Episodic memory in causal reasoning about singular events

Sofiia Rappe (sofiia.rappe@ruhr-uni-bochum.de)
Department of Philosophy II, Universitätstraße 150
4478 Bochum, Germany

Markus Werning (markus.werning@ruhr-uni-bochum.de)
Department of Philosophy II, Universitätstraße 150
4478 Bochum, Germany

Abstract

Recent literature often presents memory as ultimately dealing with the future–helping the organism to anticipate events and increase its adaptive success. Yet, the distinct contribution of episodic (as opposed to semantic) memory to future-oriented simulations remains unclear. We claim that episodic memory yields adaptive success because of its crucial role in singular counterfactual causal reasoning, which thus far has been mostly ignored in the literature. Our paper presents a causal inference model based on the predictive processing framework and the minimal trace account of episodic memory. According to our model, evaluating the cause of an event involves (i) generating an episodic memory related to the said potential cause, (ii) constructing a counterfactual scenario through inhibition of the relevant part of the past episode, and (iii) temporal evolution followed by alternative model evaluation.

Keywords: counterfactuals; causal reasoning; predictive processing; episodic memory; trace minimalism; simulation

1. Introduction

Rather than focusing on how memory represents the past, recent philosophical and psychological literature emphasizes the functional role of memory in dealing with the future—anticipating events and increasing the organism's adaptive success (see, e.g., Addis and Schacter, 2013; Boyer, 2008; Klein, 2013; Schacter, 2012; Suddendorf & Corballis, 2007; Tulving, 2005). Memory-based simulation of future events directly aids goal-directed behavior (e.g., D'Argembeau & Mathy, 2011; Sheldon et al., 2011) and farsighted decision-making (e.g., Benoit, Gilbert, & Burgess, 2011; Boyer, 2008; Peters & Büchel, 2010), contributing to one's psychological well-being (Brown et al., 2002; Crisp & Turner, 2009) and self-concept development (Conway, 2005). Nevertheless, in most of these cases, it is difficult to determine the distinct contribution of episodic memory to future-oriented simulations contrasted with the simulations that draw from general knowledge about the world. Stanley Klein goes as far as to claim that "there is no principled (or empirical) reason to suppose that semantic memory […] does not make available the same memory content […] as does episodic memory" (Klein, 2013, p.228). Although semantic memory is ultimately gained through personal experiences, these experiences do not need to be stored, reconstructed, or explicitly represented by the brain for semantic memory to be accessible. Rather, semantic memory is a result of statistical learning and avails strict or probabilistic regularities through generalization. This raises a question: What is the distinct evolutionary advantage of episodic memory if it does not uniquely afford any of its alleged future-oriented functions? In other words, what is the point of explicitly constructing a scenario of the past? Contra Klein's claim (2013), we suggest that singular counterfactual causal reasoning presents one case where semantic memory cannot fulfill the role of episodic memory. This idea challenges most theoretical literature on causality, which has not acknowledged episodic memory as a cognitive factor in causal inference beyond its involvement in evidence accumulation. We propose a cognitively plausible model of singular counterfactual causal reasoning that relies on episodic memory based on predictive processing and trace minimalism frameworks.

2. Causal reasoning and counterfactuals

Causal reasoning is instrumental in many day-to-day situations, such as evaluating possible actions, learning and transmitting tool-use faculties, and navigating social interactions (Werning, 2009). "Indeed, the ability to attain causal understanding and harness it for diagnoses, predictions, and interventions is so advantageous that it has been considered the main driving force in human evolution" (Bender, 2020).

Yet, causal reasoning is not inherently human. Although researchers disagree on the extent of causal reasoning abilities in animals, it is generally accepted that many animal behaviors, such as tool use in monkeys and birds (Call & Tomasello, 1998; Santos et al., 2006; Sterelny, 2003; Taylor et al. 2007), cannot be attributed to mere associative learning. Moreover, the emergence of similar abilities in the species distantly located on the evolutionary tree indirectly supports the adaptivity of causal reasoning.

Starting with Hume's (1748), many modern theories of reasoning (such as Pearl's probabilistic approach, 2000) focus on information integration and, specifically, the accumulated statistical regularities strongly associated with general, semantic information, which is also reflected in the empirical methods and psychological study- design (Bender, 2020). For example, it is often assumed that predicting the future (and hence deciding which action to take) comes down to simulating the relevant scenario and letting the internal causal model update as if the future events would unfold over time according to specific statistical regularities stored in the semantic memory (Beck & Rafetseder, 2019). However, empirical evidence suggests that humans are great one-shot

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

learners and that it is often sufficient for us to encounter an event once to identify its cause (Schlottman & Shanks, 1992; White, 1999; Ahn & Kalish, 2000; Sloman, 2005), in other words, human causal reasoning "makes causal attributions on… much fewer data than would be required of a statistician's calculations to reach similar degrees of confidence" (Bowers, 2021, p.2). Finally, reliance on statistical regularities (and semantic information) can only be a part of the story because basing decisions on statistical regularities is outright detrimental when the future is affected by exceptional singular events.

Consider the following example. You may always take the road to the office through the city center because it is the shortest. However, yesterday, part of the road was closed for construction, and you were late for work. Therefore, today, you decided to take another road instead. The singular event of encountering a construction must be granted special status since, statistically speaking, one failure to get to the office on time using the usual road does not outweigh the long history of prior successes. Moreover, even though the decision to take a new road may at least partially rely on some probabilistic knowledge—e.g., that road constrictions often cause traffic jams and typically last more than one day—it also requires episodic remembering of encountering the road construction in the first place and subsuming this event under a regularity is outright counterproductive for the task of getting to the office on time.

The so-called black-sawn events provide another type of striking, although much less ubiquitous, examples where no subsumption of the singular event under a general strict or probabilistic regularity is available or purposive. Black swan events are high-impact outlier events that are difficult to predict under normal circumstances, e.g., a Coronavirus pandemic or a sudden stock market crash. However, once they occur, these events, in their individual instances, can be traced to specific causes. Both the black-swan events and the more mundane "unexpected interference" scenarios like the road construction case above demonstrate the potential impact of not just semantic but also episodic memory on future-oriented decision-making.

2.1 The counterfactual theory of causation

The question arises: How does one determine the cause of an event if not through subsumption under statistical regularities? The counterfactual theory of causation, which has also been developed to cover cases of singular causal dependencies, offers one approach. Even though its roots can be traced back to much earlier authors, such as Hume (1748), the counterfactual theory of causation gained momentum only after Lewis (1973) connected it with an analysis of counterfactual conditionals through possible world semantics. The following analysis expresses the central idea:

1) Given an actual event A (the cause) and a distinct, temporally succeeding actual event B (the effect), for event A to have caused event B, the following diachronic subjunctive conditionals must hold:
   a. If A were to happen, B would happen.
   b. If A had not happened, B would not have happened.

As David Lewis put it, the counterfactual theory of causation pays tribute to thinking of a cause as "something that makes a difference, and the difference it makes must be a difference from what would have happened without it" (1973, p. 161) Lewis's important innovation was to provide truth conditions for counterfactual conditionals in terms of a similarity metric defined over the set of (accessible) possible worlds, according to which some worlds are closer to the actual world than others:

2) "If X were the case, Y would be the case" is true in the actual world iff
   a. either there are no possible X-worlds;
   b. or no X-world that is a (not-Y)-world is closer to the actual world than any X-and-Y-world.

When we apply the possible world semantics for counterfactuals in (2) to the analysis of causation in (1), we may first disregard the disjunct (2a) of the former since an A-world is possible, and we may moreover disregard condition (1a) of the latter because it becomes trivially true. We thus arrive at the following analysis of singular causation:

3) An event A is a cause of an event B just in case no (not-A-and-B-world) is closer to the actual world than any (not-A-and-not-B)-world.

In other words, if one were to "travel" in the similarity space from the actual world—in which both A and B took place—to the nearest not-A world, this should be a world in which B is also absent. Only then A is a cause of B in the actual world.

Proposition (1), in connection with proposition (2), renders the simplest and most intuitive counterfactual analysis of causation. As Lewis (1973) himself noted, this analysis faces the problem that causation is (apparently) a transitive relation, but counterfactual conditionals are generally not transitive. Lewis, therefore, replaced simple counterfactual dependency with chain-wise counterfactual dependence as a necessary condition for causation. Others have argued that, even though counterfactual dependence is not strictly transitive, it still is weakly transitive. Additional objections raised against the counterfactual theory of causation have to do with problems of overdetermination and preemption (cf. Menzies & Beebee, 2020). However, these debates may be of greater relevance to the philosophy of science and the metaphysics of causation than our primary aim to understand and model real-world human cognition about causation.

2.2 Causation counterfactual cognition

Although counterfactual theories of causation do not need to be treated as descriptive of the psychological processes (Hoerl et al., 2011), it has been suggested that making a causal judgment about whether A caused B may, in some cases, explicitly involves "considering counterfactual simulations operating over a causal model of the situation" (Gerstenberg & Stephan, 2021, p. 2) and, more specifically, simulating the not-A world closest to the actual world to test (3). In his metaphysical analysis of causation, Lewis neglects the foregoing task of pre-selecting a potential cause A. However, for a cognitively plausible account of establishing a causal counterfactual, the first step should be such preselection, possibly through subjective, heuristical means (challenge 1). Secondly, to estimate the plausibility of (3), one needs to simulate the (not-A)-world that is the closest possible to the actual world at time t1, which is a rather complex task. This is a non-trivial task. First, alternative scenarios in which the potential cause A is absent (or not-A worlds) must be generated by some means (challenge 2). In addition, these possible worlds would have to be ranked against each other with respect to their closeness to the actual world. The situation is exacerbated by the fact that worlds diverge in a time-dependent manner. If the exact single change is made at two different time points, the respective worlds will diverge to a different extent. This part of the approach immediately strikes us as cognitively too demanding to be even approximately fulfilled by a biological system with all its limitations regarding working memory. Hence, what we think may be needed instead is a mechanism for the ad-hoc generation of "the closest possible" non-A world (challenge 3). Finally, the closest non-A scenario would have to be evaluated with respect to B: Is it also a non-B scenario rather than a B-scenario (challenge 4)?

In the following chapters, we outline one plausible model of causal inference based on counterfactuals with the help of trace minimalism (Werning, 2020) and predictive processing frameworks (Clark, 2013, 2015; Friston, 2005, 2010; Hohwy, 2013; Rao & Ballard, 1999). This combined approach addresses all four challenges, including providing the necessary machinery for the ad-hoc generation of the subjectively closest possible alternative to the A-world. This is made possible by the specifics of the efficient prediction-error-minimization-based model updating.

One important note is that although Lewis's counterfactual theory was formulated in terms of the objective similarity space of possible worlds, our approach relies on reasoning with subjective probabilities. Spohn (2012) provides one possible translation between the two. The details of this translation are irrelevant to our purpose, but essentially, "the closest possible" not-A world may be understood as the least surprising (or subjectively most probable).

3. Trace minimalism and predictive processing

Predictive processing is an influential framework for explaining cognition, perception, and action, which proposes that the brain's fundamental purpose is to function as a prediction engine (Clark, 2013, 2015; Friston, 2005, 2010; Hohwy, 2013; Rao & Ballard, 1999). The brain constantly minimizes the discrepancy (prediction error) between its predictions about the state of the world (which form a broadly causal generative hierarchical model) and the incoming sensory signal (in the case of perception, but other inputs may be used for the same purposes in other cognitive processes). Prediction error minimization is achieved by updating the internal generative model or bringing the predictions to life through action (i.e., *active inference*), and the set of predictions the system settles upon defines one's experiences.

The minimalist and efficient approach to bottom-up information involved in predictive processing has largely inspired Werning's (2020) trace minimalism framework of episodic memory. Trace minimalism draws upon predictive processing and scenario construction frameworks (Cheng et al., 2016), presenting episodic remembering as a "predictive" process. In perception, the brain produces predictions about the present based on the learned statistical regularities that are checked against the sparse sensory information. In contrast, in episodic recollection (or rather past scenario construction), the brain produces predictions about the past based on the learned statistical regularities that are checked against the minimal hippocampal trace—a carrier of non-categorial and sequential hippocampal information that serves as the causal link between experience and memory.

Because trace minimalism does not require storing explicit representational content but "an informationally sparse, merely causal link to a previous experience" (Werning, 2020, p. 301), it, on the one hand, resolves some of the difficulties associated with the Causal Theory of episodic memory (Martin & Deutscher, 1966), and on the other, provides the missing link to lived experiences for simulationism. Further, Werning's account presents an effortless addition to predictive processing by introducing a new source of prediction errors, which solves a significant problem—that of mental time travel to the past and, consequently, dealing with diachronic counterfactuals and causal reasoning that involves more than mere statistical regularities.

Counterfactual hypothesizing generally requires access to the past states of the generative model (or at least its parts). In bare-bones predictive processing, as new information comes along, the representations (hypotheses, predictions) are updated, "overwriting" the past states of the generative model. Thus, there is no straightforward way to trace the model back to its previous states (Hoerl & McCormack, 2019). It would be plausible to suggest that past states of the world may be predicted based on prior knowledge, just like the close-to-present state is being predicted in perception. However, prediction reliability in perception is ensured by the prediction error from the sensory input absent in the cases of past-oriented construction. As a new source of prediction error, minimal traces allow the cognitive system to simulate not just some possible but a "very close-to-actual" world at the time of the event linked to by the memory trace. Although a predictive agent may still be unable to directly trace back their (and their environment's) past states, they can construct them anew through simulation.

4. Counterfactual inference through episodic memory traces

According to our model, because of its truth condition (3), evaluating counterfactuals of type (1) involves a) episodic memory to construct a scenario of the past (Cheng et al., 2016), b) negation of the target event A in that scenario, and c) temporal evolution of the resultant non-A scenario to discover whether the evolved models is that of a B or non-B world. We describe these steps in more detail in the following sections.

Importantly, our approach implies that because the process of generating counterfactual scenarios involves the initial generation of the relevant episodic memories and engages minimal episodic memory traces that are associated with the hippocampus, we should expect to see relevant neural activation during the generation of past-directed counterfactual scenarios as well as causal reasoning tasks that involve causal attribution based on singular events. Empirical literature indeed suggests a partial overlap in brain activity during episodic remembering and episodic counterfactual thinking (De Brigard & Parikh, 2019), including in the hippocampus (Addis et al., 2009; Van Hoeck et al., 2013).

The link between episodic memory and episodic counterfactual thinking is further supported by behavioral and neurological evidence in clinical cases such as schizophrenia (Kwok et al., 2021) and hippocampal amnesia (Hooker, Roese, & Park, 2000; Contreras et al., 2016; Mullally & Maguire, 2014). Finally, developmental evidence shows similar age-related trajectories across past and counterfactual simulations in young children and older adults regarding the degree of employment of semantic vs. episodic information (De Brigard et al., 2016; Weisberg & Gopnik, 2013; see section 5 for a more extended discussion). Little, however, has been investigated directly regarding hippocampal activation during causal reasoning. This is not particularly surprising, given the general sparseness of the literature tying causal reasoning to episodic memory. Nevertheless, Fugelsang and Dunbar (2005) demonstrated that the hippocampal gyrus is involved in at least some cases of evaluating causal theories. With that in mind, we now processed directly to our model of counterfactual causal inference involving singular exceptional events.

4.1 Constructing the scenario of the past

According to trace minimalism, episodic remembering is predicting the past. The brain produces predictions regarding past events based on semantic information, learned statistical regularities, and embodied cues, matched against the minimal hippocampal traces of previous neural activity—non-representational links to previously experienced events, which act as the source of prediction errors. Although the nature of minimal traces is yet to be investigated, computational modeling has shown that the reconstruction of visual images by combining minimal traces with semantic information is robust and reliably approximates truth (Fayyaz et al., 2022). Hence, the scenario constructed through the processes of error minimization between the statistical-knowledge-driven predictions and hippocampal traces may be regarded as a reliable simulation of the past.

Once the counterfactual exploration is triggered, for example, by low certainty of predictions regarding causal attribution in the relevant layers of the probabilistic generative model, the first step in causal reasoning through counterfactuals is to generate the episodic memory, which contains the possible cause of the target event. Although the pool of potential possible causes is strictly speaking infinite, some systematic factors or heuristics guide people's selection of potential variables to evaluate as causes, including statistical and prescriptive normative expectations, the purpose of inquiry, or the event structure of the situation (Gerstenberg & Stephan, 2021). The specific dimensions of the selection of possible causes require further investigation. However, the general agreement in psychological research is that such selection is intuitively based on what "naturally" comes to mind. The initial pool of possible causes is thus subjectively restricted. On the one hand, this presents an efficient processing strategy; on the other, it may exclude the relevant potential causes altogether, leading to wrong causal inferences (which is entirely consistent with human performance).

The specification problem is generally challenging when considering causes by omissions. When a specific action does not occur, it may be unclear what the relevant contrastive event should look like (Halpern & Hitchcock, 2015; Schaffer, 2005, 2010). This problem, however, does not come up for predictive processing. Since any predictions deemed by the system as incorrect are conditionally dependent on the predictions above and below, when the system is forced to come up with the alternative scenario, it does so holistically—settling on a set of new predictions that provide the best global fit (best minimize the prediction error) and automatically filling in the necessary details, whether the content is strictly contrastive to the rejected case or not.

Once the possible cause is selected, the relevant episodic scenario is generated based on general knowledge about the world, contextual information, and embodied cues, which are matched against the minimal episodic trace per the trace minimalism account (Werning, 2020).

4.2 Constructing alternatives to the past

The counterfactual scenario (not-A world) may be simulated by negating the target event A in the constructed scenario. This can be accomplished by inhibiting the scenario's relevant aspects. According to predictive processing and similar approaches, the content of the episodic scenario is presented by the hierarchical set of predictions represented in the brain. Inhibiting the parts of the network that represent target events leads the system to alter the scenario accordingly.

One advantage of the predictive-type Bayesian architectures is that negating a prediction through inhibition does not necessarily result in a scenario with the mere absence of the transpired event. Instead, the system re-stabilizes with different local predictions that fit in with the higher-level high-certainty predictions as closely as is deemed good enough by the system. These new predictions may or may not pertain to similar agents or acts related to the inhibited predictions. To illustrate, when inhibited, the prediction "Kelly did not do anything" may be substituted for a "positive" event like "Kelly asked for help" or "Kelly ran away" depending on the systems state and likelihood of each given prediction—in other words, inhibition does not always result in the negation in the strict linguistic sense but rather a substitute of the relevant predictions for the globally closest alternative fit (Figure 1). Importantly, this approach allows the system to generate the "subjectively" closest alternative world without generating many alternative worlds and without any explicit comparison to the actual world.

4.3 Model selection

Once the system identified the most subjectively probable counterfactual model of the world at time t1 in which A is absent, our cognitive system has to temporally evolve this counterfactual model according to learned statistical regularities (involving general causal regularities) to arrive at a model of the world at time t2 (Figure 1). The next step would be evaluating whether the resulting model at t2 is a model in which B is also absent. If it is absent, the conclusion should be that the counterfactual conditional 'if A had not happened, B would not have happened' supports the proposition that A is a cause of B (see principle 3).

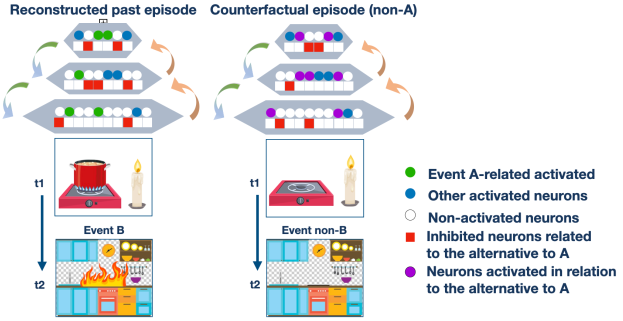
Reconstructed past episode
Counterfactual episode (non-A)
t1
t2
Event B
t1
t2
Event non-B
Event A-related activated
Other activated neurons
Non-activated neurons
Inhibited neurons related to the alternative to A
Neurons activated in relation to the alternative to A

Figure 1: Alternative scenario construction.

Our approach (Figure 2) provides an efficient way of dealing with causal relations that are too specific to be effectively subsumed under general causal rules. However, our approach is heuristics-based, which does not guarantee a correct solution, only a good enough for the task at hand, given the context and specific agent's knowledge. First, networks are probabilistic and not deterministic (even if the causes are). Second, the pre-selection of causal variables may exclude the actual cause of an event. Third, the prediction error minimization process at the root of the memory construction, counterfactual generation, and model comparison is satisficing and not optimizing. Once the error is undetectable to the cognitive system, the model will settle into a stable state regardless of the proximity of the model to the actual state of affairs. Nevertheless, it should not be surprising that this approach should still work sufficiently well for many daily situations—the simulation is rooted in the predictive generative model validated throughout the individual's life span.

To warrant the reliability of temporal evolution leading from non-A scenario at t1 to B or not-B scenario at t2, we can recruit general regularities and multiple episodic memories of previous encounters or similar situations. Each episodic memory comes with a temporal sequence of reconstructable events. Interlacing those sequences might enable us to reliably bridge the time gap between t1 and t2 in the temporal evolution of the counterfactual generative predictive model at t1 (Parra-Barrero & Cheng, 2023).

Importantly, singular counterfactual reasoning, as described by our model, is fast and efficient regarding working memory and other cognitive resources. Even if multiple episodic traces are recruited and multiple episodes of the past are constructed simultaneously, such recruitment is a case of one-shot learning of temporal sequences and not inductive inference. Unlike Lewis' original framework, our approach avoids comparisons of multitudes of multiple worlds—a task that would be too computationally demanding for a biological system to engage in. Finally, singular counterfactual reasoning does not require any special "hardware" or resources to represent or learn transtemporal or diachronic rules, regularities, or generalizations. All that is required is a hierarchical predictive model, a mechanism for storing minimal episodic traces, and a capacity to do inhibition (often associated with the pre-frontal cortex). Thus, evolutionarily, it does not require dedicated radical innovation.

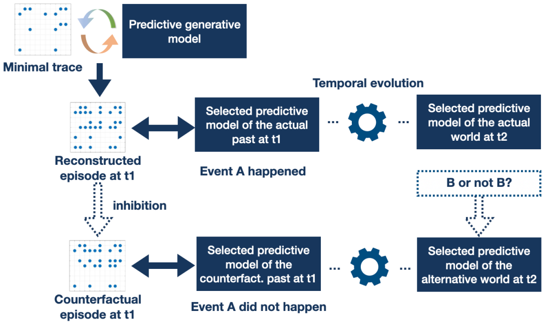
Minimal trace
Predictive generative model
Temporal evolution
Reconstructed episode at t1
Selected predictive model of the actual past at t1
Event A happened
Selected predictive model of the actual world at t2
B or not B?
inhibition
Counterfactual episode at t1
Selected predictive model of the counterfact. past at t1
Event A did not happen
Selected predictive model of the alternative world at t2

Figure 2: A model of singular counterfactual causal reasoning.

5. Developmental perspective

Updating one's causal model through counterfactual simulation has substantial benefits but may be employed selectively. Empirical research shows that reliance on memory-traces-based counterfactual simulation changes throughout one's lifetime. These age-related changes are a natural consequence of our developmental trajectory.

5.1 Counterfactual reasoning in early childhood

Empirical literature in developmental psychology suggests that children can only adequately engage in counterfactual reasoning once they are at least 5-6 years old (Rafetseder et al., 2013). Notably, at this stage, they possess episodic memory but are prone to confabulation and mixing reality and fantasy. This issue may stem from the lack of proper development of reality monitoring mechanisms necessary for simulations involving atemporal or past-oriented temporal departures from the current generative model.

The overlap between perception (stimuli-dependent content) and imagination (self-generated content) poses the problem of inferring whether the source of the neural activity is more likely to be stimulus-dependent or self-generated. However, the issue is more fine-grained. For example, Deroy and Rappe (2022) discuss the cases of extraordinary perception, where, despite a clear understanding that one's experience is real, the person still experiences a certain sense of confusion regarding the way that reality feels. They argue that our subjective sense of reality is a composite of several subjective markers, including a categorical one that can identify an experience as perceptual. However, it also involves residual feelings of control, sensory incongruence, low multisensory confidence, or feelings of wonder that introduce layers or dimensions of subjective confusion. Even though a subjective categorical marker of perceptual reality can (and evolutionarily should) be reasonably robust early in development, as metacognitive processes and multisensory congruence develop in the first few years of life, the sense of perceptual reality should become richer but also more subjectively confusing (see Goupil and Kouider, 2019). Because the subjective sense of reality and categorical reality monitoring are not independent (see Rappe and Wilkinson, 2022), this developmental process can affect reality monitoring, one's perceptual experience, and the capacity to entertain counterfactual hypotheses (e.g., Nyhout and Ganea, 2019).

The negation-through-inhibition aspect of causal reasoning with counterfactuals could present another potential challenge for young children, as the acquisition of negation as rejection ("I do not want juice") developmentally precedes negation as non-existence ("There is no juice") (Cucio 2011; Dimroth 2010; Tagliani et al., 2022; Vaidyanathan, 1991). Although negation as non-existence is typically acquired before proper reality monitoring is developed, it still presents a necessary developmental step toward counterfactual cognition. Only once all the elements of the counterfactual machinery finally take shape can a child take full advantage of counterfactual inference.

5.2 Counterfactual reasoning in later adulthood

Recent literature also highlights age-related differences in reliance on episodic memory between younger and older adults. For example, according to Addis and colleagues (2011), the contribution of episodic memory in constructing autobiographical events diminishes in older adults. The details are filled with the help of semantic, conceptual information. One possible explanation is that episodic specificity for episodic memories and future simulations is reduced with age. For example, De Brigard et al. (2017) found that younger adults generate more episodic details across all simulations of past, future, and counterfactual events than older adults (see also Levine et al., 2002; Addis et al. 2008, 2010; Gaesser et al., 2011). At the same time, older adults generate more semantic and contextual details than younger adults, which a richer generative model may explain. Interestingly, this was only true for episodic future and counterfactual simulations but not for episodic memories. One possible explanation of that effect is that while episodic memory construction is constrained and rooted in episodic traces, future and counterfactual simulations are more top-down, semantic-information-driven processes, even if they require engaging episodic memory traces at some steps.

The reduction of specificity of episodic memory in older adults may be related to the reduction of precision weightings on episodic traces for the construction of scenarios of the past (Korkki et al., 2020). Since precision weighting is commonly taken to be mediated by dopamine (Haarsma et al., 2021), the decline of precision of memory traces in older adults is indirectly supported by the studies reporting an age-related decline in striatal and extrastriatal dopamine. Furthermore, adults over 60 show multiple hippocampal alterations, including structural and connectivity changes and reduced hippocampal activity commonly associated with episodic memory deficiencies (Nyberg, 2017).

According to our proposal, such age-related changes in episodic memory should also diminish the effectiveness and degree of reliance on counterfactual cognition in decision-making. Although this hypothesis remains to be tested, Lempert et al. (2022) found that older people tended to perform worse at making adaptive episodic memory-based choices. In the social domain, older adults were more likely to engage with people they recognized, even if they episodically remembered that the previous interaction was unfair. Further, "older adults were more influenced by appearances, choosing to interact with others that are perceived as more generous, even though those perceptions did not accord with past experience" (Lempert et al., 2022, p. 7). These findings suggest that older adults rely more on statistical regularities than individual events in decision-making.

6. Conclusion

We argued that episodic memory may play a vital role in counterfactual causal reasoning involving singular events where subsumption of the said events under a set of general principles or statistical regularities is not possible or purposive. Our proposed model of singular counterfactual inference, rooted in predictive processing and trace minimalism, is efficient with regard to cognitive resources, does not require dedicated causal-reasoning-specific machinery, and allows the cognitive agent to learn from sparse data. However, the model requires a realistic proposal regarding its neural implementation. Little is known regarding the neural mechanisms of counterfactual simulation, prediction error minimization, or even the structure and principles of the hierarchical generative networks. Once some such specific proposals are formulated, new methods and tools may also be required for a more targeted, discriminating hypothesis evaluation.

Acknowledgments

This work has been supported by the grants no. 419040015, no. 419038924, and no. 397530566 from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the DFG research group Constructing Scenarios of the Past (FOR 2812).

References

Addis, D. R., Musicaro, R., Pan, L., & Schacter, D. L. (2010). Episodic simulation of past and future events in older adults: Evidence from an experimental recombination task. P*sychology and Aging*, 25, 369–376.

Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, *47*(11), 2222-2238.

Addis, D. R., Roberts, R. P., & Schacter, D. L. (2011). Age-related neural changes in autobiographical remembering and imagining. *Neuropsychologia*, *49*(13), 3656-3669.

Addis, D. R., & Schacter, D. L. (2013). Future-oriented simulations: The role of episodic memory.

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. *Psychological science*, *19*(1), 33-41.

Ahn, W. K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. *Explanation and cognition*, 199-225.

Beck, S. R., & Rafetseder, E. (2019). Are counterfactuals in and about time?.*Behavioral and Brain Sciences*, *42*.

Bender, A. (2020). What is causal cognition? *Frontiers in Psychology*, *11*, 3.

Benoit, R. G., Gilbert, S. J., & Burgess, P. W. (2011). A neural mechanism mediating the impact of episodic prospection on farsighted decisions. Journal of Neuroscience, 31, 6771–6779.

Bowers, R. I. (2021). Causal reasoning. *Encyclopedia of evolutionary psychological science*, 920-936.

Boyer, P. (2008). Evolutionary economics of mental time travel? Trends in Cognitive Sciences, 12, 219–224.

Brown, G. P., MacLeod, A. K., Tata, P., & Goddard, L. (2002). Worry and the simulation of future outcomes. Anxiety, Stress, and Coping, 15, 1–17.

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes) and human children (Homo sapiens). *Journal of Comparative Psychology*, *112*(2), 192.

Cheng, S., Werning, M., & Suddendorf, T. (2016). Dissociating memory traces and scenario construction in mental time travel. *Neuroscience & Biobehavioral Reviews*, *60*, 82-89.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Contreras, F., Albacete, A., Castellví, P., Caño, A., Benejam, B., & Menchón, J. M. (2016). Counterfactual reasoning deficits in schizophrenia patients. *PLoS One*, *11*(2), e0148440.

Conway, M. A. (2005). Memory and the self. *Journal of memory and language*, *53*(4), 594-628.

Crisp, R. J., & Turner, R. N. (2009). Can imagined interactions produce positive perceptions?: Reducing prejudice through simulated social contact. *American psychologist*, *64*(4), 231.

Cuccio, V. (2011). On Negation. What do we need to "say no"?. *Rivista Italiana di Filosofia del Linguaggio*, *4*, 47-55.

D'Argembeau, A., & Mathy, A. (2011). Tracking the construction of episodic future thoughts. Journal of Experimental Psychology: General, 140, 258–271

De Brigard, F., Giovanello, K. S., Stewart, G. W., Lockrow, A. W., O'Brien, N. M., & Spreng, R. N. (2016). Characterizing the subjective experience of episodic past, future, and counterfactual thinking in healthy younger and older adults. The Quarterly Journal of Experimental Psychology, 69, 2358–237

De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, *28*(1), 59-66.

De Brigard, F., Rodriguez, D. C., & Montañés, P. (2017). Exploring the experience of episodic past, future, and counterfactual thinking in younger and older adults: A study of a Colombian sample. *Consciousness and cognition*, *51*, 258-267.

Deroy, O., & Rappe, S. (2022). The clear and not so clear signatures of perceptual reality in the Bayesian brain. *Consciousness and Cognition*, *103*, 103379.

Dimroth, C. (2010). The Acquisition of Negation. In L. Horn (Ed.), *The Expression of Negation*. Berlin, New York: De Gruyter Mouton.

Fayyaz, Z., Altamimi, A., Zoellner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022). A model of semantic completion in generative episodic memory. *Neural Computation*, *34*(9), 1841-1870.

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, *28*(6), 764–766.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*(8), 1204-1213.

Gaesser, B., Sacchetti, D. C., Addis, D. R., & Schacter, D. L. (2011). Characterizing age-related changes in remembering the past and imagining the future. Psychology and Aging, 26,80–84.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.Goupil and Kouider, 2019)

Haarsma, J., Fletcher, P. C., Griffin, J. D., Taverne, H. J., Ziauddeen, H., Spencer, T. J., ... & Murray, G. K. (2021). Precision weighting of cortical unsigned prediction error signals benefits learning, is mediated by dopamine, and is impaired in psychosis. Molecular psychiatry, 26(9), 5320-5333.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*.

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences*, *42*.

Hoerl, C., McCormack, T., & Beck, S. R. (Eds.). (2011). *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*. Oxford University Press.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hooker, C. I., Roese, N. J., & Park, S. (2000). Impoverished counterfactual thinking is associated with schizophrenia. Psychiatry, 63, 326–335.

Hume, D. (1748/1988). An Enquiry Concerning Human Understanding. La Salle, IL: Open Court (Originally published 1748).

Klein, S. B. (2013). The temporal orientation of memory: It's time for a change of direction. *Journal of Applied Research in Memory and Cognition*, *2*(4), 222-234.

Korkki, S. M., Richter, F. R., Jeyarathnarajah, P., & Simons, J. S. (2020). Healthy ageing reduces the precision of episodic memory retrieval. *Psychology and Aging*, *35*(1), 124.

Kwok, S. C., Xu, X., Duan, W., Wang, X., Tang, Y., Allé, M. C., & Berna, F. (2021). Autobiographical and episodic memory deficits in schizophrenia: A narrative review and proposed agenda for research. *Clinical psychology review*, *83*, 101956.

Lempert, K. M., Cohen, M. S., MacNear, K. A., Reckers, F. M., Zaneski, L., Wolk, D. A., & Kable, J. W. (2022). Aging is associated with maladaptive episodic memory-guided social decision-making. *Proceedings of the National Academy of Sciences*, *119*(42), e2208681119.

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Ageing and autobiographical memory: Dissociating episodic from semantic retrieval. Psychology and Aging, 17, 677–689.

Lewis, D. (1973). Counterfactuals and comparative possibility. In *IFS: Conditionals, Belief, Decision, Chance and Time*. Dordrecht: Springer Netherlands.

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

Menzies, P., & Beebee, H. (2020). Counterfactual Theories of Causation [M/OL]. *Zalta E N. The Stanford Encyclopedia of Philosophy.*

Mullally, S. L., & Maguire, E. A. (2014). Counterfactual thinking in patients with amnesia. Hippocampus, 24, 1261–1266.

Nyberg, L. (2017). Functional brain imaging of episodic memory decline in ageing. *Journal of internal medicine*, *281*(1), 65-74.

Nyhout, A., & Ganea, P. A. (2019). The development of the counterfactual imagination. *Child Development Perspectives*, *13*(4), 254-259.

Parra-Barrero, E., & Cheng, S.. (2023). Learning to predict future locations with internally generated theta sequences. PLOS Computational Biology, 19(5), e1011101.

Pearl, J. (2000). Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, *19*(2), 3.

Peters, J., & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*(1), 138-148.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of experimental child psychology*, *114*(3), 389-404.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79-87.

Rappe, S., & Wilkinson, S. (2022). Counterfactual cognition and psychosis: adding complexity to predictive processing accounts. *Philosophical Psychology*, *36*(2), 356-379.

Santos, L. R., Pearson, H. M., Spaepen, G. M., Tsao, F., & Hauser, M. D. (2006). Probing the limits of tool competence: Experiments with two non-tool-using species (Cercopithecus aethiops and Saguinus oedipus). *Animal cognition*, *9*, 94-109.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773-786.

Schaffer, J. (2005). Contrastive causation. The Philosophical Review, 114(3), 327–358.

Schaffer, J. (2010). Contrastive causation in the law. Legal Theory, 16(04), 259–297.

Schlottman, A., and Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. Q. J. Exp. Psychol. Hum. Exp. Psychol. 44, 321–342. doi: 10.1080/02724989243000055

Sheldon, S., McAndrews, M. P., & Moscovitch, M. (2011). Episodic memory processes mediated by the medial tem

Sloman, S. A. (2005). Causal Models: How We Think About the World and its Alternatives.NewYork, NY: OxfordUniversityPress.doi: 10.1093/acprof:oso/9780195183115.001.0001

Spohn, W. (2012). The laws of belief: Ranking theory and its philosophical applications (1st ed ed.). Oxford: Oxford University Press.

Sterelny K (2003) Thought in a Hostile World. The evolution of human cognition. Blackwell, Oxford.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans?. *Behavioral and brain sciences*, *30*(3), 299-313.

Tagliani, M., Vender, M., & Melloni, C. (2022). The acquisition of negation in Italian. *Languages*, *7*(2), 116.

Taylor AH, Hunt GR, Holzhaider JC, Gray RD (2007) Spontaneous metatool use by New Caledonian crows. Curr Biol 17:1504–1507.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human. *The missing link in cognition: Origins of self-reflective consciousness*, 3-56.

Vaidyanathan, R. (1991). Development of forms and functions of negation in the early stages of language acquisition: a study in Tamil. Journal of Child Language, 18(1), 51-66.

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. Social Cognitive and Affective Neuroscience, 8, 556–564. doi:10.1093/scan/nss031

Weisberg, D. S., & Gopnik, A. (2013). Pretense, counterfactuals, and Bayesian causal models: Why what is not real really matters. *Cognitive Science*, **37**, 1368–1381. doi:10.1111/cogs.12069.

Werning, M. (2009). The evolutionary and social preference for knowledge: How to solve Meno's problem with reliabilism. *Grazer Philosophische Studien*, *79*(1).

Werning, M. (2020). Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of philosophy and psychology*, *11*, 301-333.

White, P. A. (1999). Toward a causal realist account of causal understanding. Am. J. Psychol. 112, 605–642. doi: 10.2307/1423653