MEMORY, CONFABULATION, AND EPISTEMIC FAILURE

Umut BAYSAN

ABSTRACT: Mnemonic confabulation is an epistemic failure that involves memory error. In this paper, I examine an account of mnemonic confabulation offered by Sarah Robins in a number of works. In Robins' framework, mnemonic cognitive states in general (e.g., remembering, misremembering) are individuated by three conditions: existence of the target event, matching of the representation and the target event, and an appropriate causal connection between the target event and its representation. Robins argues that when these three conditions are not met, the cognitive state in question is an instance of mnemonic confabulation. Here, I argue that this is not true. There are mnemonic cognitive states which don't meet any of these conditions, and they are not cases of mnemonic confabulation. On a more positive note, I argue that mnemonic confabulation requires it to be a failing on behalf of either the subject or her mnemonic system that these conditions are not met.

KEYWORDS: confabulation, epistemic failure, memory, misremembering

1. Introduction

Confabulation is an epistemic failure, and in paradigmatic cases, it involves failure of remembering. In what ways a subject's remembering has to fail in order for her to count as confabulating a memory is a venue for philosophical debate. This paper aims to contribute to this debate. In what follows, I examine an account of confabulation proposed by Sarah Robins,[1] and argue that although the background philosophical framework that Robins has developed is commendable, her treatment of confabulation yields some counterintuitive results. In her work, Robins focuses on *mnemonic confabulation* (instead of confabulation *simpliciter*), and here I will follow her in doing that (except when I explicitly state otherwise).

[1] Sarah Robins, "Misremembering," *Philosophical Psychology* 29 (2016): 432-447, "Confabulation and Constructive Memory," *Synthese* (2017): 1-17, and "Mnemonic Confabulation" (unpublished manuscript).

2. Robins on Mnemonic Cognitive States

Mnemonic confabulation is the type of confabulation that involves a memory error. To understand Robins' treatment of mnemonic confabulation, let's explore the ways she contrasts mnemonic confabulation with other mnemonic cognitive states.

The paradigm case of a mnemonic cognitive state is successfully remembering a past episode. (Taking remembering to be a factive mental state, I will henceforth omit "successfully."[2]) Suppose Jude remembers that Sue bought him cufflinks for his 31st birthday. Let's say that Jude very vividly remembers the occasion with its relevant details: its being his birthday, and in fact being his 31st birthday, that Sue gave him a navy blue box just when the main dish was being served, that he opened the box and there were a pair of cufflinks, and so on. According to Robins, this instance of remembering involves, *first*, the fact that the target event did take place (i.e., Jude had a 31st birthday which he celebrated with his partner, and Sue gave him cufflinks on that very occasion); *second*, that Jude has a mental representation of the target event in a way that matches the target event with respect to its relevant aspects (i.e., the event is represented as a birthday celebration event, as well as a Sue-gifting-Jude-cufflinks event); and *third*, the right kind of causal connection between the target event and the mental representation of the target event (i.e., Jude's birthday celebration and the cufflinks-giving event are among the causal antecedents of the representation). Let's call these three conditions TARGET, MATCHING, and CAUSATION respectively.

The first two conditions (TARGET and MATCHING) can account for the fact that remembering is a factive mental state: the mental state that underlies remembering correctly represents the remembered event which indeed took place in the way it is represented. The third condition (CAUSATION) rules out cases where the representation and the target event match, but not for the right reasons. Suppose it is true that Sue bought Jude cufflinks for his 31st birthday, but Jude never came to know this. But a demon implants a chip in his brain which physically realizes a mnemonic mental representation that has the relevant aspects of the target event. Although he has an accurate representation of a past event, this

[2] I assume, without argument, that remembering is factive. I acknowledge that this could be debated.

shouldn't count as remembering. CAUSATION rules out this case from being a case of remembering.

Another type of mnemonic cognitive state is misremembering. Imagine that Jude seems to remember that Sue bought him cufflinks with blue prints on them, whereas in fact the prints were burgundy. Apart from this difference, let's say that Jude's mental representation of his 31st birthday is accurate. Although, in this case, he still remembers certain aspects of the target event, overall, his cognitive state counts as misremembering. In this case, his representation still picks out the target event (i.e., his 31st birthday celebration), but it misrepresents some of its aspects. In cases of misremembering, despite the satisfaction of TARGET (i.e., the target event exists), MATCHING is not satisfied (i.e., the content and the target event do not match). Thus, misremembering is not factive. The difference between remembering and misremembering is similar to the difference between veridical perceptions and perceptual illusions.[3] In both veridical perceptions and perceptual illusions, a target object does exist, but whereas in the former, the target object is perceived as it is (i.e., it does have the sensory properties it appears to have), in the latter, the target object is misperceived (i.e., it doesn't have some of the sensory properties it appears to have).

Now, we are in a position to understand Robins' account of mnemonic confabulation. Mnemonic confabulation is also a mnemonic cognitive state, and it is different from both remembering and misremembering in important aspects. Suppose that Sue did not buy Jude anything for his 31st birthday, and in fact, they did not even have any celebration. Now suppose Jude seems to have a memory of Sue giving him cufflinks on his 31st birthday. So, he has a mental representation of an event which has the aspects of a 31st birthday event, a celebration dinner event, a cufflinks-gifting event, and so on. According to Robins, this would be an example of mnemonic confabulation. This state is importantly different from both remembering and misremembering. Unlike in cases of both remembering and misremembering, the target event does not exist. There is no 31st birthday celebration regardless of whether cufflinks were given or not. So, TARGET is not satisfied. Given that the target event does not exist, MATCHING is not satisfied either: the target event and the content of the representation do not match (simply because the target event doesn't exist). And furthermore, since these two conditions are not met, CAUSATION fails also: there is no right kind of causal relation between a target event and the content of the representation. Thus,

[3] Robins, "Confabulation and Constructive Memory."

mnemonic confabulation differs from remembering and misremembering in the sense that all three conditions for mnemonic cognitive states fail to be satisfied (whereas in remembering all three are satisfied, and in misremembering, at least the first one is satisfied). Here, Robins compares mnemonic confabulation to cases of perceptual hallucinations. In perceptual hallucinations, there is no target object although there is a representation of an object with some sensory properties. Likewise, in mnemonic confabulation, there is no target event despite the fact that there is a representation of an event with certain aspects.

What we see here is an elegant framework which locates different sorts of mnemonic cognitive state in one table. Mnemonic states are representational states, and in individuating different kinds of mnemonic cognitive states, the relevant parameters are TARGET, MATCHING, and CAUSATION. The account is also in line with a more general framework according to which memory and other mnemonic states are understood in terms of causation.[4] The following falls out from Robins' account as a characterisation of mnemonic confabulation:

> **(MC)** A subject mnemonically confabulates some putative past event if and only if she has a mnemonic representation which meets none of TARGET, MATCHING, and CAUSATION.[5]

Although this is an elegant framework and arguably successful in explaining contrasting remembering and misremembering, I believe that **(MC)** yields counterintuitive results, which I shall highlight next.

3. Some Counterintuitive Results

I believe that a central aspect of confabulation is missing in **(MC)**. I will illustrate this by giving an example which would count as mnemonic confabulation according to **(MC)**, and then argue that it shouldn't. That will show that being a mnemonic mental representation that satisfies none of TARGET, MATCHING, and CAUSATION is *not sufficient* for being an instance of mnemonic confabulation.

[4] See also Sven Bernecker, "A Causal Theory of Mnemonic Confabulation," *Frontiers in Psychology* 8 (2017): 1205.

[5] Or when it does meet TARGET, it is purely accidental that it does. Bernecker's (*ibid.*) account of confabulation emphasises this possibility. He suggests that "a piece of confabulation may even be entirely correct. It is possible that a patient fantasizes correctly by telling a story that, by sheer luck, represents the objective reality" (*ibid.*, 5). In fact, Bernecker uses this and similar considerations to argue that the real mark of mnemonic confabulation is the failure of what I have in this paper called CAUSATION.

After considering and responding to possible ways this objection could be replied, I will remark on what I think is missing from this account.

Consider the following (very dull) story which I shall call *the flapjack case.* At t_1, I am sitting in a café, sipping my coffee, and it appears to me that there is a piece of flapjack on a plate on the table opposite of me. Actually, there is no flapjack on the table, and in fact there is nothing on the table. So, I am hallucinating a piece of flapjack on the opposite table. (It doesn't matter what causes this hallucination; to fix ideas, let's suppose it is a malicious demon behind this very uninteresting trick.) There is nothing suspicious about there being some flapjack on a table; I am in a café after all, and many cafes do serve flapjack. So, I have no reason to doubt the veridicality of this experience or reflect much on it. Days pass, now the time is t_2. I am sitting in another café, some stranger approaches and offers me a flapjack. This prompts me to recall an experience I had recently. Then I form the mental representation of a flapjack-on-the-opposite-table event that happened at t_1.

According to **(MC)**, the mnemonic cognitive state I am in at t_2 should count as mnemonic confabulation. The target event doesn't exist; there wasn't a flapjack-on-the-opposite-table event. Since there was no such event at t_1, the content of the mental representation at t_2 doesn't match a target event at t_1; and for the same reason, there is no right kind of causal connection between an event at t_1 and the representation at t_2. So, there is a mental representation that doesn't meet any of TARGET, MATCHING, and CAUSATION. However, intuitively, it is not right that this case is a case of mnemonic confabulation. Therefore, **(MC)** doesn't capture the essence of mnemonic confabulation.

One might think that the target event does exist in the flapjack case; it is just not a flapjack-on-the-opposite-table event. As a sitting-in-café event at t_1, the target event does exist, it might be argued. Whether this response is viable partly depends on how to individuate events. If the target event is essentially a café event, then the target event indeed exists. But in the representation of this event at t_2, the salient feature of it is that it is a flapjack-on-the-opposite-table event, which suggests that it is more appropriate to take it as an essentially flapjack-on-the-opposite-table event. Nevertheless, even if the target event is essentially a sitting-in-café event and hence that the target event at t_1 does exist, Robins' account runs into a different problem. For the sake of entertaining this response, let's accept that the target event does exist at t_1 as a sitting-in-café event, but it is misrepresented at t_2 as a flapjack-on-the-opposite-table event. From the characterisation of

mnemonic cognitive states Robins gives, this state then should be categorised as misremembering: TARGET is satisfied, MATCHING is not. But this is an equally implausible consequence. Misremembering should be a failure of remembering. Here, there is nothing that indicates that the failure has anything to do with remembering. The event at t_1 has always been a flapjack event for me.

4. Further Possible Responses

We have seen an example of a mnemonic cognitive state which fails all three conditions TARGET, MATCHING, and CAUSATION, yet it makes little sense to categorise it as a case of mnemonic confabulation. Before remarking on what I think is missing from this example to make it a case of confabulation, let me address two possible responses that Robins can offer.

First, Robins can argue that, in the flapjack case, the target event does indeed take place. This is not because there is a sitting-in-café event (*à la* the misremembering response discussed above), but it is due to the fact that the target event is a sensory experience. That is, it is true that there is no external object (i.e., flapjack) at t_1, but there is a sensory object that exists at t_1. So, Robins can argue that the flapjack case doesn't count as mnemonic confabulation according to **(MC)** because TARGET is satisfied. Let's call this *the sensory event response*.

I don't think that the sensory event response is a satisfactory one. Why posit internal sensory objects just to get around this particular type of counterexample? If we are to posit an internal sensory object to explain the flapjack case, why not take the target event in cases of remembering also as internal sensory events? If we are to hold that when I remember that I saw a deer in the forest what I remember is not a deer but instead an internal sensory object (a deer-like sense-datum), why not also hold that when I take myself to see a deer in the forest, what I in fact perceive is a deer sense-datum? This is not the place to give an argument against the sense-data theory, but it is worth noting that this response comes with the burden of making a case for sense-data.

Regardless of any qualms about the metaphysics of sensory objects, it is not clear that it is a good move to suggest that target events are sensory events in *all* mnemonic cognitive states (which we should do if we want to make the sensory event response sound less *ad hoc*). We would then be holding that, in a case of remembering, there is an external event, which then causes a sensory event, which in turn is accurately represented (and representation is caused in the right way by the sensory event). What is problematic with this is that the representation would

be a *representation of a sensory event*, *not a representation of an external event*. Introspectively, I find this hard to believe. On a more theoretical point, it is problematic to think that at every time I seem to remember a physical event *e*, I actually remember a sensory event *e**, but I take myself to remember *e* (or I am in a position to take myself to remember *e*). If remembering, as a mental state kind, is to accommodate this possibility, it shouldn't be categorised as a factive mental state kind. There surely are cases when we recall our sensory experiences; but when we do so, we remember them *as sensory experiences*. If we don't remember them as sensory experiences, we don't remember them *simpliciter*. I believe these difficulties make it very difficult to motivate the sensory event response.

So much for the sensory event response. What about Robins' second possible move? Robins can bite the bullet and hold that the flapjack case is indeed a case of mnemonic confabulation. However, note how different this case would be from more paradigmatic cases of mnemonic confabulation. In the paradigmatic cases of mnemonic confabulation, the inaccurate representation is a failing on part of the mnemonic system. Mnemonic confabulation is a failure of remembering. In the flapjack case, there is no failure of remembering. If there is any failure, it has to do with the forming of the experience at t_1 in the first place.

I believe what we have just seen reveals what is missing in the characterisation of mnemonic confabulation in **(MC)**. It is evident (from the fact that she compares mnemonic confabulation to perceptual hallucinations) that Robins (rightly) treats mnemonic confabulation as an epistemically unideal kind of mental state. Mnemonic confabulation involves a form of failure. However, **(MC)** doesn't give us any clue as to where that failure lies.

5. Normativity in Confabulation

I mentioned that I am following Robins in focusing on mnemonic confabulation rather than confabulation *simpliciter*. In this section and next, I shall relate the discussion so far to the concept of confabulation, broadly understood. My intuitions regarding the flapjack case are motivated by the fact that I take mnemonic confabulation to be a specific kind of confabulation. I think a theory of mnemonic confabulation would be unattractive if it couldn't accommodate the fact that mnemonic confabulation is a type confabulation. But what is confabulation more generally?

The nature of confabulation is a matter of dispute among philosophers of cognitive science, philosophers of psychiatry, psychiatrists, and others.[6] Whereas some researchers restrict the term "confabulation" to epistemically problematic mental states which have to do with memory,[7] others take confabulation to be a more general epistemic failure which involves false beliefs regardless of whether these false beliefs concern putative past events or not.[8] But surely, not every false belief counts as confabulation.[9] If it did, then why need the category of confabulation over and above the category of false belief? Then, what is the additional component in confabulation on top of a false belief?

Researchers seem to agree that one of the things that mark the difference between a merely false belief and a confabulatory mental state is that in the latter, there is failing on behalf of either the subject or the subject's mnemonic system where there ideally shouldn't be. Turnbull and colleagues suggest that, in confabulation, "false beliefs and opinions about the world … are manifestly incorrect."[10] Being manifestly incorrect, these beliefs are beliefs that the subject should not have formed or retained. Hirstein suggests that when a subject confabulates that P, her belief that P is ill-grounded and moreover that subject should (but does not) know that her belief is ill-grounded.[11] These suggest that confabulation, if it is to be separated from a merely false belief, involves a normative element.

What do I mean by a normative element? When I say that a confabulated belief is a belief that should not have been formed or retained, am I suggesting that the subject had an obligation not to form that belief? If the idea of obligation is

---

[6] For discussion, see Lisa Bortolotti's *Delusions and other Irrational Beliefs* (Oxford: Oxford University Press, 2010), 43-50.

[7] For example, Aikaterini Fotopoulou, "False-Selves in Neuropsychological Rehabilitation: The Challenge of Confabulation," *Neuropsychological Rehabilitation* 18 (2008): 541-565.

[8] Oliver H. Turnbull, Sarah Jenkins, and Martina L. Rowley, "The Pleasantness of False Beliefs: An Emotion-Based Account of Confabulation," *Neuro-Psychoanalysis* 6 (2004): 5-45, William Hirstein, *Brain Fiction: Self-deception and the Riddle of Confabulation* (Cambridge, MA: MIT Press, 2005), Linda Örulv and Lars-Christer Hydén, "Confabulation: Sense-Making, Self-Making and World-Making in Dementia," *Discourse Studies* 8 (2006): 647-673, and G.E. Berrios, "Confabulations", in *Memory Disorders in Psychiatric Practice*, eds. G.E. Berrios and J.R. Hodges (New York, NT: Cambridge University Press, 2000), 348-368.

[9] Also, as noted in footnote 5, it is possible for the content of a confabulation to be accidentally true. I may confabulate that P whereas P happens to be true (as in Gettier cases).

[10] Turnbull, Jenkins, and Rowley, "The Pleasantness of False Beliefs," 6.

[11] Hirstein, *Brain Fiction*, 187.

linked to that of responsibility, does this mean that, when S confabulates a belief, S *could have* believed otherwise? Ideally, I would like not to make any of these commitments. After all, it is plausible that, at least in some cases of confabulation in the clinical population, subjects could not have done otherwise. This suggests that the sense of normativity here is different from the sense of normativity that underlies moral responsibility. Nevertheless, it is clear that, in cases of confabulation (likewise in cases of delusions and irrational beliefs), there is a sense in which either subjects or their mnemonic systems depart from some epistemic norms.

Could CAUSATION in Robins' account not be viewed as a normative requirement? In cases of remembering, representations must be *appropriately* caused. It might be thought that the appropriateness of the causal connection could underwrite the normative element that I argue is missing in Robins' account. The problem with this suggestion is that, in **(MC)**, CAUSATION fails purely in virtue of the failure of TARGET. Its failure has nothing to do with the causal connection appropriateness. So, overall, I don't think that **(MC)** captures the required normativity.

6. Concluding Remarks and a Proposal

It strikes me as evident that **(MC)** fails to have the normative element required from an account confabulation. In concluding, let me highlight four possible ways this may have bearing on Robins' account. First, and least desirably, one could just argue that because mnemonic confabulation, as discussed above, fails to be a form of confabulation due to failing to meet a normativity criterion, Robins' framework should be abandoned altogether. Although I am mentioning this possibility, let me make it explicit that this is not the recommendation I am making; there are less radical ways to resolve the issue at hand. Second, and less undesirably, one could just accept that mnemonic confabulation is a very different type of cognitive state compared to confabulation simpliciter. At this point, it may be merely a terminological dispute as to whether mnemonic confabulation should be called as such. Third, and relatedly, one could argue that mnemonic confabulation is a type of confabulation, but the class of confabulatory mental states are very diverse. If this is the preferred option, one should also be prepared to respond to some worries with respect to whether confabulation, as a mental state kind, is a natural kind or not. And finally, one could agree with the message of the previous section, and accept that confabulation has to have a normative element. If mnemonic

confabulation is a type of confabulation, then mnemonic confabulation must have a normative element too. If that is the case, the spirit of Robins' account can be retained, but can be supplemented with a normative criterion. One way of doing so would be to hold the following:

> **(MCN)** A subject mnemonically confabulates some putative past event if and only if she has a mnemonic representation which meets none of TARGET, MATCHING, and CAUSATION, and it is a failing on behalf of either the subject or her mnemonic system that none of these conditions is met.[12]

This might not be the only way the problem I have highlighted could be solved, but it is one way of solving it, and I hope it is helpful way of doing so.[13]

[12] (**MCN**) is an account of mnemonic confabulation, but admittedly it fails to accommodate the possibility of veridical mnemonic confabulation, a possibility that one might want to consider as per footnotes 5 and 9 above. To get around this problem, we can add a disjunctive clause in **(MCN)** to the effect that when TARGET and MATCHING are met, it is only accidental (as in Gettier cases) that they are met.

[13] Thanks to Kathy Puddifoot for her helpful comments on a previous version of this paper.