Two Informational Theories of Memory: a case from Memory-Conjunction Errors

Danilo Fraga Dantas
Federal University of Santa Maria (UFSM)
First draft, forthcoming in Disputatio

Abstract
The causal and simulation theories are often presented as very distinct views about declarative memory, their major difference lying on the causal condition. The causal theory states that remembering involves an accurate representation causally connected to an earlier experience (the causal condition). In the simulation theory, remembering involves an accurate representation generated by a reliable memory process (no causal condition). I investigate how to construe detailed versions of these theories that correctly classify memory errors (DRM, "lost in the mall", and memory-conjunction errors) as misremembering or confabulation. Neither causalists nor simulationists have paid attention to memory-conjunction errors, which is unfortunate because both theories have problems with these cases. The source of the difficulty is the background assumption that an act of remembering has one (and only one) target. I fix these theories for those cases. The resulting versions are closely related when implemented using tools of information theory, differing only on how memory transmits information about the past. The implementation provides us with insights about the distinction between confabulatory and non-confabulatory memory, where memory-conjunction errors have a privileged position.

Keywords
Episodic memory, memory-conjunction errors, causal theory of memory, simulation theory of memory, information theory.

1 Introduction

In this paper, I use memory errors to examine the main theories of declarative memory: causal theory (Bernecker 2010; Robins 2016b), which states that remembering requires having an accurate representation standing in a causal connection with an earlier

experience (the causal condition), and simulation theory (Michaelian 2016b), which states that remembering requires having an accurate representation generated by a reliable memory process (no causal condition). These theories are often presented as very distinct views about memory, their major difference lying on the causal condition. I start by discussing how to model different versions of these theories using tools of theory of reference. The causal theory is presented in two versions: direct (Bernecker 2010) and indirect (Robins 2016b). The simulation theory is only presented in an indirect version (Michaelian 2016b). Then I test how the resulting models classify memory errors as misremembering (inaccurate memories) or confabulation (roughly, fabricated memories).

The errors I survey result from three experimental paradigms: DRM (Deese 1959; Roediger and McDermott 1995), "lost in the mall" (Loftus and Pickrell 1995), and memory-conjunction errors (Reinitz et al. 1992). The first two errors are used to discuss minor difficulties for the theories; the main action comes from the analysis of memory-conjunction errors. Neither causalists nor simulationists have paid due attention to memory-conjunction errors. As a result, the existing versions of these theories have difficulties with these cases. Memory-conjunction errors are cases of misremembering due to ambiguity in target selection, but both theories share the background assumption that an act of remembering has one (and only one) target. The background assumption is reflected in the accuracy condition of the (direct and indirect) causal and simulation theories and in the counterfactual causal condition of (some) causal theories.

The surveyed direct causal theory cannot be fixed for the problem cases because of both its direct and counterfactual features. Robins' theory and Michaelian's theories, on the other hand, can be worked out. The accuracy condition should be modeled using a plural target selection function and the causal condition must be reduced to a probabilistic (minimal) condition. The resulting theories are closely related when implemented using tools of information theory, differing only on how memory transmits information about the past. The implementation provides us with insights about the distinction between

confabulatory and non-confabulatory memory, where memory-conjunction errors have a privileged position.

In §2, I use tools of the theory of reference to model the accuracy condition of the causal (direct and indirect) and simulation theories. The causal and reliability conditions are modeled using counterfactuals and probabilities (respectively). I survey how Robins and Michaelian distinguish misremembering from confabulation. In §3, I survey the experimental results of DRM, "lost in the mall", and memory-conjunction errors and investigate how the causal and simulation theories classify these errors as misremembering or confabulation. In §4, I fix the causal and simulation theories for the problem cases and implement the resulting theories using tools of information theory.

2 Setting the stage

Declarative memory is the capacity of retrieving information that can, at least in principle, be brought to consciousness. Researchers of the field usually agree that declarative memory has two poles. Bernecker (2010: 14) distinguishes object-, property-, and event-memory from propositional memory; Robins (2020: 122) and Michaelian (2016b: 31) work with the standard distinction between episodic and semantic memory. These classifications use different criteria and exhibit categories with different intensions and extensions. For example, while the distinction between object-, property-, and event-memory and propositional memory builds upon the targets of acts of remembering, the distinction between episodic and semantic memory builds upon the experience involved in such acts. Episodic (but not semantic) memory involves mental reenactments of earlier experiences, "mental time travel" (Tulving 1985). In the following, «remembering» refers to the mental experience involved in the episodic memory.

Most often, the mental reenactment in a «remembering» involves mental representations. Mental representations (in general) have two structural features: targets and content (Cummins 1996). Robins (2020: 123) offers an account of these features for «remembering»s specifically. The targets of a «remembering» is what it is attempting to represent. The content of a «remembering» is how it represents its

targets[1]. The accuracy of a «remembering» can be measured from the (mis)match between features of targets and how they are represented in the content. Most of the researches on declarative memory share the background assumption that a «remembering» has one (and only one) target. The assumption is made explicit by the use of the definite article 'the' for qualifying 'target'. For example, "*the* target is a particular event or experience in the representer's personal past" (Robins 2020: 123)[2].

Bernecker works with propositional memory; Michaelian and Robins, with episodic memory. Most cases of episodic memory are not cases of propositional memory, in the sense of having targets that are not propositions, but Bernecker (2017: 3) claims that some cases of propositional memory are episodic, in the sense of being accompanied by (episodic) «remembering»s. I use the expression 'memory of individuals' as a means of focusing on the cases of episodic memory whose targets are concrete individuals: people, objects, events (including mental events), etc. In the following, all «remembering»s target concrete individuals. Since all cases of memory of individuals are episodic, this emphasis will not be a problem in the discussion about simulation theory. The same is not true for causal theory, since Bernecker, but not Robins, works with propositional memory. My conclusions may not apply to propositional memory.

2.1 Causal theory

The causal theory states that successfully remembering an individual requires having an accurate representation of that individual standing in an appropriate causal connection with an earlier experience of it. This is the classical formulation of the causal theory, where "he" successfully remembers "something past" only if:

1 The content of a «remembering» may encapsulate mental images, conceptual content and other phenomenological features, such as the feeling of pastness (Robins 2020: 124). For simplicity, I refer to all these features as 'representations'.

2 Likewise, Michaelian (2016b: 103) states that "simulation is a matter of drawing on a range of past experiences to produce a… representation of *the* target episode". Bernecker (2010: 163) states that "this raises the question to what degree a memory report must correspond to *the* target stimulus to count as accurate" (my emphases).

> 1. Within certain limits of accuracy he represents that past thing. 2. If the thing was 'public', then he observed what he now represents. If the thing was 'private', then it was his. 3. His past experience of the thing was operative in producing a state or successive states in him finally operative in producing his representation. (Martin and Deutscher 1966: 166)

In the following, 'accuracy condition' refers to (1) and 'causal condition' refers to (3), where the notion of being "operative" refers to a causal connection[3].

The definition of the accuracy condition depends on how the theory describes the content and target selection. The causal theory has two versions: direct (Bernecker 2008) and indirect (Robins 2016b). I focus on the direct version because the proposed causal and simulation theories are indirect, where the difficulties of a direct theory will motivate some of their features, but I will also discuss a provisional indirect causal theory. Bernecker describes his theory as being "direct" and opposes it to the "representative" (indirect) theories. His "direct theory" is an unstable because it implements opposing intuitions: "We do not remember the past by virtue of being aware of an image presenting the past to us, rather our awareness of the past is direct", but "memory is indirect in the sense that it involves internal representations" (2008: 67 and 8). For this reason, Aranyosi (2020: 8) thinks that Bernecker's theory "does not deserve the name 'direct realism about memory'". I will not discuss this issue here.

What I will try to do is to reconcile Bernecker's opposing intuitions: a «remembering» involves mental representations, but grants direct access to its targets. As I understand it, 'direct access' should mean access not (completely?) dependent on mental representations. In this case, the content a «remembering» could be modeled using a device analogous to a proper name in a causal theory of reference (Kripke 1980)[4]:

3 'The earlier experience condition', expressed in (2), is not particularly controversial. This condition is presupposed in all cases through this paper, regarding both the causal and simulation theories.

4 I am not asserting that the content of a «remembering» of individuals is a proposition (see fn. 1). This is a simplified model used to highlight some properties of

$$P_1a \,\&\, P_2a \,\&\, \ldots \,\&\, P_na,$$

where the predicates P_1, P_2, …, P_n model the properties represented in the «remembering». The direct character of the theory is modeled using the proper name a in the content of the «remembering». This seems to be appropriate because, in a causal theory of reference, a proper name refers to an individual via a causal chain of uses. The reference is not dependent on properties attributed to the referent[5]. If the margin of error is null, then & refers to the logical conjunction. I do not claim to have modeled Bernecker's theory because, if not for other reasons, Bernecker's is a theory of propositional memory. This is a model of a 'Bernecker-like' theory for memory of individuals. Also, this should not be seen as a model of a relational theory of memory (e.g. Debus 2008)[6].

Target selection for the direct causal theory is very simple because the targets of a «remembering» are referred in its content. If a «remembering» has content of the form $P_1a \,\&\, P_2a \,\&\, \ldots \,\&\, P_na$, then its target is the individual a. The direct causal theory seems to be inescapably committed to the background assumption for two reasons. First, a «remembering» represents its targets as one (and only one) individual. This is supposed to be a general feature of «remembering»s, which all theories should express (Perrin et al. 2020: 5)[7]. Second, that

the theory that are relevant to our discussion.

5 The reference of a causal proper name does not depend on attributions of properties, but it can involve attributions of properties. For example, the proper name 'Holy Roman Empire' attributes properties to some, but refers to something that is "neither holy, Roman nor an empire" (Kripke 1980: 26). The same holds for a direct theory of memory: a «remembering» represents its targets as having properties, but which target is selected is not determined by how it is represented.

6 A relational «remembering» would have individuals themselves as constitutive parts. A "direct" «remembering» has a causal naming devices as constitutive parts.

7 "As far as episodic memory is concerned, it characterizes the particularity of experienced and remembered events that are represented by the memories, for instance such or such particular event of going to the university on a specific day (e.g. the time I meet an old friend on the tram as I was heading towards the university), in contrast with the iterative event of doing so. On standard accounts of the content of episodic memories, they typically represent singular events in this sense". That this assumed by "the vast majority of studies" (Perrin et al. 2020: 5).

the subject is using different naming devices *in the same context* should be transparent to her[8]. Note that, for other theories, a «remembering» representing its targets as one (and only one) individual does not entail the background assumption that it must have one (and only one) target.

The accuracy condition for the direct causal theory is as follows: a «remembering» with content of the form P_1a & P_2a & … & P_na and, consequently, target a is accurate if and only if (iff) P_1a & P_2a & … & P_na. In the following, I work under two simplifications about accuracy. First, since memory is a generative process (Schacter and Addis 2007), the accuracy condition should include a margin of error for &. I will partially ignore this issue until §4.1. Second, the accuracy of a «remembering» is often measured against an earlier experience of that individual, but I will measure accuracy directly in relation to the targeted individual. This simplification is equivalent to the assumption that the earlier experience is accurate[9].

The causal condition states that there must exist a causal connection between a «remembering» and an earlier experience of its targets (see condition 3 of Martin and Deutscher). However, I will often talk about the causal connection being between a «remembering» and its targets, which is equivalent to supposing that the earlier experience (e.g. perception) is also causal. I will be more cautious when the earlier experience is not causal (e.g. imaginations)[10]. The causal theories almost

8 In the Paderewski case (Kripke 1979), a subject is unaware of her using the same proper name in different contexts. I don't think that it is possible for a subject to be unaware that she is using different proper names in the same context because these names would need to exhibit syntactical differences. The prospects for a relational theory are better here because even if the use of two different proper names in the same context is transparent to the subject, which individuals are being referred to isn't.

9 I will specify the character of the earlier experience when it is relevant (e.g. when the targets are previously imagined individuals). Please note that the difference between «remembering» a previously imagined individual (e.g. an imagined person) and «remembering» the act of imagining that individual, which «remembering» a mental event and not the individual.

10 The causal theory is ambivalent about whether a «remembering» of an individual that was previously imagined can be non-confabulatory. The issue is that imaginations often target non-existent/occurrent individuals. These are «remembering»s caused by

always demand the causal connection between a «remembering» and its targets to be sustained by an *appropriate* causal chain of memory traces, which are supposed to have some features, such as not being "deviant". These provisos are necessary for distinguishing successful remembering from relearning. I will briefly discuss memory traces in §4.3.

In addition to the chain of memory traces, some versions of the causal theory demand the causal connection to instantiate a counterfactual relation (e.g. Bernecker 2010: 98): "If *S* hadn't represented at t_1 that p^* he wouldn't represent at t_2 that p", where $t_1 < t_2$ and p^* and p refer to the content of the earlier experience and of the «remembering» (respectively). Other versions of the causal condition only demand the causal connection to retain information about the targets: "Remembering occurs when a person retains the capacity to represent information acquired from past events" (Robins 2017: 2). The counterfactual condition entails the informational condition, but not the other way around (see §4.2). I will focus the counterfactual condition because the proposed causal theory has an informational causal condition and the difficulties of the first will motivate features of the second. The direct causal theory with counterfactual condition does not correctly classify the errors in §3 as misremembering or confabulation.

Robins (2020: 124) proposes a taxonomy of memory errors that uses the accuracy and causal conditions to distinguish cases of successful remembering, misremembering, and falsidical confabulation [Table 1][11].

	Accuracy condition	Causal condition
Successful remembering	Yes	Yes
Misremembering	No	Yes
Falsidical confabulation	No	No

Tab. 1: Causal taxonomy for (some) memory errors (Robins 2020: 124).

A misremembering is an inaccurate «remembering» that is caused by its targets: "When a person misremembers, her report is inaccurate, yet this inaccuracy is explicable only on the assumption that she has

an earlier experience of their targets without being caused by the targets themselves.

11 Robins also discuss cases of relearning, but I will not deal with those cases.

retained information from the event her representation mischaracterizes" (Robins 2016a: 434). In the model, a misremembering occurs when a «remembering» with content of the form P_1a & P_2a & … & P_na is caused by the target a, but P_ia is false for a given number of is (depending on the margin of error).

The term 'confabulation' was introduced in the beginning of the 20th century for describing subjects with Korsakoff's syndrome (see Berrios 1998 for a review). Afterwards, the term was applied to a wider range of phenomena, such as "the involuntary and unconscious… recollection of episodes, which never actually happened, or which occurred in a different temporal-spatial context to that being referred to by the patient" (Dalla Barba 2002: 28). The causal theory describes confabulation as a «remembering» not caused by its targets[12]. A falsidical confabulation is a confabulation with inaccurate content. In the model, a falsidical confabulation occurs when a «remembering» with content of the form P_1a & P_2a & … & P_na is not caused by the target a and P_ia is false for a given number of is (depending on the margin of error).

2.2 Simulation theory

The simulation theory states that successfully remembering an individual requires having an accurate representation of that individual generated by a reliable memory process (Michaelian 2016b: 97). The first condition is the accuracy condition; I refer to the second as 'reliability condition'. In the simulation theory, a successful remembering of an individual may be generated using information from an earlier experience of the individual, but also other experiences of the subject, general knowledge about the world, etc (no causal condition).

In the literature, the simulation theory is only presented in an indirect version (e.g. Michaelian 2016b)[13]. In an indirect theory, "one

12 "Mnemonic confabulation… occurs when there is no relation between a person's seeming to remember a particular event or experience and any event or experience from their past – either because there is no such event in their past or because any similarity to such an event is entirely coincidental" (Robins 2020: 122).

13 The prospects for a direct simulation theory depends on a mode of reference that is neither causal nor descriptive. In the first case, a 'direct simulation theory'

remembers something… with a mediating image which represents that thing" (Bernecker 2008: 65). In this case, the indirect content of a «remembering» would be a form of descriptive content, which could be modeled using a device analogous to the unique existential quantification (Whitehead and Russell 1910: 173):

$$\iota x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx),$$

where the quantifier ιx should be read as 'the x'. The expression $\iota x(Px)$ should be interpreted as being equivalent to $\exists x(Px \,\&\, \forall y(Py \rightarrow x = y))$. Previous considerations about the predicates P_1, P_2, …, Pn and the margin of error for & apply here.

The indirect content is related to a description theory of reference (Searle 1958). The use of descriptions to model the indirect character of the theory is adequate because the meaning of a description is spelled out without mentioning individuals (not even the individuals, if there are some, that satisfy the description). The use of *definite* descriptions is adequate because a «remembering» represents its targets as one (and only one) individual (see fn. 7). The existence of a margin of error is even more pressing for an indirect theory than it is for a direct one. Often we, creatures with limited cognitive resources, have vague but successful «remembering»s, which represent its targets as one and only one individual ('the x'), but fail to target one (and only one) individual[14].

Target selection for an indirect theory may use an ι function[15]. If both the functions work in the same way, target selection for an indirect

would collapse into a causal theory; in the second case it would collapse into an indirect theory. The indexical mode of reference seems to be related to the relational theory.

14 Vague remembering is not a problem for a direct causal theory because reference is secured by a device analogous to a causal proper name. In fact, the vaguer a «remembering», the more easily it is accurate. This modeling has the consequence that the direct causal theory predicts successful «remembering»s with less representational content than the indirect theories.

15 For simplicity, I refer to both the content and the target selection functions as, but these are different functions: the former returns truth-values; the latter, individuals. The difference will be important in §4.1.

theory would work as follows: if a «remembering» has content of the form $\imath x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$, then its target is 'the x' that uniquely satisfies this description (this is an expression of the background assumption). In this case, the «remembering» does not have a target when either zero or more than one x satisfy the description. In general, the accuracy condition for an indirect theory is as follows: a «remembering» with content of the form $\imath x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$ and targets $y_1, \ldots, y_j, \ldots$ is accurate iff $\imath y_i(P_1y_i \,\&\, P_2y_i \,\&\, \ldots \,\&\, P_ny_i)$ for all y_j. This accuracy condition is trivial if both functions work in the same way: a «remembering» is accurate iff it has one (and only one) target.

For a final point about indirect theories, we could define a provisional indirect causal theory with content and target selection modeled using $\imath$ functions as above, and a counterfactual causal condition. This theory would not correctly classify the memory errors in §3 as misremembering or confabulation.

The reliability condition is as follows: a successfully remembering must be generated by a reliable memory process. Roughly, a reliable memory process is one that tends to produce mostly accurate representations (Michaelian 2016a: 6). Michaelian insists that this tendency should be understood as a modal (and not a purely statistical) notion. This modal reading does not preclude the existence of interesting connections between reliability and probability. For example, it is reasonable to presuppose that a reliable memory process, working under normal conditions, produces accurate «remembering»s with a ratio higher than .5. The (initial) simulation theory would have content and target selection modeled using an $\imath$ function as above, and the reliability condition. This theory correctly classifies the memory errors in §3 as misremembering or confabulation, but its description of these cases is unsatisfactory.

Michaelian (2016a: 8) proposes a taxonomy of memory errors that uses the accuracy and the reliability conditions to distinguish cases of successful remembering, misremembering, and falsidical confabulation [Table 2][16].

16 Michaelian also discusses cases of relearning and veridical confabulation, where a confabulation ends up (by mere luck) representing its targets accurately. I will not deal

	Accuracy condition	Reliability condition
Successful remembering	Yes	Yes
Misremembering	No	Yes
Falsidical confabulation	No	No

Tab. 2: Simulationist taxonomy of (some) memory errors (Michaelian 2016a).

A misremembering occurs "when the reliability condition is met but the accuracy condition is not". In the model, a misremembering occurs when a «remembering» with content of the form $\exists x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$ is generated by a reliable memory process, but some of its targets y_j are such that P_iy_j is false for a given number of is (depending on the margin of error). A falsidical confabulation "occurs when neither the reliability condition nor the accuracy condition are met" (Michaelian 2016a: 7). In the model, a falsidical confabulation is a confabulatory «remembering» with content of the form $\exists x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$, but some of its targets y_j are such that P_iy_j is false for a given number of is (depending on the margin of error). Since it is reasonable to presuppose that human memory systems in normal functioning and conditions implement reliable memory processes, the simulationist notion of confabulation refers to situations where human memory systems are not reliable, which is more akin to the initial use of the notion.

3 Memory errors

The study of memory errors can provide insights about the mechanisms at work in both successful and unsuccessful memory. The results in "the science of false memory", as it has come to be developed (see Brainerd and Reyna 2005 for a review), describe enduring characteristics of normal, rather than pathological, memory. For this reason, theories of memory must predict or, at least, be compatible with these results. I survey three experimental paradigms: DRM (Deese 1959; Roediger and McDermott 1995), "lost in the mall" (Loftus and Pickrell, 1995), and memory-conjunction errors (Reinitz et al. 1992). Robins cites DRM

with those cases.

and "lost in the mall" as exemplar cases of misremembering and falsidical confabulation (respectively). Michaelian diagnoses both as misremembering. Neither Robins nor Michaelian have studied memory-conjunction errors.

3.1 DRM

The DRM (Deese 1959; Roediger and McDermott 1995) paradigm is composed of two phases. In the study phase, the subjects are presented with lists of semantically related words. In the test phase, the subjects are asked to «remember» as many studied words as possible. In a typical result, subjects report «remembering» a "lure" word, which was not studied but is the 'semantic focus' of a studied list. This is an example of a list of words used in the paradigm ('king' is the lure):

> (King) queen, England, crown, prince, George, dictator, palace, throne, chess, rule, subjects, monarch, royal, leader, reign (Roediger and McDermott 1995).

Causal theory

The correct diagnosis for cases of DRM should be one of misremembering because their explanation involves "an appeal to a particular past event that has been distorted" (Robins 2016a: 434). Nevertheless, it is difficult to reach this diagnosis within a causal theory because a misremembering *of the lure* would need to be caused by an earlier experience of the lure, which was not seen in the study phase.

The 'natural description' of cases of DRM is that the words seen in the study phase cause the misremembering of the lure in the test phase. This description is not available for a direct causal theory because a 'misremembering' with content of the form P_1a & P_2a & … & P_na, where a refers to the lure, would need to be caused by the lure, but the lure was not seen in the study phase. This would be the description of a confabulation. If the content of the misremembering is about another word (not the lure), this would not be a misremembering of the lure.

The direct causal theorist needs to adopt a more convoluted description. In the 'list description', the content of the misremembering

of the lure would be about a list as a whole. The content of the «remembering» would have the form P_1a & P_2a & … & P_na, where a refers to the list seen in the study phase and each of the P_1, P_2, …, P_n describes a word in the list, except for one (e.g. P_n), which describes the lure. This would be a case of misremembering because the «remembering» would be caused by the list seen in the study phase, but its content would be inaccurate because the lure was not in the list (e.g. P_n a is false). However, this description is not adequate because subjects in DRM studies report «remembering» the event of studying the lure (and not of the list as a whole)[17].

As I can see it, there remains two options for a direct causal theorist. The 'temporal misattribution description' consists in the «remembering» being about the lure but being caused by some other experience of the lure (e.g. reading the lure in a book years ago). The result would be of misremembering because the «remembering» represents the lure as being seen in the study phase[18]. This description is artificial because of the appeal to an indefinite event of study, but it is used for other cases of misremembering (e.g. Levine 1997). The 'reality-monitoring error description' consists in the «remembering» being about the lure, but being caused by an act of imagining the lure (e.g. during the study phase, see fn. 9). The result would be of misremembering because the «remembering» represents the lure as being studied (and not imagined). The reality-monitoring error description is more natural than the temporal misattribution one and it is also used for other cases of misremembering (see Johnson 1997).

The provisional indirect causal theory does no better. The content of the «remembering» would have the form $\imath x(P_1x$ & P_2x & … & $P_nx)$, where P_1, P_2, …, P_n represent the properties of the lure (including being seen in the study phase). Either 'the x' that satisfies the description is not

17 "Our subjects confidently recalled and recognized words that were not presented and also reported that they remembered the occurrence of these events" (Roediger and McDermott 1995: 812).

18 I sometimes appeal to the property of having been seen in the study phase as being represented in the content of a «remembering». This 'representation' may refer not to an explicit representation but to the feeling of pastness, ownership, etc associated with the phenomenology of episodic memory (Perrin et al. 2020 and fn. 1).

the lure or it is (depending on the margin of error). In the fist case, this would be a confabulation. In the second, a successful remembering.

Simulation theory

The simulationist diagnosis for cases of DRM is of misremembering (Michaelian 2016a: 9). The content of the «remembering» would have the form $\exists x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$, where the P_1, P_2, …, P_n represent the properties of the lure (including being seen in the study phase). Since the lure was not studied and the studied words do not have most of the properties of the lure, the content would be inaccurate, however, since cases of DRM occur in human memory systems in normal functioning and conditions[19], the diagnosis would be of misremembering. The diagnosis is correct, but there is something unsatisfactory about this description: this would be a case of misremembering 'of the lure', which does not target the lure (in fact, which does not have targets at all).

3.2 Lost in the mall

Loftus and Pickrell (1995: 721) recruited 24 pairs of relatives. The older relatives provided their youngers with four stories about the youngers' childhood (three of which were true; the fourth was a fiction about getting lost in the mall). The youngsters were asked to «remember» the four events. In the most well known result (Loftus et al. 1996), Chris, a 14-year-old boy, was informed by his older brother, Jim, that, when he was five, he was lost in a mall in the city of Spokane, Washington, where his family often went shopping, and that he was crying heavily when he was rescued by an elderly man and reunited with his family. Over five days and two interviews, Chris had «remembering»s of being rescued by a "really cool" man, of being scared that he would never see his family again, etc.

19 The DRM effect has been replicated extensively and can be obtained from different forms of similarity (categorical, phonological, orthographic); kinds of stimuli (pictures, faces, dot arrays); intervals between the study and recognition phases (hours, days, months), etc (see Robins 2016a: 434, for the references).

Causal theory

The 'official' causal diagnosis for cases of "lost in the mall" (LTM) is of falsidical confabulation (Robins 2016a: 434). There is a 'natural description' of these cases available for the direct causal theory. The content of Chris' «remembering» would have the form P_1a & P_2a & … & P_na, where a refers to the event of getting lost in the mall and P_1, P_2, …, P_n represent the properties of this event (including being part of Chris' past). Since there isn't an event with all these properties, the content is inaccurate. For the same reason (as the description goes), that event cannot be the cause of Chris' «remembering».

Loftus and Pickrell (1995: 724) adopt the reality-monitoring error description for cases of LTM[20]. In this case, the content of Chris' «remembering» would have the form P_1a & P_2a & … & P_na, where a refers to an imagined event of getting lost in the mall and P_1, P_2, …, P_n represent the properties of that event (including being part of Chris' past). The «remembering» would be inaccurate because the imagined event is not part of Chris' past, but it would be caused by an act of imagining that event (e.g. during the study phase, see fn. 9). This would be a diagnosis of misremembering. The choice between the first description over the second amounts to the restriction of the causal condition so that acts of imagining cannot cause non-confabulatory «remembering»s. In this case, the reality-monitoring error description would not be available, not only for cases of DRM, but for all cases described in the literature (e.g. Johnson 1997). These cases would need to be described as a temporal misattribution of an indefinite event of study.

The provisional indirect causal theory does no better. The content of Chris' «remembering» would have the form $\exists x(P_1x$ & P_2x & … & $P_nx)$, where P_1, P_2, …, P_n represent the properties of this event (including

20 "The development of the false memory of getting lost may evolve first as the mere suggestion of getting lost leaves a memory trace in the brain. Even if the information is originally tagged as a suggestion rather than a historical fact… as time passes and the tag that indicates that getting lost in the mall was merely a suggestion deteriorates. The memory of a real event, visiting the mall, becomes confounded with the suggestion that you were once lost in the mall" (Loftus and Pickrell 1995: 724).

being part of Chris' past). Either 'the x' that satisfies this description does not exist or is an imagined event (depending on the margin of error). In the fist case, this would be a confabulation. In the second, either an imagination of 'the x' cannot properly cause «remembering»s or it can. In the first case, again, this would be a confabulation. In the second, a successful remembering.

Simulation theory

The simulationist diagnosis for cases of LTM is of misremembering (Michaelian 2016a: 6). The content of Chris' «remembering» would have the form $\exists x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$, where P_1, P_2, …, P_n represent the properties of the event of getting lost in the mall (including being part of Chris' past). This content is inaccurate because 'the x' that satisfies this description is not part of Chris' past, but cases of LTM occur in human memory systems in normal functioning and conditions (which are close to the ecological in LTM experiments). Then the diagnosis is of misremembering[21]. The difference in diagnosis between the causal and simulation theories is partly due to the different notions of confabulation adopted: where simulationist confabulation applies more directly to pathological cases, causalists insist in using a broader notion. I find it plausible to describe cases of LTM as misremembering or falsidical (causalist) confabulation[22].

21 This would be a case of misremembering without targets, which is reasonable here since the event of getting lost in the mall (the intended target) did not occur.

22 The causal and simulation theories diverge in diagnosis for reality-monitoring errors occurring in human memory systems in normal functioning and conditions. I do not find it plausible to describe all reality-monitoring errors in the literature as falsidical confabulations.

3.3 Memory-conjunction errors

Fig. 1: The left and middle panels show potential study stimuli, and the right panel shows a potential conjunction stimulus constructed from them. (Reinitz et al. 1992: 6)

Memory-conjunction errors (MCEs) occur when subjects study a number of related items and have a «remembering» with content constructed from elements of more than one item. Reinitz et al. (1992 E. 6) have tested 48 subjects using line drawings of human faces. In the study phase, each subject was presented with six randomly selected faces. In the test phase, the subjects were presented with eight stimuli in a recognition test: two previously studied faces (target stimuli); two faces constructed by combining the features of two studied stimuli (conjunction stimuli); two faces in which features of one studied stimulus are combined with unstudied features (feature stimuli); two faces entirely constructed from unstudied features (new stimuli). The subjects should answer to the question: "Was this one of the faces you studied?". The relative frequencies of recognition for target, conjunction, feature, and new stimuli were .71, .52, .19, and .13 (respectively). Suppose that subject *S* was presented with the first two faces in Figure 1 and asked to «remembers» the third. Let the first, second, and third faces be *a*, *b*, and *c* (respectively). Then *c* is composed of features of *a* and *b* in equal parts (conjunction stimulus, see appx.).

Causal theory

MCEs are problem cases for the direct causal theory for three reasons. First, the direct causal content yields different (conflicting) descriptions

of MCEs, where there is no reason whatsoever for choosing among these descriptions. Suppose that the causal diagnosis for MCEs is of misremembering. The only two descriptions of the content of *S*' «remembering» available for the causalist are: P_1a & P_2a & … & P_na and P_1b & P_2b & … & P_nb. There is no reason for deciding which of these formulas describe the content of *S*' «remembering» because *c* is composed of features of *a* and *b* in equal parts. This is problematic because these descriptions are not equivalent from the point of view of the causal theory, since they pose different causal histories: these are «remembering»s caused by *a* [alternatively, *b*] but not necessarily by *b* [*a*]. The second reason has to do with MCEs being cases of ambiguous misremembering, where (part of) the inaccuracy is caused by multiple targets[23]. The direct causal theory is inescapably committed to the background assumption (see §2.1) and, consequently, the description of MCEs as ambiguous «remembering»s is not available to the causalist.

The third reason has to do with the counterfactual version of the causal condition. The counterfactual condition states that "If S hadn't represented at t_1 that *p** he wouldn't represent at t_2 that *p*" (Bernecker 2010: 98), where this clause should be interpreted as in Lewis (1973b). In Lewis' analysis, a counterfactual is true iff its consequent is true in all possible worlds maximally similar to the actual world where its antecedent is true. The results of Reinitz et al. show that there is a (frequentist) probability of 29% of *S* «remembering» *c* if she didn't see a [alternatively, *b*] in the study phase (see appx.). In this case, it is reasonable to suppose that *S* «remembers» *c* in at least some of the maximally similar worlds where *S* didn't see *a* [*b*] in the study phase[24].

23 Schooler and Tanaka (1991) distinguish between composite recollections, «remembering»s representing features from multiple sources, and compromise recollections, «remembering»s in which previously experienced features are combined to produce a recollection that represents a perceptual or semantic averaging of the studied items. MCEs are cases of composite recollection, but cases compromise recollection can also give rise to ambiguous «remembering»s.

24 Consider the "standard resolution of vagueness" (Lewis 1979:. 472). The situations where *S* «remembers» *c* without seeing *a* [alternatively, *b*]: (1 and 3) do not involve violation of laws (of Physics, Psychology, etc), (2) maximize the space-temporal match with the actual world (*S* «remember» *c* in the actual world), and (4)

Then the counterfactuals 'if *S* hadn't seen *a* [*b*] in the study phase, she wouldn't «remember» *c* in the test phase' are both false and neither *a* nor *b* fulfill the counterfactual condition. The resulting diagnosis would be of confabulation[25].

It is not reasonable to diagnose MCEs as confabulations (even in the causalist broader notion). Causalist confabulation "occurs when there is no relation between a person's seeming to remember a particular event or experience and any event or experience from their past" (Robins 2020: 125). But MCEs are causally anchored in past experiences of the subject (the study of *a* and *b*). In fact, had not *S* seen *b* [alternatively, *a*] in the study phase, the causal theory would happily describe *S*' «remembering» as a misremembering[26]. Also, if *c* were composed of more features of *a* [*b*], there would be reason for choosing the first of the two initial descriptions and, again, the direct causal theory would happily describe *S*' «remembering» as a misremembering. It is not reasonable that seeing more faces in the study phase (or the composition of the third face being slightly different) transforms a misremembering in a confabulation.

secure similarity of all other particular facts.

25 It is worthwhile to note that Bernecker's counterfactual condition is a version of Lewis' causal dependence (not causation). Lewis (1973a) himself thinks that causal dependence, although sufficient, is not necessary for causation. Lewis' causation is such that *a* [alternatively, *b*] is a cause of *c* when there is a causal chain (e.g. of memory traces) leading from *a* [*b*] to *c*, where each element of the causal chain is causally dependent on the former. This notion is of little help here. Whenever the *a*-memory trace interacts with the *b*-memory trace to form the first *c*-memory trace in the causal chain that leads to *c*, the first *c*-memory trace is not causally dependent on either the *a*-memory trace or the *b*-memory trace for the same reasons that c is not causally dependent on *a* or *b*. Lewis (1973a) is aware of the limitation of his theory in dealing with cases of overdetermination (e.g. MCEs), which he attempts to dismiss: "I shall not discuss symmetrical cases of overdetermination, in which two overdetermining factors have equal claim to count as causes. For me these are useless as test cases because I lack firm naive opinions about them" (Lewis 1973a: fn. 11). I discuss cases of overdetermination in §4.2.

26 The content of *S*' «remembering» would have the form P_1a & P_2a & … & P_na, where, for half of the P_is, it is false that P_ia (inaccurate), but the «remembering» would be caused by *a* [alternatively, *b*].

The provisional indirect causal theory does no better. Similarly to the direct causal theory, MCEs would be cases of confabulation because of the counterfactual condition. But the provisional theory has its own problems with ambiguous «remembering»s. The indirect content of a «remembering» (e.g. *S*' «remembering» of *c*) would have the form $\imath x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$. If both functions work in the same way, cases where more than one *x* satisfy the description (e.g. *a* and *b*) are cases where the «remembering» has no targets. Cases of ambiguous misremembering would be always diagnosed as confabulations.

Simulation theory

The simulationist diagnosis for MCEs is of misremembering. The content of *S*' «remembering» would have the form $\imath x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$, where P_1, P_2, …, P_n represent the features of *c*. No matter the margin of error, either both *a* and *b* satisfy the description or neither does. In either way, the content would be inaccurate, but cases similar to MCEs occur in human memory systems in normal functioning and conditions[27]. Then the diagnosis would be of misremembering. The diagnosis is correct, but there is something unsatisfactory about this description: if both the functions work in the same way, either both *a* and *b* satisfy the description or neither does. Either way, the «remembering» does not have targets. Then all cases of ambiguous misremembering are described as misremembering without targets.

4 Discussion

In the previous sections, I have shown that DRM, LTM, and MCEs are problem cases for the direct and provisional indirect causal theories. The simulation theory classifies these cases correctly, but its descriptions are unsatisfactory. The most problematic cases were MCEs. The source of the difficulty is the background assumption that a «remembering» has one (and only one) target, which is reflected in the accuracy and

27 For example, a witness to a crime might claim to have seen an individual's face when she had in fact seen several faces that, when taken together, contained most of the facial features of the accused individual. Brown et al. (1977) tests this possibility in conditions closer to the ecological.

counterfactual conditions of these theories. The direct causal theory cannot be fixed for these cases because of both its direct content, which is inescapably committed to the background assumption, and its counterfactual condition, which has problems with cases of overdetermination (see fn. 25). Robins' and Michaelian's theories can be fixed for these cases. The margin of error for accuracy should include the uniqueness claim and the target selection function should return all the individuals with the highest degree of satisfaction. The counterfactual condition should be reduced to a probabilistic condition. The resulting theories classify correctly the errors in §3.

4.1 Accuracy condition

The direct causal theory cannot deal with ambiguous misremembering because it is inescapably committed to the background assumption. This is not the case for the indirect theories because it is not transparent to the subject whether a descriptive content is satisfied by one or more individuals. There is a related problem for the provisional indirect causal and simulation theories: when we misremember, we misremember something(s), but if both $\imath$ functions work in the way described above, then a «remembering» with content of the form $\imath x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$ is accurate iff it has one (and only one) target (and it is inaccurate iff it has no targets). This is unfortunate because there are cases of ambiguous misremembering and of vague but successful remembering.

The use of an $\imath$ function for modeling the accuracy condition has the consequence that all vague «remembering»s are inaccurate and, consequently, not successful. We, creatures with limited cognitive resources, most often have vague «remembering»s, whose content is not detailed enough to single out one (and only one) individual. If the accuracy condition demands the content of a «remembering» to single out one (and only one) individual, then most of our «remembering»s would be inaccurate. For example, suppose that I know twins that are identical except for a minimal difference. My «remembering» will hardly single out one of them, especially if their difference is minimal enough to pass any non-null margin of error. Nevertheless, I should be able to successfully remember one of them. The key to avoid this

problem is to note that $\iota x(Px)$ is equivalent to $\exists x(Px \,\&\, \forall y(Py \rightarrow x = y))$. The uniqueness claim $(\forall y(Py \rightarrow x = y))$ is as a clause of the full formula as any $P_i y$ and, as such, it should be included in the margin of error for &. If the margin of error is non-null, the accuracy condition may return 'accurate' when two or more individuals satisfy the description (but not when no one does due to the wider scope of the existential quantifier).

The adjustment of the margin of error is not able to solve the corresponding problem for target selection. Suppose, again, that I know twins that are identical except for a minimal difference. I should be able to successfully remember one of them, but any (non-null) margin of error that selects one of them selects both. In order to deal with this problem, the target selection function should be plural in the sense of returning all individuals that satisfy the description with the highest degree of satisfaction. The plural target function selects the right twin in our example. Note that targets outside the margin of error may be selected. This is what happens in cases of inaccurate «remembering»s.

The simulation theory should be construed using the accuracy condition and target determination as described above. The resulting theory provides consistent diagnoses for the cases in §3. For DRM, the content of the «remembering» would represent the properties of the lure (including being seen in the study phase). The target would be the lure because this is the word with highest degree of satisfaction of the description. The content would be inaccurate because the lure was not seen in the study phase. For LTM, the content of Chris' «remembering» would represent the properties of the event of getting lost in the mall (including being part of Chris' past). The target would be that event. The content would be inaccurate because the event is not part of Chris' past. For MCE, the content of *S*' «remembering» would represent the properties of *c* (including being seen in the study phase). The targets would be *a* and *b* because these are the studied faces with highest degree of satisfaction of this description. *S*' «remembering» would be inaccurate because $P_i a$ [$P_i b$] is false for half of the P_i in the description. Since these «remembering»s are generated by reliable processes, these would be diagnoses of misremembering.

4.2 Causal condition

The counterfactual version of the causal condition demands the relation between non-confabulatory «remembering»s and their targets to be one of counterfactual dependence. This condition is unable to deal with ambiguous «remembering»s because often these are cases of overdetermination and counterfactual theories of causation have problems with those cases (see fn. 25). MCEs are cases of overdetermination because the study of *a* [*b*] is individually sufficient for the «remembering» of *c* (feature stimuli). The problem counterfactual theories of causation have with overdetermination is that "causes are thought to be necessary for their effects, in the way that counterfactual analyses are placed at center stage, but neither cause, given the occurrence of its companion, is necessary in overdetermination cases" (Hall and Paul 2013: 146-7). This feature of the counterfactual condition is related to the background assumption.

Historically, causalists have paid attention to the distinction between successful remembering and relearning, but much less to cases of ambiguous «remembering»s. This focus has consequences to which cases of redundant causation that the theory can deal with. The three main cases of redundant causation are joint causation, preemption, and overdetermination. Cases of joint causation are not particularly difficult, but preemption and overdetermination often present issues to causal theories. Since relearning often involves preemption, causal theory of memory is shaped to deal with those cases, but much less with overdetermination. Bernecker discusses four cases of redundant causation, where the first is of overdetermination (the second is of preemption):

> [S]uppose you are taking part in a family reunion and are trying to remember a distant relative's name. At the very moment when you are about to remember that the relative is called 'Bert' your partner who is next to you blurts out "Bert, good to see you!". Your memory that the person's name is 'Bert' is causally overdetermined by your previous knowledge of this fact and by your partner's blurting. …Intuitively, however, you do remember that the person is called 'Bert'. (Bernecker, 2008, p. 48)

This is a sui generis case of overdetermination because your trying to remember and your partner's utterance are, in some sense, competing causes[28]: if your trying to remember had not occurred, the result would be a case of relearning (not of successful remembering). This setting is typical of cases of preemption. For this reason, Bernecker has at his disposal a very simple strategy for distinguishing between *these* cases of overdetermination (which give rise to successful remembering) and cases of preemption (relearning): if your partner's utterance happens before (as in Bernecker's second case), there is relearning; there is successful remembering otherwise. This solution is not available for MCEs because their causes are not competing, but joint (studying both *a* and *b* enhances the probability of «remembering» *c*). In this respect, MCEs are closer to Bernecker's third and fourth cases (of joint causation). However, MCEs do not fit those cases completely because the study of *a* [alternatively, *b*] is not an *indispensable* part of a jointly sufficient condition for the «remembering» of *c*.

MCEs are mixed cases of overdetermination and joint causation. These are cases of overdetermination because the study of *a* [alternatively, *b*] alone can cause the «remembering» of *c* (19%, feature error), but these are cases of joint causation because studying both *a* and *b* enhances that probability (52%, conjunction error). This pattern of causal influence is better modeled using probabilistic notions. Reichenbach (1956) and Suppes (1970) propose a probabilistic notion of causation, where *a* that occurs at t_1 is a cause of *b* that occurs at t_2 iff: (1) $t_1 < t_2$; (2) $p(b|a) > p(b)$; and (3) there is no event *a'* occurring at $t_0 \leq t_1$ such that *a'* screens *a* off from *b* (i.e. $p(b|a, a') = p(b|a')$)[29]. Bernecker dismisses this probabilistic notion of causation and poses two challenges to it: (i) "by how much does the conditional probability of the occurrence of the recounting have to exceed the probability of the occurrence of the recounting in general?"; (ii) "it doesn't seem likely

28 Your trying to remember and your partner's utterance are both causes of your «remembering», which is, in this sense, overdetermined. These are competing in the sense of being causes of a case of successful remembering and relearning respectively.

29 These were the leading theories of probabilistic causation of the 20th century, but they have been supplanted by the causal modeling approaches (Pearl et al. 2016).

that we will ever be in a position to give precise values for the probabilities in question" (2010: 95).

The probabilistic notion of causation can be used in a theory that can deal with the errors in §3. For simplicity, I focus on the uncontentious part of these theories: if *a* that occurs at t_1 causes *b* that occurs at t_2, then $t_1 < t_2$ and $p(b \mid a) > p(b)$ (leaving out the screening off condition). This uncontentious part defines a minimal causal condition, in the sense that all causal conditions should entail it. The minimal causal theory is construed using accuracy condition and target selection as defined in §4.1 and the minimal causal condition as defined above. If we assume the idealization that the representation of each feature $P_i x$ is caused by one (and only one) individual *a* such that $P_i a$, then target determination for this theory would be related to the notion of causal dominance in the hybrid theory of reference (Evans 1973)[30]. This leaves us with the following analogies between theories of memory and reference: direct causal theory (Bernecker 2008) and causal theory of reference (Kripke 1980); minimal causal theory and hybrid theory of reference (Evans 1973); simulation theory (Michaelian 2016a) and description theory of reference (Searle 1958).

The minimal causal theory provides consistent diagnoses for the errors in §3. For DRM, the content of the «remembering» would represent the properties of the lure (including being seen in the study phase). The target would be the lure (imagined in the study phase) because this is the word with the highest degree of satisfaction of the description. The content would be inaccurate because the lure was not seen in the study phase. For LTM, the content of Chris' «remembering» would represent the properties of the event of getting lost in the mall (including being part of Chris' past). The target would be that event

30 According to Evans, a proper name refers to the dominant causal source of the information associated with it. Consider a case where ancient documents containing interesting mathematical proofs are discovered. Inscribed in these documents is the name 'Ibn Kahn', which is now mistakenly taken to refer to the mathematician. In fact, the person originally named 'Ibn Kahn' was the scribe who copied the proofs. Evans claims that present uses of 'Ibn Kahn' refer to the mathematician because the dominant causal source of the information associated with the name is the mathematician. Michaelian has suggested this relation to me in personal conversation.

(imagined in the study phase), but the content would be inaccurate because that event is not part of Chris' past. If we accept that imagined individuals can cause non-confabulatory «remembering»s, these would be two diagnoses of misremembering. For MCEs, the content of *S*' «remembering» would represent the properties of *c*. The targets would be *a* and *b* because these are the studied faces with highest degree of satisfaction of the description. *S*' «remembering» would be inaccurate because $P_i a$ [alternatively, $P_i b$] is false for half of the P_i in the description, but it would be caused by both *a* and *b* because $p(c \mid a) = p(c \mid b) = .3718 > p(c) = .3024$ (see appx.). This would be a misremembering diagnosis.

4.3 Two informational theories of memory

The literature on memory is full of informational talk. Causalists and simulationists seem to agree that "the episodic memory system is… designed to draw on information originating in past experience to simulate possible episodes" (Michaelian 2016b: 103)[31]. In sum, "there is general agreement that the human memory is meant to not only store but also process the encoded information" (Bernecker 2017: 4), although I am not sure of the agreement about the 'storing' part. There seems to exist an informational common core to causal and simulation theories, but little effort has been done to make sense of the informational character of declarative memory. In this section, I implement the minimal causal and simulation theories using tools of information theory. These theories end up to be are closely related, differing only on how memory transmits information about the past.

Classical information theory (Shannon 1948) provides a measure of the amount of information associated with a proposition. The theory measures the amount of information using the principle that the more probable a proposition is, the less information it carries. The amount of information associated with a proposition *b* ($h(b)$) is calculated as follows[32]:

31 Likewise, Robins (2017: 2) states that "remembering occurs when a person retains the capacity to represent information acquired from past events".

32 Until now, I have used the symbols *a*, *b*, and *c* to refer to faces (objects). For simplicity, I will use these same symbols to express the propositions that faces *a*, *b*, and

$$h(b) = -log_2(p(b)),$$

where $p(b)$ is the probability of b being true (e.g. b is the proposition that the event of S studying face b has occurred, see fn. 32). Point-wise mutual information (pmi) is a measure of the amount of information shared between two propositions. The pmi between propositions b and a ($i(b; a)$) is calculated as follows:

$$\begin{aligned} i(b; a) &= h(b) - h(b \mid a) \\ &= h(a) - \mathrm{h}(a \mid b). \end{aligned}$$

The amount of pmi between b and a is such that $-\infty \leq i(b; a) \leq min(h(b), h(a))$. If $i(b; a) < 0$, b and a are negatively correlated[33]. If $i(b; a) = 0$, b and a are independent. If $i(b; a) > 0$, b and a are positively correlated. In an intuitive reading, $i(b; a)$ means 'the amount of information that b transmits about a' (Dretske1981: 15-6)[34].

The minimal causal and reliability conditions are equivalent to two closely related informational conditions. The minimal causal condition states that S' «remembering» b with a given target is non-confabulatory only if the probability of S having b given that S had a previous experience a of that same target is higher than the unconditional probability of S having b ($p(b \mid a) > p(b)$). This condition is equivalent to the following informational condition:

c (respectively) exist and that the event of S «remembering» (studying, etc) faces a, b, and c (respectively) have occurred, etc. This will be convenient because, while the causal condition is defined in terms of earlier experiences, the reliability condition is defined in terms of targets. I hope the context will disambiguate the uses.

33 Philosophical work about pmi is scarce. For example, does negative pmi has meaning beyond negative correlation? Is pmi a real quantity, as, for example, the amount of mutual information, which is measured in bits? These issues will be relevant for a discussion in the Conclusions.

34 "We are now asking about the informational value of situation r, but we are not asking about $I(r)$. We are asking how much of $I(r)$ is information received from or about s. I shall use the symbol $I_s(r)$ to designate this new quantity. The r in parentheses indicates that we are asking about the amount of information associated with r, but the subscript s is meant to signify that we are asking about that portion of $I(r)$ that is information received from s. …$I_s(r)$ is a measure of the information in situation r about situation s" (Dretske1981: 15-6).

$$
\begin{aligned}
p(b \mid a) > p(b) \; &\leftrightarrow log_2(p(b \mid a)) > log_2(p(b)) \\
&\leftrightarrow -log_2(p(b)) > -log_2(p(b \mid a)) \\
&\leftrightarrow h(b) > h(b \mid a) \\
&\leftrightarrow h(b) - h(b \mid a) > 0 \\
&\leftrightarrow i(b; a) > 0
\end{aligned}
$$

In this interpretation, a «remembering» b with target a is non-confabulatory only if it transmits a positive amount of information about an earlier experience of a ($i(b; a) > 0$). For MCEs, the «remembering» of c transmits a positive amount of information about the earlier experience of face a. The same holds for b ($i(c; a) = i(c; b) = .2981 > 0$, see appx.). This result is consistent with a case of non-confabulatory «remembering», which corroborates the diagnosis of the minimal causal theory, but opposes those of the direct and provisional causal theories. This *absolute* informational condition (the analogy is with absolute/incremental Bayesian confirmation theories, see Huber 2007: §4c) vindicates the intuition that (non-confabulatory) memory transmits information about the past. This interpretation also provides answers to Bernecker's challenges. The answer to the first challenge is that $p(b \mid a)$ may be higher than $p(b)$ by any amount because any amount is enough for memory transmitting information about the past ($i(b; a) > 0$). The answer to the second challenge is the Appendix: we are able to calculate the relevant probabilities (if not for ecological) for laboratory cases and then work the other cases from analogy.

The reliability condition states that a «remembering» is non-confabulatory iff it is generated by a reliable memory process, where a memory process is reliable when it tends to produce accurate «remembering»s. It is reasonable to presuppose that a reliable memory process, under normal conditions, produces accurate «remembering»s with probability higher than 50%. Consequently, if a reliable memory process produces a «remembering» with content of the form $\gamma x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$ and target a, then there is a probability higher than 50% that $\gamma x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$. Since γxPx entails $\exists xPx$, there is a probability higher than 50% that $\exists x(P_1x \,\&\, P_2x \,\&\, \ldots \,\&\, P_nx)$. Since a is one of the individuals with the highest degree of satisfaction of this description, it is reasonable to expect that a is among the individuals

that verify the existential claim. Consequently, it is reasonable to expect that there is a probability higher than 50% that *a* exists/occurred with those features. In sum, if a «remembering» *b* with target *a* is generated by a reliable memory process, then it is reasonable to expect that $p(a \mid b) > .5$. This condition is also equivalent to an informational condition:

$$\begin{aligned} p(a \mid b) > .5 \quad & \leftrightarrow p(a \mid b) > p(\neg a \mid b) \\ & \leftrightarrow log_2(p(a \mid b)) > log_2(p(\neg a \mid b)) \\ & \leftrightarrow -h(a \mid b) > -h(\neg a \mid b) \\ & \leftrightarrow h(a) - h(a \mid b) > h(a) - h(\neg a \mid b) \\ & \leftrightarrow i(b; a) > i(b; \neg a) \end{aligned}$$

In this interpretation, if a «remembering» *b* with target *a* is non-confabulatory, then it is reasonable to expect that the «remembering» transmits more information about the existence/occurrence of *a* than about the opposite ($i(b; a) > i(b; \neg a)$). For MCEs, the «remembering» of *c* transmits more information about *a* being seeing in the study phase than about the opposite. The same holds for *b* ($i(c; a) = i(c; b) = .2981 > i(c; \neg a) = i(c; \neg b) = -.0348$, see appx.). This result is consistent with a case of non-confabulatory «remembering», which corroborates the simulationist diagnoses in §3.3 and §4.1. The simulationist would be an *incremental* informational condition (Huber 2007: §4c), which also vindicates the intuition that (non-confabulatory) memory transmits information about the past (at least it usually does, see Conclusions).

The informational reading reveals the importance of MCEs for the theory of memory: MCEs are limiting cases of non-confabulatory memory in a very precise sense. We are working with the data from Reinitz et al. (1992: E. 6), where the misremembering *c* has two targets (*a* and *b*), $i(c; a) = i(c; b) = .2981 > 0$ and $i(c, \neg a) = (c, \neg b) = -.0348$. This is consistent with a case of non-confabulatory «remembering» for both the minimal causal and simulation theories. However, there are results of MCEs where «remembering»s have more than two targets (e.g. Leding et al. 2007, where MCEs have three targets). Suppose that there are results of MCEs that maintain the structure of Reinitz et al., but that have three, four, …, *n* targets. If these cases behave as those of Reinitz et al., it is expected that, as *n* increases, both $i(c; a)$ and $i(c; \neg a)$

approach 0 and the limiting case would be of falsidical confabulation for both the minimal causal and simulation theories[35].

For a final point, the minimal causal condition is minimal in the sense that a causal theorist may adopt extra causal conditions. For example, causal theorists often demand causal connections to be sustained by an appropriate causal chain of memory traces. The adoption of extra causal conditions would further distinguish causal and simulation theories, even in informational implementation. However, MCEs also posit problems for this extra causal condition. Reinitz et al. argue that their results provide evidence for a distributed account of memory traces[36]. The problems non-minimal causal conditions exhibit when used to explain the causal relations between «remembering»s and earlier experiences in MCEs tend to reappear when they are used to explain the causal relations within a chain of discrete memory traces (see fn. 25). The appropriate causal chains would need to be of *distributed* memory traces. However, Robins argues that distributed memory traces do not provide a way to track their causal history and are incompatible with causal theories[37]. I will not discuss this issue here.

35 Reinitz et al. does not contain enough data for the calculation of $i(c; a)$ and $i(c; \neg a)$ for $n > 2$ (e.g. about feature stimuli with different proportions of old and new features). It is worthwhile to produce experimental results for the case $n = 3$ and check whether it has the same structure as the case $n = 2$ (i.e. with the frequency of feature stimuli lying between that of the conjunction and new stimuli). This would be evidence for the limiting behavior supposed above.

36 "The results provide strong evidence against any model that proposes that retrieval involves the activation of a single memory trace that represents a previously experienced stimulus… In such a system, there is no reason to expect that memory conjunction errors would occur, since memories are not composed of smaller features. In contrast, memory conjunction errors would be predicted if memories for related stimuli were stored as overlapping representations in which stimulus features constituted the representational units, as distributed memory models propose". Roughly, distributed models propose that "memory traces for previously experienced stimuli are represented as a set of units that roughly correspond to stimulus features" (Reinitz et al. 1992: 9 and 1).

37 "Distributed traces do not have individually distinguishable causal histories. …Individual traces do not leave a lasting, distinctive mark on the network by which their unique causal influence on a subsequent representation could be detected,

5 Conclusions

The causal and simulation theories are often presented as very distinct views about declarative memory, but the versions of these theories that can classify the memory errors in §3 are closely related when implemented using tools of information theory. These theories differ only on whether the informational criterion is absolute ($i(b; a) > 0$, causal theory) or incremental ($i(b; a) > i(b; \neg a)$, simulation theory). This difference has implications for the epistemology of memory. For example, while both theories are externalist, in the sense of not being transparent to the subject whether a «remembering» is confabulatory or not, the simulation theory has an extra layer of fallibilism. While the causal theory entails that non-confabulatory «remembering»s transmit information about the past, simulation theory is compatible with a non-confabulatory «remembering» with target a not transmitting (positive) information about a (when $i(c; \neg a) < i(c; a) < 0$).

'Transmitting information about the past', here, is a property not of the content of a «remembering», but of the process of generating it. Consequently, veridical confabulations (usually) do not transmit information about the past, which seems to be a correct consequence. Another consequence is that, although the distinction between non-confabulatory and confabulatory «remembering»s is well-defined, non-confabulatory «remembering»s can be less or more distant from being confabulatory. This seems to be correct because memory is a generative process and the proximity to confabulation is a measure of when this process goes wrong. A complete information theory of memory would consider not only with the amount of information transmitted by the process of generating a «remembering», but also by its content. This theory would be able to deal with the memory errors in §3 directly: how much of the amount of information transmitted by the content of a «remembering»is about its target?[38] An advantage of this theory would

much less distinguished from the influence of any other distributed pattern" (Robins 2016b: 16-7).

38 For example, the amount of (relevant) information in the «remembering» of *c* is of 6 bits (because there are 64 equally probable faces) whereas the amount of

be the possibility of treating the accuracy and informational conditions in a unified way[39]. Another advantage would be the possibility of investigating both poles of declarative memory using the same tools. After all, episodic (object-, event-, etc) and semantic (propositional) memory have contents that transmit information (about the past)[40].

Danilo Fraga Dantas
Federal University of Santa Maria (UFSM)
dfdantas@ucdavis.edu

Appendix

Each subject in Reinitz et al. (1992, E. 3-6) has studied 6 faces chosen randomly from a pool of 64 faces, constructed from the crossing of 8 sets of hair-and-mouth with 8 sets of eyes-and-nose. Let the studied faces be *aa*, *ab*, ...*hh*, where the first and second letters refer to hair-and-mouth and eyes-and-nose sets (respectively). Let *Sab* mean 'face *ab* was studied' and *Rab* mean 'face *ab* was recognized'. Let *Sxy* be a shorthand for $Saa \vee Sab \vee \ldots \vee Shh$ and $\neg Sxy$ be a shorthand for $\neg Saa \wedge \neg Sab \wedge \ldots \wedge \neg Shh$ (similarly for *Rxy* and $\neg Rxy$). The results of E. 6 are as follows: $p(Rab \mid Sab) = p(Rab \mid Sab, Sxy) = .71$ (target stimuli); $p(Rab \mid \neg Sab, Say, Sxb) = .52$ (conjunction stimuli); $p(Rab \mid \neg Sab, Say, \neg Sxb) =$

information that it transmits about a is of .2981.

39 There are some issues with the development of this theory. For example, there is an incongruence between the unit of information in the content of a «remembering» (e.g. bits) and the pmi used to measure the amount of information about the past. For example, bits are always positive and do not have a maximum, whereas the results pmi may be negative and are bounded by a maximum ($-\infty \leq i(b; a) \leq min(h(b), h(a))$). It is also not clear whether this theory should be developed using the classical or a semantic notion of information (see Floridi 2019).

40 I would like to thank Kourken Michaelian, César Schirmer dos Santos, and the members of UFSM's Philosophy of Memory Lab for their helpful comments. I am deeply grateful to Jaime Rebello and Paulo Faria, who were responsible for my learning about memory and Philosophy of Language. A previous version of this paper was presented on *Santa Maria-Grenoble Memory Workshop* (October 2018). I thank all the participants to this event (including André Sant'Anna) for their stimulating remarks. This work was funded by CAPES.

$p(Rab \mid \neg Sab, \neg Say, Sxb) = .19$ (feature stimuli); and $p(Rab \mid \neg Say, \neg Sxb) = .13$ (new stimuli). Let the faces *a* and *b* in §3.3 be the faces *aa* and *bb* (respectively). Face *c* is *ab* (hair-and-mouth *a* and eyes-and-nose *b*). The amount of information that the recognition of *c* transmits about the study of *a* ($i(Rab; Saa)$) and about the absence of study of *a* ($i(Rab; \neg Saa)$) are calculated as follows, where $C(n, m) = [n * (n-1) * \ldots * (n-m-1)]/m!$ is the number of combinations of size m from n elements[41]:

$$
\begin{aligned}
p(Rab \mid Saa) \quad &= \quad p(Rab \mid Saa, Sab) * p(Sab \mid Saa) + \\
&\qquad p(Rab \mid Saa, \neg Sab, Sxb) * p(\neg Sab, Sxb \mid Saa) + \\
&\qquad p(Rab \mid Saa, \neg Sxb) * p(\neg Sxb \mid Saa) \\
&= \quad .71 * C(62, 4)/C(63, 5) + \\
&\qquad .52 * [C(62, 5) - C(55, 5)]/C(63, 5) + \\
&\qquad 2 * \{.19 * [C(56, 6) - C(49, 6)]/C(64, 6)\} + \\
&\qquad .13 * C(49, 6)/C(64, 6) \\
&= \quad .71*.01 + .52*.23 + 2*(.19*.25) + .13*.19 \\
&= \quad .3024
\end{aligned}
$$

Roughly, the probability of face *ab* being studied ($p(Sab)$) is the ratio between the number of situations where *ab* is seen, i.e. where the set of six shown faces are composed of *ab* plus five other faces drawn from the 63 remaining ($C(63, 5)$), and the total number of cases where any six faces are shown from the pool of 64 faces (C(64, 6)). Similar reasoning applies for the other probabilities.

$$
\begin{aligned}
p(Rab \mid Saa) \quad &= \quad p(Rab \mid Saa, Sab) * p(Sab \mid Saa) + \\
&\qquad p(Rab \mid Saa, \neg Sab, Sxb) * p(\neg Sab, Sxb \mid Saa) + \\
&\qquad p(Rab \mid Saa, \neg Sxb) * p(\neg Sxb \mid Saa) \\
&= \quad .71 * C(62, 4)/C(63, 5) + \\
&\qquad .52 * [C(62, 5) - C(55, 5)]/C(63, 5) + \\
&\qquad .19 * C(55, 5)/C(63, 5) \\
&= \quad .71 * C(62, 4)/C(63, 5) + \\
&\qquad .52*[C(62, 5) - C(55, 5)]/C(63, 5) +
\end{aligned}
$$

41 The results for *b* are the same, substituting *bb* for *aa*.

$$
\begin{aligned}
& .19*C(55, 5)/C(63, 5) \\
= \;& .71*.08 + .52*.43 + .19*.50 \\
= \;& .3718
\end{aligned}
$$

$$
\begin{aligned}
p(Rab \mid \neg Saa) = \;& p(Rab \mid \neg Saa, Sab)*p(Sab \mid \neg Saa) + \\
& p(Rab \mid \neg Saa, \neg \mathrm{Sab}, Say, Sxb)*p(\neg Sab, Say, Sxb \mid \neg Saa) + \\
& p(Rab \mid \neg Saa, Say, \neg Sxb)*p(Say, \neg Sxb \mid \neg Saa) + \\
& p(Rab \mid \neg Say, Sxb)*p(\neg Say, Sxb \mid \neg Saa) + \\
& p(Rab \mid \neg Say, \neg Sxb)*p(\neg Say, \neg Sxb \mid \neg Saa) \\
= \;& .71 * C(62, 5)/C(63, 6) + \\
& .52*[C(62,6)\text{-}C(56,6)\text{-}C(55,6)+C(49,6)]/C(63,6) + \\
& 2 * \{.19 * [C(55, 6) - C(49, 6)]/C(63, 6)\} + \\
& .13 * C(49, 6)/C(63, 6) \\
= \;& .71*.10 + .52*.21 + .19*.22 + .19*.27 + .13*.21 \\
= \;& .2952
\end{aligned}
$$

$$
\begin{aligned}
i(Rab; Saa) &= h(Rab) - h(Rab \mid Saa) \\
&= -log_2[p(Rab)] + log_2[p(Rab \mid Saa)] \\
&= -log_2(.3024) + log_2(.3718) = .2981
\end{aligned}
$$

$$
\begin{aligned}
i(Rab; \neg Saa) &= h(Rab) - h(Rab \mid \neg Saa) \\
&= -log_2[p(Rab)] + log_2[p(Rab \mid \neg Saa)] \\
&= -log_2(.3024) + log_2(.2952) = -.0348.
\end{aligned}
$$

References

Aranyosi, István. 2020. Preteriception: memory as past-perception. *Synthese*.

Dalla Barba, Gianfranco. 2002. *Memory, Consciousness and Temporality*. Dordrecht: Springer.

Bernecker, Sven. 2008. *The Metaphysics of Memory*. Dordrecht: Springer.

Bernecker, Sven. 2010. *Memory: A Philosophical Study.* Oxford: Oxford University Press.

Bernecker, Sven. 2017. A causal theory of mnemonic confabulation. *Frontiers in Psychology* 8: 1207.

Berrios, German. 1998. Confabulations: a conceptual history. *Journal of the History of the Neurosciences* 7(3): 225–41.

Brainerd, Charles and Valerie Reyna. 2005. *The Science of False Memory*. Oxford: Oxford University Press.

Brown, Evan; Kenneth Deffenbacher, and William Sturgill. 1977. Memory for faces and the circumstances of encounter. *Journal of Applied Psychology* 62(3): 311–8.

Cummins, Robert. 1996. *Representations, Targets, and Attitudes*. Cambridge: MIT Press.

Debus, Dorothea. 2008. Experiencing the past: a relational account of recollective memory. *Dialectica* 62(4): 405–432.

Deese, James. 1959. On the prediction of occurrence of certain verbal intrusions in free recall. *Journal of Experimental Psychology* 58: 17–22.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.

Evans, Gareth. 1973. The causal theory of names. *Proceedings of the Aristotelian Society* 47: 187–208.

Floridi, Luciano. 2019. Semantic conceptions of information. In *The Stanford Encyclopedia of Philosophy,* ed. by Edward Zalta. URL = <plato.stanford.edu/entries/information-semantic>.

Hall, Ned and Laurie Paul. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Huber, Franz. 2007. Confirmation and induction. In *Internet Encyclopedia of Philosophy,* ed. by Fieser, James and Bradley Dowden. URL = <iep.utm.edu/conf-ind>.

Johnson, Marcia. 1997. Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society B: Biological Sciences* 352(1362): 1733–45.

Kripke, Saul. 1979. A puzzle about belief. In *Meaning and Use*, ed. by Avishai Margalit. Dordrecht: Reidel.

Kripke, Saul. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

Leding, Juliana; James Lampinen, Norman Edwards, and Timothy Odegard. 2007. The memory-conjunction error paradigm: normative

data for conjunction triplets. *Behavior Research Methods* 39(4): 920–5.

Levine, Linda. 1997. Reconstructing memory for emotions. *Journal of Experimental Psychology: General* 126(2): 165–77.

Lewis, David. 1973a. Causation. *Journal of Philosophy* 70(7): 556–67.

Lewis, David. 1973b. *Counterfactuals*. Hoboken: Wiley-Blackwell.

Lewis, David. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4): 455–76.

Loftus, Elizabeth; James Coan and Jacqueline Pickrell. 1996. Manufacturing false memories using bits of reality. In *Implicit Memory and Metacognition*, ed. by Reder, Lynne. London: Psychology Press.

Loftus, Elizabeth and Jacqueline Pickrell. 1995. The formation of false memories. *Psychiatric Annals* 25(12): 720–5.

Martin, Charles and Max Deutscher. 1966. Remembering. *Philosophical Review* 75: 161–96.

Michaelian, Kourken. 2016a. Confabulating, misremembering, relearning: the simulation theory of memory and unsuccessful remembering. *Frontiers in Psychology* 7: 1857.

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

Pearl, Judea; Madelyn Glymour, and Nicholas Jewell. 2016. *Causal Inference in Statistics: A Primer.* Hoboken: Wiley.

Perrin, Denis; Kourken Michaelian, and André Sant'Anna. 2020. The phenomenology of remembering is an epistemic feeling. *Frontiers in Psychology* 11: 1531.

Reichenbach, Hans. 1956. *The Direction of Time*. Berkeley: University of California Press.

Reinitz, Mark; William Lammers, and Barbara Cochran. 1992. Memory-conjunction errors: miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition* 20(1): 1–11.

Robins, Sarah. 2016a. Misremembering. *Philosophical Psychology* 29(3): 432–47.

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

Robins, Sarah. 2017. Contiguity and the causal theory of memory. *Canadian Journal of Philosophy* 47: 1–19.

Robins, Sarah. 2020. Mnemonic confabulation. *Topoi* 39: 121–32.

Roediger, Henry and Kathleen McDermott. 1995. Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21: 803–14.

Schacter, Daniel and Donna Addis. 2007. The ghosts of past and future: a memory that works by piecing together bits of the past may be better suited to simulating future events than one that is a store of perfect records. *Nature* 445(7123): 27.

Schooler, Jonathan and James Tanaka. 1991. Composites, compromises, and CHARM: what is the evidence for blend memory representations? *Journal of Experimental Psychology: General* 120(1): 96–100.

Searle, John. 1958. Proper names. *Mind* 67(266): 166–73.

Shannon, Claude. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423.

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality.* Amsterdam: North-Holland Pub. Co.

Tulving, Endel. 1985. Memory and consciousness. *Canadian Psychology* 26 (1): 1–12.

Whitehead, Alfred and Bertrand Russell. 1910. *Principia Mathematica, Vol. 1*. Cambridge: Cambridge University Press.