cognitive Science*, 11(4), 687-709.

McDowell, J., & AE911Truth Staff. (2015). 60 structural engineers cite evidence for controlled demolition. Architects & Engineers for 9/11 Truth.

https://www.ae911truth.org/evidence/technical-articles/articles-by-ae911truth/199-60-structural-engineers.

Meade, M. L., Nokes, T. J., & Morrow, D. G. (2009). Expertise promotes facilitation on a collaborative memory task. *Memory*, 17, 39–48.

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*, 24(3), 323–42.

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.

K. Michaelian. (Forthcoming). Confabulation: False belief, causalist, epistemic, explanationalist, and simulationist accounts. In A. Sant'Anna, C. McCarroll, & K. Michaelian (Eds.), *Current Controversies in Philosophy of Memory*. Routledge.

Michaelian, K., & Arango-Muñoz, S. (2018). Collaborative memory knowledge: A distributed reliabilist perspective. In M. Meade, C. B. Harris, P. Van Bergen, J. Sutton, & A. J. Barnier (Eds.), *Collaborative Remembering: Theories, Research, Applications* (pp. 231–247). Oxford University Press.

Michaelian, K., Perrin, D., & Sant'anna, A. (2020). Continuities and discontinuities between imagination and memory: The view from philosophy. In A. Abraham (Ed.), *The Cambridge Handbook of the Imagination* (pp. 293–310). Cambridge University Press.

Michaelian, K., & Sant'Anna, A. (Forthcoming). Memory without content? Radical enactivism and (post)causal theories of memory. *Synthese*.

Michaelian, K., & Sutton, J. (2018). Collective memory. In M. Jankovic & K. Ludwig (Eds.), *The Routledge Handbook of Collective Intentionality* (pp. 140–151). Routledge.

Michaelian, K., & Sutton, J. (Forthcoming). Collective mental time travel: Remembering the past and imagining the future together. *Synthese*.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–59.

Olick, J. K. (1999). Collective memory: The two cultures. *Sociological theory*, 17(3), 333-348.

Perrin, D. (2018). A case for procedural causality in episodic recollection. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions in the Philosophy of Memory* (pp. 33–51). Routledge.

Perrin, D., & Michaelian, K. (2017). Memory as mental time travel. In S. Bernecker & K. Michaelian (Eds.), *The Routledge Handbook of Philosophy of Memory* (pp. 228–239). Routledge.

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

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, 5, 649–63.

Rehg, W., & Staley, K. (2008). The CDF collaboration and argumentation theory: The role of process in objective knowledge. *Perspectives on Science*, 16(1), 1–25.

Ren, Y., & Argote, L. (2011). Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences. *Academy of Management Annals*, *5*(1), 189-229.

Robin, F. (2010). Imagery and memory illusions. *Phenomenology and the Cognitive Sciences*, 9, 253–262.

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

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*. Oxford University Press. Second edition.

Shanton, K. (2011). Memory, knowledge and epistemic competence. *Review of Philosophy and Psychology*, 2(1), 89–104.

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

Staley, K. W. (2007). Evidential collaborations: Epistemic and pragmatic considerations in "group belief". *Social Epistemology*, 21(3), 321–35.

Sutton, J. (1998). *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge University Press.

Tait, A. (2016). "The movie that doesn't exist and the Redditors who think it does." *New Statesman*. https://www.newstatesman.com/science-tech/internet/2016/12/movie-doesn-t-exist-and-redditors-who-think-it-does. Accessed 19/01/2019.

Talland, G. A. 1961. Confabulation in the Wenrnicke-Korsakoff syndrome. *Journal of Nervous and Mental Disease*, 132: 361–381.

Talland, G. A. 1965. *Deranged Memory*. New York: Academic Press.

Tanesini, A. (2018). Collective amnesia and epistemic injustice. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos, & D. Pritchard (Eds.), *Socially Extended Epistemology* (pp. 195–219). Oxford University Press.

Theiner, G. (2013). Transactive memory systems: A mechanistic analysis of emergent group memory. *Review of Philosophy and Psychology*, 4, 65–89.

Theiner, G. (2018). Groups as distributed cognitive systems. In M. Jankovic & K. Ludwig (Eds.), *The Routledge Handbook of Collective Intentionality* (pp. 233–248). Routledge.

Theiner, G., & Sutton, J. (2014). The collaborative emergence of group cognition. *Behavioral and Brain Sciences*, *37*(3), 277–278.

Tollefsen, D. P. (2015). *Groups as Agents*. Polity Press.

Tuomela, R. (1992). Group beliefs. *Synthese*, 91(3), 285–318.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne*, 26(1), 1-12.

u/AscendedMinds. (2017). "Scientists believe Parallel Universes ARE interacting. Is this the cause of the 'Mandela Effect'?" Reddit r/MandelaEffect. https://www.reddit.com/r/MandelaEffect/comments/6eq5k7/scientists_believe_parallel_universes_are/. Accessed 19/01/2019.

u/AutoModerator. (2018). Did you discover a possible new Mandela effect? Post it here! (weekly discussion) (2018-06-24)." Reddit r/MandelaEffect. https://www.reddit.com/r/MandelaEffect/comments/8tfyqc/did_you_discover_a_possible_new_mandela_effect/. Accessed on 19/01/2019.

u/Denominax. (2017). Mandela effect wiki. Reddit r/MandelaEffect. May 18, 2017. https://old.reddit.com/r/MandelaEffect/wiki/index. Accessed on 19/01/2019.

u/Destielluh. (2017). "Shazam / Shazaam with Sinbad was real and here is all the movie information." Reddit r/Shazaam. https://www.reddit.com/r/Shazaam/comments/5m8o02./shazam_shazaam_with_sinbad_was_real_and_here_is/dfwk010. Accessed: 19/01/2019.

u/DonnaGail. (2017). Shazam / Shazaam with Sinbad was real and here is all the movie information." Reddit r/Shazaam.

https://www.reddit.com/r/Shazaam/comments/5m8o02/shazam_shazaam_with_sinbad_was_real_and_here_is/ddjzs8p. Accessed on 19/01/2019.

u/EpicJourneyMan. (2016). “The Sinbad genie movie - Complete analysis.” Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/55f5rt/the_sinbad_genie_movie_complete_analysis/. Accessed on 19/01/2019.

u/EpicJourneyMan. (2018). “All of my coworkers remember Shazam.” Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/7nhdjt/all_of_my_coworkers_remember_shazam/ds73ebi. Accessed on 19/01/2019.

u/ezydown. (2017). 254 confirmed mandela effects: List. Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/6p6zwd/254_confirmed_mandela_effects_list/dknskwm/. Accessed on 19/01/2019.

u/Fae_Leaf. (2018). All of my coworkers remember Shazam. Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/7nhdjt/all_of_my_coworkers_remember_shazam/. Accessed on 19/01/2019.

u/Jhoobie. (2017). What is the Mandela effect? Reddit r/OutOfTheLoop. https://www.reddit.com/r/OutOfTheLoop/comments/5m644b/what_is_the_mandela_effect/dc1ib3d. Accessed on 19/01/2019.

u/manafrmhvn. (2018). All of my coworkers remember Shazam. Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/7nhdjt/all_of_my_coworkers_remember_shazam/ds6josi. Accessed on 19/01/2019.

u/melossinglets. (2017). 254 confirmed Mandela effects: list.” Reddit r/Mandela Effect. https://www.reddit.com/r/MandelaEffect/comments/6p6zwd/254_confirmed_mandela_effects_list/dko45es. Accessed on 19/01/2019.

u/shazaamthemovie. (2017). “Shazam / Shazaam with Sinbad was real and here is all the movie information.” Reddit r/Shazaam. https://www.reddit.com/r/Shazaam/comments/5m8o02/shazam_shazaam_with_sinbad_was_real_and_here_is/. Accessed on 19/01/2019.

u/ThadeusOfNazereth. (2016). “What is going on with Mother Teresa?” Reddit r/MandelaEffect. https://www.reddit.com/r/MandelaEffect/comments/516nen/what_is_going_on_with_mother_teresa/. Accessed on 19/01/2019.

u/TimmyTheShpee. (2018). “Monopoly man, I swear to god he has a monocle.” Reddit r/MandelaEffect. https://www.reddit.com/r/MandelaEffect/comments/9mhqny/monopoly_man_i_swear_to_god_he_has_a_monocle/. Accessed 19/01/2019.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen, & G. R. Goethals (Eds.), *Theories of Group Behavior* (pp. 185– 208). Springer.

Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology*, 61, 923– 9.

Weldon, M. S. (2000). Remembering as a social process. *Psychology of Learning and Motivation*, 40, 67–120.

Wray, K. B. (2001). Collective belief and acceptance. *Synthese*, 129(3), 319–33.