Title
Dynamics of Counterfactual Retrieval

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Wang, Feiyi
Aka, Ada
Bhatia, Sudeep

Publication Date
2021

Peer reviewed

Dynamics of Counterfactual Retrieval

Feiyi Wang (feiyiw@sas.upenn.edu)
Department of Psychology, 425 S. University Ave,
Philadelphia, PA, 19104-6241 USA

Ada Aka (adaaka@wharton.upenn.edu)
Psychology and Marketing, 425 S. University Ave,
Philadelphia, PA, 19104-6241 USA

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, 425 S. University Ave,
Philadelphia, PA, 19104-6241 USA

Abstract

People often think about counterfactual possibilities to an event and imagine how it could have been otherwise. The study of how this occurs is central to many areas of cognitive science, including decision making, social cognition, and causal judgment; however, modeling the memory processes at play in naturalistic counterfactual retrieval has been difficult. We use established memory models to evaluate and compare multiple mechanisms that could be involved in counterfactual retrieval. Our models are able to capture nuanced dynamics of retrieval (e.g. how retrieved counterfactuals cue subsequent counterfactuals), and can predict the effects of retrieval on evaluations and decisions. In doing so, we show how existing theories of counterfactual thinking can be combined with quantitative models of memory search to provide new insights about the formation and consequences of counterfactual thought.

Keywords: counterfactual thinking; memory; computational models; vector semantics

Introduction

Counterfactual thinking, or the ability to imagine alternative possibilities to one's experience, is ubiquitous (Byrne, 2016; De Brigard & Parikh, 2019; Phillips, Morris, & Cushman, 2019). Once in mind, counterfactual thoughts have a wide array of effects on cognition and behavior. For example, judgments of causality depend on salient counterfactuals, and counterfactual assessment is a key component in cognitive models of causal judgment (Gerstenberg & Tenenbaum, 2017; Sloman & Lagnado, 2015; Wells & Gavanski, 1989). In social settings, counterfactuals determine judgments of responsibility and the moral evaluations of acts (Greene et al., 2004; Zultan et al., 2012). Finally, decisions rely critically on counterfactuals, with desirable counterfactuals reducing the judged desirability of target choice outcomes (Loomes & Sugden, 1982; Mellers et al., 1997; Stewart et al., 2006).

Unsurprisingly, understanding how counterfactual thoughts are retrieved from memory is of considerable interest to cognitive scientists, psychologists, and neuroscientists. Given the limited capacity of working memory, counterfactual thoughts that spontaneously come to mind are only a sample of the vast possibilities that an individual can imagine. Recent literature has shown that subjective desirability and probability influence what comes to mind by default (Bear et al., 2019) or in decision tasks (Griffiths & Tenenbaum, 2006). This led some researchers to propose that the default counterfactual thoughts are thoughts that are both highly desirable and probable (i.e., have a high likelihood of occurrence) (Phillips et al., 2019).

In the well-known norm theory, Kahneman and Miller (1986) argued that people create imagined alternatives of an experience by mentally simulating its exceptional aspects as normal ones. Interestingly, people have been found to be more likely to generate counterfactual thoughts after a "near-miss" situation, in which the perceived probability that a counterfactual would have actually occurred is high (Roese & Epstude, 2017). Judgments of counterfactual plausibility have been further shown to be mediated by the perceived similarity between counterfactuals and the actual event (De Brigard et al., 2021).

In most of the above settings, the generation of counterfactual thoughts requires retrieval from memory. Yet, formal models of the memory processes at play in counterfactual generation have not been developed. Such models are widely used in the study of list recall (Polyn, Norman, & Kahana, 2009; Raaijmakers & Shiffrin, 1981) and semantic memory search (Abbott, Austerweil, & Griffiths, 2012; Hills, Jones, & Todd, 2012). In the domain of counterfactual generation, such models can be used to parameterize the effects of distinct retrieval mechanisms and cues and quantitatively test which mechanisms play the largest role.

Such models are also necessary to describe and predict the sequences of counterfactual thoughts that come to mind in response to a particular outcome or event. As retrieved counterfactuals can cue the subsequent retrieval of semantically related counterfactuals (i.e., semantic clustering), the dynamics of retrieval may be quite complex and may have subtle effects on downstream tasks like causal judgment and decision making. Semantic clustering is a well-known phenomenon in memory research (Howard & Kahana, 2002; Bhatia, 2019; Hills et al., 2012), yet its role in counterfactual generation and evaluation has not been examined or controlled for in prior work.

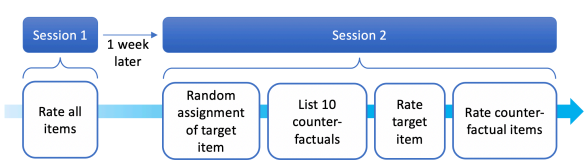
Finally, parametric cognitive models of counterfactual generation can be used to characterize the structure of variability in counterfactual thoughts across individuals, explain influences of task and context, and predict counterfactuals generated in response to novel stimuli with high accuracy. For this reason, modeling the memory processes involved in counterfactual retrieval can add substantial rigor and detail to our understanding of counterfactual thought.

In this paper, we build a formal parametric model of counterfactual retrieval. Our model takes the form of a Markov random walk over items in memory (Abbott et al., 2012). The Markov random walk is a basic model of memory search that emerges as a special case from more complex models (such as Polyn et al., 2009; Hills et al., 2012; and Raaijmakers & Shiffrin, 1981). In our application of this model to counterfactual retrieval, we assume that people generate counterfactuals sequentially and that the probability of transitioning from one counterfactual to the another depends on a number of variables, including variables studied in prior work (such as desirability, probability, and similarity to the target). Critically, we also allow for the effect of new variables (such as semantic relatedness with the previously retrieved item) that are implicated in memory search but have not been studied in counterfactual generation tasks.

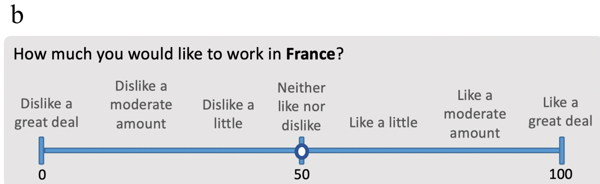
We test our model using open-ended counterfactual generation tasks. In our tasks, participants are shown one target item from a particular event, and then asked to list 10 items that came to mind as they consider the target item. After recalling these items, participants are also asked to evaluate the target item. To ensure the effects we observe are not unique to the scenario employed, we conduct three experiments using a similar procedure but different word pools and evaluation contexts. By fitting our model to data from these experiments, we aim to formally model the determinants and consequences of counterfactual thinking.

Method

a



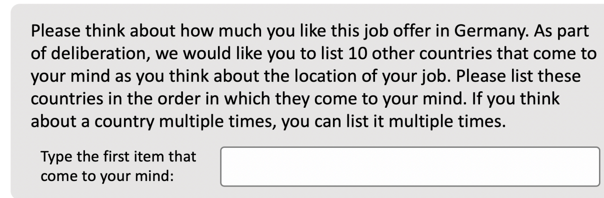
b



c

Figure 1: (a) Schematic of the task design. (b) Example of a desirability rating question in the first session. (c) Example of the counterfactual generation task in the second session.

Participants



Participants in Study 1 ($N = 53$; mean age = 20; 55% female), participants in Study 2 ($N = 53$; mean age = 32; 56.6% female), and participants in Study 3 ($N = 40$; mean age = 20; 80% female) performed the experiment online. All participants in Study 1 and 3 were undergraduate students at the University of Pennsylvania. Participants in Study 2 were recruited via Prolific Academic and received monetary compensation at a rate of $9/hr.

Procedure

All experiments gave participants a hypothetical scenario involving various outcomes. Study 1 used a job offer scenario, in which participants were told that they received a job offer in a given country and were asked to retrieve counterfactual countries that came their minds. Study 2 used a vacation travel scenario, in which participants were told that they had won a sweepstake to travel to a given country and were asked to retrieve counterfactual countries that came their minds. Finally, Study 3 used a food tasting scenario, in which participants were told that they were offered to taste a given fruit or vegetable and were asked to list other fruits or vegetables that came to their minds.

Our three experiments had very similar procedures, so we report them together here. Each experiment contained two sessions: (i) the baseline ratings, and (ii) counterfactual generation. These two sessions were separated by a week to reduce the potential effects of memorization (Figure 1a). In the first session, participants were asked to give their subjective desirability and probability ratings for a large set of items that could be at play in the experimental scenario (for example, in Study 1 participants were asked to rate the desirability of job offers in 193 countries, and their probability of taking job offers in these countries). Ratings were made on a scale from 0 to 100 (see Figure 1b; 0 corresponds to extremely undesirable/improbable, 50 to neither desirable/probable nor undesirable/improbable, and 100 to extremely desirable/probable).

In the second session, participants were randomly shown one of four target items. Next, participants were asked to list 10 counterfactual items that came to their mind as they considered the target item in the scenario (Figure 1c). Participants listed these counterfactuals in the order in which they came to mind on 10 successive screens. Next, participants were taken to two separate screens where they rated the target item in terms of its desirability and probability, respectively. All ratings were on the same scale as described earlier, and these questions were self-paced.

For Study 1 (the job offer study), we created a word pool of countries using the 193 member states of the United Nations (as of February 7, 2020). To generate the target items, we first retrieved 300-dimensional semantic vector representations of each country using Google's Word2Vec model (Mikolov et al., 2013). We applied multidimensional scaling on these vector representations to visualize all items on a two-dimensional graph. By inspecting this graph, we selected target items from different clusters that emerged in the visualizations. Our target items for Study 1 were Germany, Kenya, Guatemala, and Saudi Arabia. Study 2 (the vacation travel study) used the same word pool as Study 1, and we selected France, Costa Rica, Japan and South Africa as the target items. For Study 3 (the tasting scenario), we generated a word pool of all vegetables and fruits existing in the Word2vec semantic space, and selected strawberry, passionfruit, collard, and zucchini as the target items based on the multidimensional scaling solution.

Results

Qualitative Patterns

Desirability and Probability We first attempted to test the effect of desirability and probability (Phillips et al., 2019) on counterfactual generation. To do this, we computed the likelihood that each item gets listed as a counterfactual in the second session (i.e., retrieval probability) and correlated it with participants' baseline desirability and probability ratings from the first session. As illustrated in Figure 2a, an item that was considered more desirable was more likely to come to mind as a counterfactual thought. Similarly, retrieval probability increased with probability (Figure 2b). Overall, we observed a significant positive relationship between each item's retrieval probability and average desirability ratings (r = .677, p <.001 for Study 1; r = .652, p < .001 for Study 2; r = .492, p < .001 for Study 3) as well as probability ratings (r = .805, p < .001 for Study 1; r = .772, p < .001 for Study 2; and r = .538, p < .001 for Study 3).

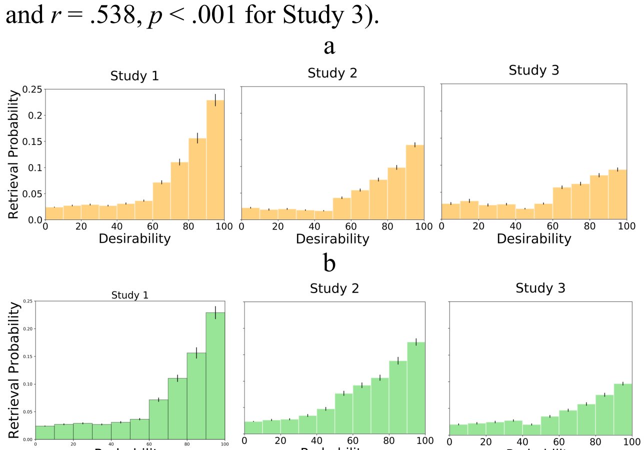


Figure 2. Plotting the probability that an item gets retrieved as a counterfactual as a function of (a) baseline desirability ratings, and (b) baseline probability ratings. Study 1 (left), Study 2 (middle), Study 3 (right). Each participant's ratings were divided into 10 percentile bins and then averaged across all participants. Errors bars display ± one standard error.

Similarity with the Target Outcome Our experimental paradigm asked participants to list the counterfactuals that came to their mind as they evaluated a target outcome. Semantic similarity with the prompt has been implicated in such processes, and we attempted to test for its effect using semantic representations in Google's Word2Vec space. Recall that we presented participants with target items from one of four distinct clusters in a two-dimensional decomposition of the Word2Vec semantic space. If similarity with the target cues retrieval, the counterfactuals listed by our participants should be more similar to an item when it is the target item, compared to when it is not. We found that this is indeed the case, as shown in Figure 3 which plots the semantic similarity to each target item across four target conditions.

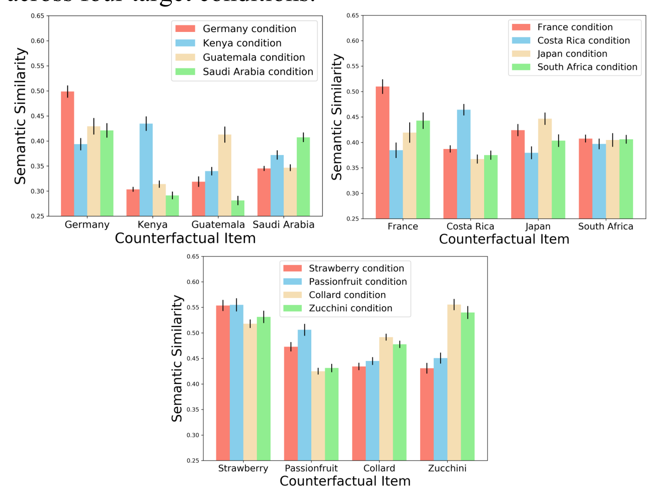


Figure 3. Semantic similarity of counterfactuals to a target item for each condition. Study 1 (top right), Study 2 (top left), Study 3 (bottom). Errors bars display ± one standard error.

We also conducted paired t-tests to compare (i) the semantic similarity between each counterfactual and their corresponding target item, and (ii) the average of semantic similarities between each counterfactual and three other items used as the target item in other conditions but not in the tested condition. For all twelve experimental conditions across the three studies, the results of the paired t-tests have p values of less than 0.001. This illustrates that, across participants, counterfactuals were higher in semantic similarity to an item when it was the target item relative to when it was not for all experiments and target conditions.

Semantic Clustering We tested for the effect of semantic similarity with previously retrieved counterfactuals on subsequent retrieval. A positive effect of similarity (leading to semantically clustered recalls) has been documented in prior work on memory search, though this effect has not been tested in counterfactual generation tasks. To perform this test, we measured conditional response probabilities (CRP) in retrieval using the method proposed by Howard and Kahana (2002), and once again we measured semantic similarity using the Word2Vec model. We calculated CRP with ten equally sized bins, with the first bin corresponding to the smallest similarity between possible pairs of items and the last bin corresponding to the largest similarity between all possible pairs of items (see Howard & Kahana, 2002, for details of this method).

For all three studies, the average CRP for the last similarity bins were substantially higher than the average CRPs for the remaining bins (Figure 4). This indicates that after one counterfactual comes to mind, participants were especially likely to think about other counterfactuals that are semantically related to that just-recalled counterfactual.

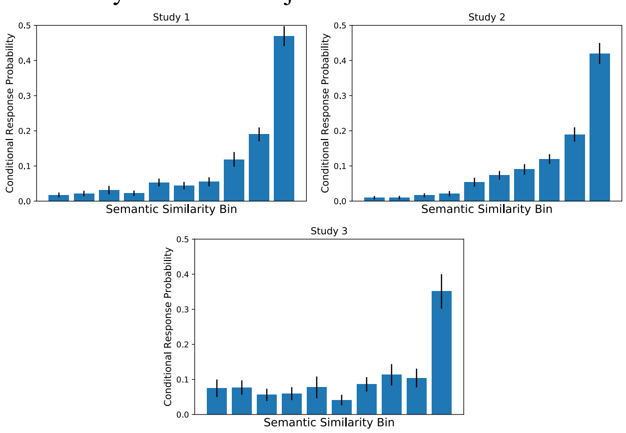


Figure 4: Plotting average conditional response probabilities for each bin number. Study 1 (top left), Study 2 (top right), Study 3 (bottom). Errors bars display ± one standard error.

Quantitative Fits

Markov Model The previous tests provided evidence for how various variables such as desirability of item, probability of item, semantic similarity with target, and semantic similarity with previously retrieved items influence counterfactual thinking. In addition, we expected that corpora frequency influences counterfactual retrieval because it has been shown in the memory literature that higher frequency words are more likely to be recalled (Aka, Phan, & Kahana, 2020). These variables have been previously implicated in counterfactual generation and memory search, and we found that they have a persistent effect in our task as well.

We also, however, attempted to model the effects of these variables more formally. In order to do so, we used a Markov memory model. In our model, we assumed that the "activation" of a counterfactual item at a given point in time was a linear function of that item's desirability, probability, corpora frequency, and its semantic similarity with the previously listed item, and its semantic similarity with the target item. In addition, we allowed for additional effects at the start of retrieval (thus the first retrieval could involve a disproportionately higher influence of desirability or similarity with the target). Probabilities of listing each counterfactual alternative were obtained by passing the item activations through a logit link function.

Model Fitting We fit our model using maximum likelihood estimation (with Nelder–Mead method) in Python and each model was fit 30 times to stabilize parameters. Next, we incrementally dropped each of the variables and conducted likelihood ratio tests between the full model and each of the constrained models. This method allowed us to evaluate whether the dropped variable resulted in an inferior fit of the model. Negative log likelihoods and *p*-values of the likelihood ratio tests are reported in Figure 5.

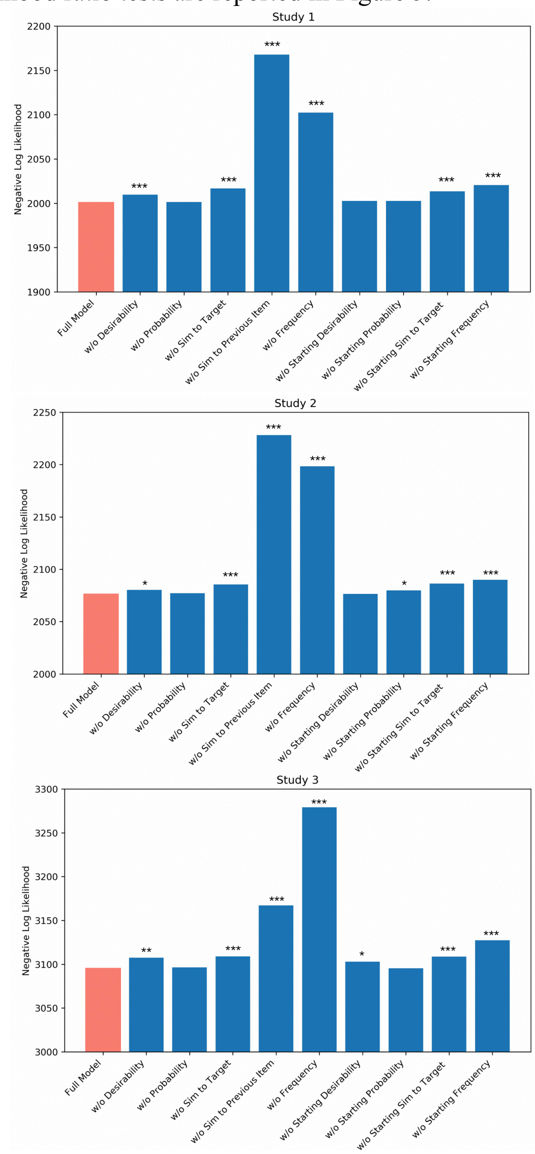


Figure 5: Plotting negative log likelihood for all models. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$.

In all three studies, we found that desirability had a positive contribution to counterfactual retrieval. More specifically, results from a likelihood ratio test between a model with and a model without desirability as a variable showed that adding desirability significantly improved the fit of the Markov memory model ($p < .001$ for Study 1; $p = .018$ for Study 2; $p = .004$ for Study 3). However, probability did not have a significant contribution to the model. There was a correlation of 0.935 ($p < .001$) between baseline desirability and probability ratings for Study 1, a correlation of 0.926 ($p < .001$) for Study 2, and a correlation of 0.949 ($p < .001$) for Study 3. Since these two variables are very highly correlated in our data, including them simultaneously in a model can lead to collinearity issues in the estimation of their effects. We also found that desirability and probability did not have a disproportionately larger or smaller effect at the start of retrieval.

Importantly, we demonstrated that memory-related mechanisms influence counterfactual retrieval. Previously we showed that counterfactuals that are more semantic similar to the target outcome are more likely to be retrieved. Here, we confirmed this effect by showing that adding semantic similarity to the target as a variable resulted in a much superior fit of the model, both at the start of the retrieval ($p < .001$ for all three studies) and for later retrieved counterfactuals ($p < .001$ for all three studies). Using the same approach, we also examined whether semantic similarity with previously retrieved counterfactuals influence subsequent ones. Previously, we found a significant correlation between conditional response probabilities (CRP) and semantic similarity in all of our studies. Similarly, model results also indicate that a retrieved counterfactual semantically cues the retrieval of the next counterfactual.

As discussed earlier, corpora frequency has been found to influence retrieval in memory recall tasks. Here, we showed that corpora frequency also had a positive contribution to counterfactual retrieval. More specifically, results from likelihood ratio tests showed that including corpora frequency as a variable in the Markov memory model significantly increases the model's fit, both at the start of counterfactual retrieval ($p < .001$ for all three studies) and for subsequent retrieval ($p < .001$ for all three studies). Hence, results from our models illustrated that memory processes play a significant role in counterfactual retrieval.

Predicting Post-Counterfactual Evaluations Our counterfactual generation task was followed by an evaluation task in which participants rated the desirability of the target item. Counterfactual alternatives have previously been implicated in such evaluations and we attempted to model the effects of counterfactuals using established theories from decision-making research (Loomes & Sugden, 1982; Mellers et al., 1997; Stewart et al., 2006). To do this, we relied on the insight that the list of counterfactuals serves as a reference class, and comparisons against this list increase or decrease subsequent evaluations of the target item. In the context of desirable counterfactuals, a target item is considered as less desirable, whereas in the context of undesirable counterfactuals, the item is considered as more desirable. To test for this effect in our data, we calculated the difference between the baseline desirability of the target and the average baseline desirability of the listed counterfactuals (baseline desirability measures were elicited in the first session, in the absence of a particular target item). We then regressed the rated desirability of the target in the second session on this difference measure. The effect of the difference variable was highly significant in all three studies: $F(1,51) = 12.66$, $p < .001$ in Study 1, highly $F(1,51) = 19.52$, $p < .001$ in Study 2, and $F(1,38) = 26.44$, $p < .001$ in Study 3. We are currently in the process of building more sophisticated decision models that take into account the order in which counterfactuals are retrieved, and are capable of implementing more complex comparisons, including those based on rank or loss aversion. We will present the results of this analysis elsewhere.

Discussion

People often wonder about how their experiences could have turned out differently by engaging in counterfactual thinking. These counterfactual thoughts influence our causal and moral judgments and guide our decisions. Despite their importance, the memory processes that underlie the generation of counterfactual thoughts have not yet been formally modeled. We attempted to build a model of counterfactual retrieval and quantitatively characterize the dynamics of counterfactual thinking. Specifically, we used a Markov memory model that specifies retrieval as a random walk over items in memory. Our model was able to formalize the effect of desirability, probability, similarity with target outcome, similarity with retrieved counterfactuals, and corpora frequency, and subsequently predict the sequences of counterfactuals listed by participants.

First, we found that subjective desirability and probability influence the likelihood that an item comes to mind as a counterfactual. This result replicates prior work in cognitive science. For example, Phillips et al. (2019) suggested that, across diverse tasks, the alternative possibilities that people consider by default are biased toward what is valuable and probable. We provided both qualitative and quantitative evidence to support this proposal, and by fitting parameters of a formal memory model, were able to characterize how desirability and probability guide retrieval.

Second, our findings revealed an underexplored relationship between counterfactual thinking and semantic memory. Although research in neuroscience has indicated an overlap between recalling experiences and imagining counterfactuals (e.g., Schacter et al., 2015), past work has not studied this relationship using computational memory models. Inspired by the free recall paradigm in memory research, we devised a novel task in which participants listed counterfactual thoughts in response to a particular target outcome or event. Replicating prior results in memory research (Aka & Bhatia, in press; Bhatia, 2019; Hills et al., 2012; Howard & Kahana, 2002), we observed a strong effect of semantic similarity with the target item, semantic similarity with the previously retrieved item, and corpora frequency. Thus, counterfactuals that were closely related to the event or outcome in consideration and to previously generated counterfactuals, as well as highly frequent, were more likely to come to mind.

Building on our model, future studies can investigate the effects of individual differences in counterfactual thinking. One promising domain for such an analysis involves aging. Researchers can fit separate models for elderly populations and parametrically specify the effects of aging on counterfactual retrieval. This will shed light on the specific set of memory mechanisms that are damaged with age. Prior work has found that older people are more likely to confuse counterfactual simulations for remembered events (Gerlach et al., 2014). By examining the parameters that increase the perceived similarity between counterfactuals and experiences in memory, researchers may be able to obtain new insights about age-related differences in cognition and behavior.

Of course, the promise of our model extends beyond age-related cognitive impairments to other types of disorders. For example, by correlating individual-level model parameters inferred from counterfactual retrieval data with neural data, researchers can better understand brain regions implicated in disruptions to counterfactual simulation. These studies can facilitate the development of interventions that improve real-world cognition in impaired populations.

To conclude, we present a novel approach to modeling retrieval dynamics in counterfactual thought. By building a Markov model of counterfactual retrieval, we combine insights from cognitive science and memory research to investigate the mental processes involved in counterfactual thinking. Our work opens up many potential research questions with substantial real-world applications, and we look forward to future work that uses our modeling framework to understand the determinants and consequences of counterfactual thought.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *NIPS* (pp. 3050-3058).

Aka, A. & Bhatia, S. (in press). What I like is what I remember: memory modulation and preferential choice. Journal of Experimental Psychology: General.

Aka, A., Phan, T. D., & Kahana, M. J. (2020). Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind?. *Cognition*, *194*, 104057.

Bhatia, S. (2019). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627-640.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135-157.

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

De Brigard, F., Henne, P., & Stanley, M. L. (2021). Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition*, *209*, 104574.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731-1744.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767-773.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval?. *Journal of Memory and Language*, *46*(1), 85-98.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, *92*(368), 805-824.

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, *44*(2), 421-428.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423-429.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*, 3111-3119.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026-1040.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, *116*(1), 129.

Quillien, T. (2020). When do we think that X caused Y?. *Cognition*, *205*, 104410.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In Advances in experimental social psychology. Academic Press.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of learning and memory*, *117*, 14-21.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223-247.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1-26.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of personality and social psychology*, *56*(2), 161.

Zultan, R. I., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429-440.