Causal inference shapes counterfactual plausibility

Tadeg Quillien[1] (tadeg.quillien@gmail.com), **Aba Szollosi**[2], **Neil Bramley**[2], **Christopher G. Lucas**[1]
[1]School of Informatics, University of Edinburgh, [2]Department of Psychology, University of Edinburgh

Abstract

When we reason about what could have been, some possibilities seem plausible, and others far-fetched. According to a recent theory, counterfactual possibilities are plausible if they can be generated by making local, probabilistic adjustments to the causes of what actually happened. We provide evidence that people think about counterfactuals in this way even when they have to infer the causes of what happened. We told participants about the diet of a fictional animal, and then asked them simple counterfactual questions. For example, given that the animal has eaten 1 berry today, how much food could it plausibly have eaten instead? When the amount of food eaten by the animal licensed an inference about a causally upstream variable, participants inferred the state of this variable and used it to guide their counterfactual plausibility judgments. More generally, the distribution over counterfactual values derived from participants' judgments was remarkably similar to the distribution predicted by the model.

Keywords: causality; counterfactuals; computational modeling

Introduction

It usually takes John between 20 and 30 minutes to drive to work. Today, John drives to work in 40 minutes. If it had taken John less than 40 minutes to go to work today, how much time would it have taken?

It is not clear whether there is a normatively correct answer to that question. Yet some answers (e.g. "33 minutes") seem intuitively better than others ("2 minutes"). Understanding these intuitions is important: many everyday judgments involve thinking about other ways things could have happened, and these judgments depend on what possible counterfactuals people consider to be plausible or relevant (Phillips et al., 2015; Bernhard et al., 2022; Lucas & Kemp, 2015; Quillien & German, 2021; Quillien & Lucas, 2023; Icard et al., 2017; Henne & O'Neill, 2022; Byrne, 2016; Kahneman & Miller, 1986; Lassiter, 2017a; Goldman, 1976).

What makes a counterfactual plausible? To preview our findings, we predict and find that people's judgments involve a compromise between the prior probability of a given value and its closeness to what actually happened, in ways that respect causal structure. In terms of our example above, people's judgments appear to be a compromise between A and B (See Figure 1a for illustration):

- A) "It would have taken John 25 minutes to drive to work." (typical value)
- B) "It would have taken John 39 minutes to drive to work." (nearby value)

We also ask why counterfactual plausibility judgments are pulled toward what actually happened. It is commonplace that counterfactual thoughts focus on 'nearby' possible worlds (e.g. Lewis, 1973; De Brigard et al., 2021), but what counts as a nearby world to the human mind?

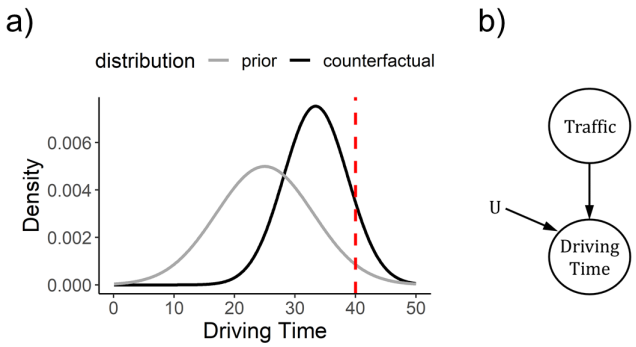


Figure 1: **a)** Our framework predicts that plausible counterfactual values (black line) are influenced by the variable's prior distribution (grey line) and its actual-world value (dashed red bar). **b)** Causal model for our driving time example. *U* represents unknown factors besides traffic that also affect driving time.

We consider two possible accounts. On the first account, counterfactual plausibility is driven by a low-level, non-causal sense of similarity. People think that 39 minutes is a plausible value simply because 39 is numerically close to 40 minutes (John's actual-world driving time).

On the second account, actual-world information affects plausibility via causal inference. When people learn that John took longer than usual to drive to work, they infer the possible causes of that event (e.g. there probably was a lot of traffic). When they simulate other possible ways that things could have happened today, they start by simulating the causally upstream variable (i.e. whether there is traffic), and then they simulate driving time in function of how much traffic there is. People tend to only make small modifications to how much traffic there was, and driving time is pulled toward its actual-world value as a result.

In this particular example and in most cases, the two accounts make similar predictions. Here we find evidence for the causal inference hypothesis, in a setting where the hypothesis predicts that events are sometimes pulled *away* from their actual-world values.

The Extended Structural Model

Our predictions are guided by the idea that counterfactual reasoning involves computations over causal models. In particular, we test some assumptions of the Extended Structural Model of counterfactual reasoning (ESM; Lucas & Kemp, 2015), a psychological model that builds on Pearl's structural analysis of counterfactuals (Pearl, 2000).

A causal model represents a given aspect of the world in terms of variables linked by causal relationships. For our driving time scenario, Driving Time is causally influenced by Traffic, and a variable U represents all other (unknown) factors that can influence driving time (see Figure 1b). The causal model also encodes information about how causes generate their effects, and about the base rate of variables whose causes are not explicitly represented in the model (not shown in Figure 1).

The ESM can be understood in terms of a process that *simulates possibilities* from a causal model.[1] Under this framwork, answering a counterfactual conditional question (e.g. "If John had driven to work in less than 40 minutes, how much time would he have taken?") consists in simulating many different counterfactual possibilities, and discarding those possibilities that contradict the premise (for example, possibilities where John takes 40 minutes or more to drive to work)[2].

To simulate a possibility from a causal model, we first sample the value of the causally upstream variables (e.g. Traffic), then we determine the value of the causally downstream variables (e.g. Driving Time) as a function of the values of their causes. A crucial assumption of the ESM is that people simulate the counterfactual value of upstream variables by making probabilistic, local adjustments to their actual-world value (Lucas & Kemp, 2015). For instance, if people think there was a lot of traffic in the actual world, they will tend to imagine counterfactual possibilities that also have a lot of traffic, although the amount of traffic varies stochastically from one simulation to the next.

Note that if people cannot observe the actual-world value of a variable, they can infer this value (or a posterior probability over this value) from the state of the variables that they do observe. For instance, people might infer that there was probably a lot of traffic from the observation that John took a lot of time to drive to work.

Combining these two assumptions (local probabilistic adjustments to what actually happened, and inferences about the actual-world value of unobserved variables), we expect that causal inference might play a role in anchoring counterfactual reasoning to the actual world. When observing an event, people infer the possible causes of this event. Then, they simulate counterfactual possibilities by making local adjustments to the (inferred) state of these causes.

The current experiment

If people simulate possibilities by making local adjustments to the actual world, their judgments should be biased by what actually happened. But such a bias might also have simpler explanations. People might bias their answers toward actual-world values because of basic anchoring effects (Tversky & Kahneman, 1974), or because they are guided by a low-level, non-causal sense of similarity (Lewis, 1973). In contrast, we argue that judgments are anchored to the actual world in part because people infer the causes of what they observed. We therefore designed an experiment for which our causal inference account makes different predictions than these simple accounts.

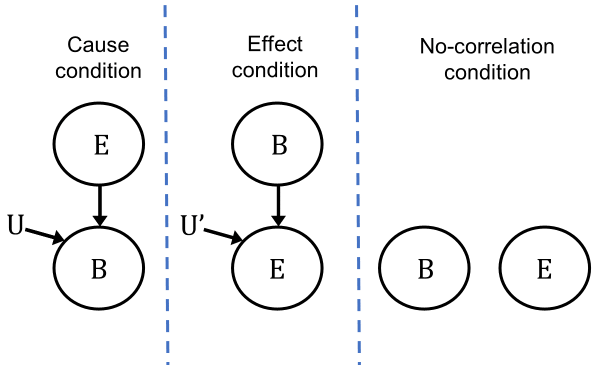


Figure 2: Graphs of the causal models used in each condition. E: presence of the Enzyme; B: number of Berries eaten. U variables (not explicitly mentioned to participants) account for stochasticity in the causal relationships.

Methods

Participants played the role of scientists studying the food habits of an animal species (the yorgis) on an alien planet. The yorgis eat berries, and on certain days they have an enzyme called XRD in their bloodstream. Participants were taught a probabilistic causal model of the relationship between XRD and food consumption. We told them that having XRD in its blood makes a yorgi food consumption *more extreme* (i.e. it makes the animal eat either very few or very many berries), and we also taught them the probability distribution over the possible number of berries that a yorgi can eat on a given day, depending on whether it has XRD in its blood or not.

After teaching the probabilistic causal model to participants, we asked them a series of counterfactual questions. For example, given that today the animal ate 1 berry, how many berries would it have eaten if it had eaten more than 1 berry?

In our main experimental condition (the Cause condition), we told participants that XRD has a causal influence on food consumption. We also included two control conditions where we reverse or remove the causal relationship between the enzyme and food consumption (see Figure 2). In the Effect condition, food consumption has a causal influence on the enzyme's presence (rather than the other way around), but otherwise the covariation between food consumption and enzyme is exactly the same as in the Cause condition. In the No Correlation condition, there is no relationship between food consumption and the enzyme, but the marginal probability distribution of both variables is the same as in the other two conditions. Condition was manipulated between-subject.

[1] Note that we are not making a process-level claim; i.e. we are agnostic about whether people actually perform simulations. Our (functional-level) account is simply easier to formulate in terms of a simulation-based procedure.

[2] Here we model counterfactual premises as observations, although the broader theory can also accommodate interventions, see Lucas & Kemp (2015).

Predicted differences between the conditions

Suppose that in the actual world, today the animal ate only 1 berry. In the Cause condition, plausibility judgments should be somewhat U-shaped. That is, participants should favor counterfactuals where the animal eats either low or high amounts of food. This is because the animal probably has the enzyme in its blood today (this explains why it ate little food). If people simulate counterfactuals that look like what they think the actual world is like, they should tend to simulate counterfactuals where the animal also has the enzyme. In these counterfactual worlds, the animal is also likely to eat either very little or a lot of food. On the other hand, non-causal accounts of counterfactual similarity predict that people should be simply biased toward low values, because they are most similar to what actually happened.

In the control conditions, our account assumes that people simulate counterfactual values for the Berries variable before they simulate counterfactual values for the enzyme variable (or independently of it). Therefore, counterfactual plausibility judgments should not be affected by inferences about the enzyme. Judgments should simply be biased towards the variable's actual-world value.

Procedure

After completing a consent form, participants were given explicit instructions about the causal structure. For example, in the Cause condition, they were told that XRD makes an animal's food consumption more extreme: when the animal has XRD in its blood, it eats either a very low or very high amount, and when it does not have XRD in its blood, it eats a moderate amount. Then they completed a comprehension question (asking about the causal direction between XRD and food consumption—participants failing this question were excluded from analysis). In order to teach participants the joint probability distribution induced by the causal model, we then asked them to observe, for each of 40 days, the number of berries that the animal had eaten, and whether it had the enzyme in its blood or not.

During this distribution learning phase, participants saw the data for each day on a separate page. Each page first displayed information about whether XRD was present in the animal on that day (in words, as well as with a picture of the enzyme above the animal's picture—on days where XRD was absent there was empty space above the animal's picture). Then after 500 ms the screen also displayed information about the number of berries eaten by the animal (in words, as well as with a picture of the berries). After another 500 ms a button appeared at the bottom of the screen that allowed participants to proceed to the next page. In the Effect condition, the order of appearance was reversed, and the berries appeared onscreen before the enzyme. (We presented the effect variable after the cause variable in order to emphasize the direction of the causal relationship. We chose to present samples for 40 days on the basis of previous research on distribution learning; Yeung & Whalen (2015)).

Figure 3a (purple bars) shows the frequency with which each amount of berries was presented, in the absence and in the presence of XRD, in each condition. The enzyme was present in half the trials. The order of presentation of the stimuli was randomized for each participant.

After the distribution learning phase, participants answered four counterfactual conditional questions. The questions were phrased as follows:

"On another day you see the yorgi eat [number] berries. You have not tested whether it has enzyme XRD in its blood or not. If the yorgi had eaten [less / more] than [number] berries on that day, how many berries do you think it would have eaten? Please use the slider next to each number to indicate how much you agree that the yorgi would have eaten that number of berries."

The numbers used for the actual-world amount of berries were 1,5,6,10. Participants were asked to imagine the animal eating more berries than it actually did when the number was 1 or 5, and to imagine less berries when the number was 6 or 10. Participants were implicitly asked to enter a distribution over possible counterfactual values: they answered the question using n horizontal sliders, one for each possible number of berries consistent with the counterfactual premise. For example, to the question about "more than 5 berries" participants answered using five sliders labeled "x berries" with x ranging from 6 to 10. Each slider was initially anchored in the middle, and sliders were otherwise unlabeled. The order of presentation of the four conditionals was randomized.

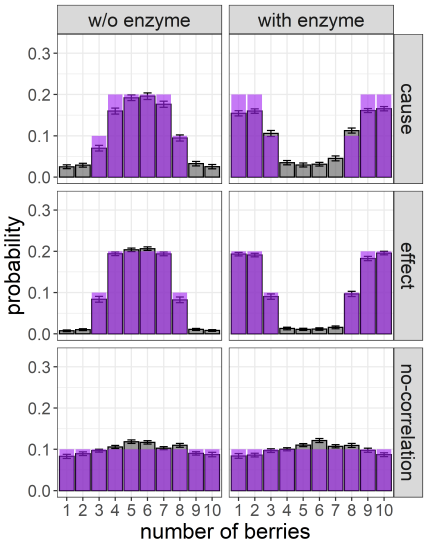
To check that participants had been able to learn the probability distributions during the learning phase, we then asked them to reproduce the probability distribution over the number of berries eaten conditional on the presence of the enzyme, and then conditional on its absence. These questions used a similar response format as the counterfactual conditionals: on each page, participants set the values of 10 sliders labeled from "1 berry" to "10 berries".

Finally, participants completed a short demographic questionnaire, were thanked for their participation and redirected to Prolific for compensation.

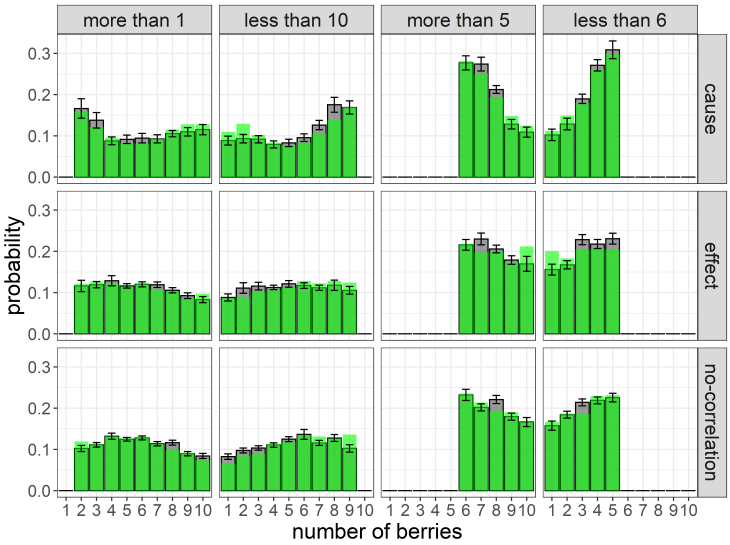
Before analysis, we standardize participants' ratings so that the value of all slider ratings for a given distribution sum to 1 (this transformation was preregistered). Data and code for modeling and analysis are available at https://osf.io/fzgmn/?view_only=6106ba12602d4f11bb721ccda14e949c. Interested readers can try out the experiment at http://eco.ppls.ed.ac.uk/~tquillie/countenz/.

Participants

We recruited 294 US residents (142 female, 5 other) from Prolific. Participants were compensated £0.90. Following our preregistered exclusion criteria (see https://osf.io/cs76p/?view_only=dc55598ae32f492c843e3a2219bd2822), we excluded 26 participants who failed the comprehension check, and 61 participants whose performance in the distribution-



(a) **Results for the conditional distribution learning questions**. Purple bars show the ground truth distribution (i.e. the frequency of each value in the training data) and grey bars show participants' probability judgments. Left (right) panels display probabilities in the absence (presence) of the enzyme. Error bars represent SEM.



(b) **Results for the main task.** Grey bars show mean human judgments of counterfactual plausibility, and green bars show model predictions. Each vertical panel represents a different counterfactual conditional, and each horizontal panel a different condition. The value in the vertical panel names also indicate the variable's actual-world value, e.g. for the "more than 1" question the animal eats 1 berry in the actual world. Error bars represent standard error of the mean.

Figure 3: Results.

learning task was below a threshold[3], yielding a final sample of 207 participants.

Results

Participants learned the conditional distributions. See Figure 3a. Average human judgments for the conditional distributions were highly correlated with the ground truth distribution, $r(58) = .97$, $p < .001$. This correlation is artificially inflated by the exclusion of participants whose learning was below-chance, but it confirms that participants in our final sample learned the causal model (a necessary assumption for our later analyses to be meaningful).

Plausibility judgments are anchored to the variable's actual-world value. Grey bars in Figure 3b display participants' mean answers to the counterfactual conditionals. Counterfactual values that are close to the actual-world value tend to be judged as more plausible. We can test this more formally by looking at the conditionals "more than 1" and "less than 10", and focus on answers for counterfactual values 2 to 9, as participants made judgments about these values for both conditionals. If judgments are anchored in the variable's actual-world value, then participants should think that (e.g.) '2 berries' is a more plausible counterfactual value when the animal has eaten 1 berry in the actual world, compared to when the animal has eaten 10 berries in the actual world.

Specifically, we looked at the interaction between the actual-world and the counterfactual value, in linear mixed models with actual-world amount and counterfactual amount as predictors, random slopes (for both predictors) and random intercepts, and participants as random effects. This interaction is significant in the Cause condition ($p < .001$) and the No-Correlation condition ($p = .01$), but not in the Effect condition ($p = .15$), although it is in the predicted direction.

Why are judgments anchored in the actual world? According to our hypothesis, counterfactual reasoning potentially involves inferences about variables that are causally upstream of the target variable. As outlined above, this predicts that people's judgments should be different depending on the causal structure they have been taught. We test this prediction next.

Participants generate different distributions depending on causal structure. In the Cause condition, distributions tended to be skewed toward extreme values when the evidence suggests the presence of the enzyme (the animal ate 1 or 10 berries), and toward moderate values when the evidence suggests the absence of the enzyme (the animal ate 5 or 6 berries); see Figure 3b. There was no such pattern (or a weaker one) in the Effect and No-Correlation conditions. We also find that – as predicted – the distributions in the two control conditions (Effect and No-correlation) were very similar.

Bootstrapping tests suggest that the distributions generated

[3]We excluded these participants because our experiment is not focused on people's ability to learn distributions, and we need to assume that participants have learned the correct causal model for our analyses to be meaningful. We assessed participants' performance using the Kullback-Leibler divergence between a participant's reported distribution and the ground truth distribution, and set the performance threshold to the KL divergence between a random (i.e. uniform) responder and the ground truth distribution in the Cause and the Effect condition.

by participants in the Cause condition were significantly different than in the Effect (all ps $< .003$) and the No-Correlation condition (all ps $< .001$)[4]. By contrast, there was no evidence for a difference between the Effect and No-Correlation condition, all ps $> .46$.

Computational modeling

Our results provide support for the qualitative predictions we derived from the ESM (Lucas & Kemp, 2015). Here we test whether the model can also explain people's judgments at a quantitative level.

The ESM borrows the following two assumptions from Pearl's structural model analysis (Pearl, 2000). First, people's representations of the world can be modeled with Structural Causal Models (see Pearl, 2000, for an introduction to SCMs). That is, people represent the world in terms of variables linked by causal relations. The causal relations are deterministic: even though the observed relationships between the Enzyme and the Berries variable is stochastic, the apparent stochasticity is assumed to come from unobserved background factors U.

Second, people infer the actual-world value of unobserved variables from the actual-world value of observed variables. For instance, when observing the actual-world value b_a of the Berries variable (the number of berries eaten), people infer the actual-world values of Enzyme and of U. People make these inferences on the basis of Bayes' rule, by inverting the generative model that we describe in the next section.

Generative model

The presence of the enzyme is represented by a binary variable E, with $Pr(E=1)=\frac{1}{2}$. We make the assumption (not explicitly instructed to participants) that the number of berries is generated by a function $F(u,e)$ which returns the u-th percentile of the distribution $Pr(B|E=e)$[5].

This assumption implies that U is uniformly distributed in $[0,1]$. To model counterfactual re-sampling of U it will be useful to think of U as representing the combined effect of many independently operating background processes. We will assume that these background processes combine additively to produce an effect W which is normally distributed with mean np and variance $np(1-p)$, where n is a large number representing the number of independent background processes that can have an influence on U (we set $n=1000$ without loss of generality)[6] and p represents the probability that a given process operates, which we arbitrarily set to $p=1/2$. We then obtain U by transforming W to make it uniform, via a function Q that maps every percentile of W to the corresponding percentile of U.

Counterfactual re-sampling

In the Cause condition, the model simulates counterfactuals according to Algorithm 1. In the control conditions, the model works in the same way, except that computation for the value of B is equivalent to the computation for the value of U in the Cause condition; the value of E is not relevant for computing B and is not simulated.

For each simulation, the model first samples each exogenous variable (i.e. parentless node in the graph) from a mix of its prior distribution and its posterior belief about the variable's actual value.

Counterfactual re-sampling for E works in the standard way specified by the ESM (see Lucas & Kemp, 2015). We sample E from the posterior $Pr(E_a|B=b_a)$ with probability s_E, and from the prior $Pr(E)$ with probability $1-s_E$, where s_E is a free parameter.

Because U is a continuous variable, we re-sample it using a different process—which can be seen as a natural extension of the ESM to continuous variables. To sample the value of U, the model first simulates a latent variable W, which represents the number of independent background processes contributing to U that are 'active'. This value is sampled from a distribution that is biased by the inferred actual-world value of U. Formally, it is sampled from a normal distribution with mean $\mu_w = w_a p^+ + \neg w_a p^-$, and standard deviation $\sigma_w = w_a p^+(1-p^+) + \neg w_a p^-(1-p^-)$, where $p^+ = s_U + (1-s_U)p$, $p^- = (1-s_U)p$, and $\neg w_a = n - w_a$, with w_a the inferred actual-world value of W, and s_U a free parameter. This particular distribution can be shown to reflect the result of a process where we re-sample the outcome of every independent background process from a mix of its prior distribution and its (inferred) actual-world value. The sampled counterfactual value of U can then be computed as $u = Q(w)$.

Finally, the value of B is set according to the functional equation $b = F(e,u)$. For each condition and each counterfactual question, we generate model predictions by collecting the value of B in 10^5 samples, and discarding the samples that do not respect the counterfactual premise (e.g. for the 'more than 5 berries' conditional, samples where $b_i \leq 5$).

Model fitting

We fit the values of s_U and s_E to the human data by maximizing the correlation between model predictions and average human judgments. We find $s_U = .09$; $s_E = .64$.

[4] We compute a bootstrap test between two conditions by randomly re-sampling (with replacement) the participants assigned to the two conditions n times (where n is the total number of participants assigned to either condition), and arbitrarily dividing these re-sampled participants into two new subgroups. For each subgroup we compute the mean judgments, and then we compute the Hellinger distance between the mean judgments in each subgroup. Repeating this simulation process a large number of times, we can approximate the expected distribution of the Hellinger distance between two samples assuming they come from the same distribution. We compute a p-value as the number of such simulations where the Hellinger distance is larger than the empirically observed distance. Each p-value is based on 10000 simulations.

[5] Because we measured participants' conditional probability distributions for $Pr(B|E=1)$ and $Pr(B|E=0)$—see Figure 3a—we use these empirically derived distributions (averaged at the group level) in our modeling (see pre-registration).

[6] Repeating the analysis across a wide range of different values (20 to 10000) shows negligible sensitivity to the value of n.

Algorithm 1 Counterfactual simulation.

Infer $Pr(U_a|b_a)$ and $Pr(E_a|b_a)$ ▷ Infer actual-world state
for $i \leftarrow 1$ **to** m **do** ▷ Sample m counterfactuals
  Sample e_i from
    $Pr(E_a|b_a)$ with prob s_E
    $Pr(E)$ otherwise
  Sample u_a from $Pr(U_a|b_a)$
  Set $w_a = Q^{-1}(u_a)$
  Sample $w_i \sim \mathcal{N}(\mu_w, \sigma_w)$ ▷ See main text
  Set $u_i = Q(w_i)$
  Set $b_i = F(e_i, u_i)$
  if b_i inconsistent with counterfactual premise **then**
    Discard sample
  end if
end for
return simulated counterfactuals b_i, ..., b_m.

Results

Figure 3b displays best-fitting model predictions (in green) alongside participants mean judgments (grey). The model has a good fit to participants' average judgments, r(82) = .96, p < .001.

We also compare the model with lesioned versions that constrain the re-sampling to be based entirely on either the prior distribution or the actual-world posterior. That is, these models constrain s_U and s_E to be either 0 or 1. There are 4 such versions (one for each combination of s_U and s_E in $\{0,1\}^2$). We also consider a model that does not engage in causal inference (which we implement by constraining the full model to make the same predictions in the Cause condition as it does in the control conditions), and a random baseline which divides probability mass equally between all counterfactual values.

Because most of these alternative models have fewer free parameters than the ESM, we compare the models using the Bayesian Information Criterion[7], which penalizes models with more free parameters, see Figure 4. Bayes Factors derived from these BICs indicate that the ESM has a significantly better fit to the data than all alternatives, all BFs $> 10^4$.

General Discussion

Our findings add to the evidence that humans use a causal representation of the world to reason about counterfactuals (Pearl, 2000; Rips, 2010; Lassiter, 2017b; Vandenburgh, 2022).

[7]To compute the BICs, we compute the log-likelihood of the data at the aggregate level, using average distribution as the unit of analysis. For each distribution (e.g. the distribution of counterfactual plausibility for "less than 1" in the "cause" condition), we computed the average distribution by computing the average across participants for each rating. Then we calculated the log-likelihood of this distribution under a Dirichlet with vector of parameters $k * [\alpha_1, ..., \alpha_n]$, where n is the number of relevant ratings (i.e. the number of sliders participants had to click on), and k is a free parameter determining the stochasticity of the distribution (lower values of k correspond to more noisy responses). We compute the total log-likelihood as the sum of all log-likelihood across distributions.

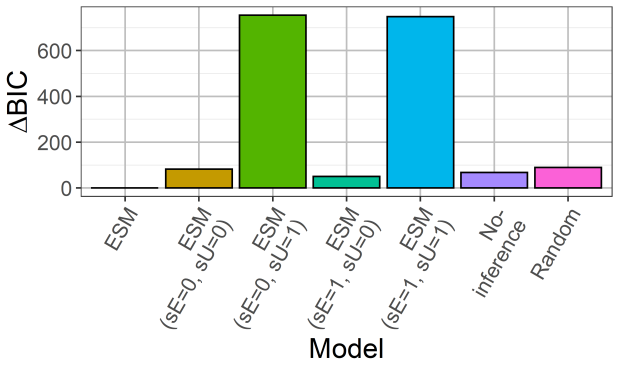


Figure 4: Difference in BIC relative to the best-fitting model.

On the other hand, Pearl's classic model of counterfactual reasoning cannot account for our results. When there are many possible ways to make a counterfactual premise true, Pearl's account does not specify which one we should choose (see Ciardelli et al., 2018; Lassiter, 2017a). There are for instance many possible ways to make true the premise "if the animal had eaten less than 10 berries": make the animal eat 1 berries, 2 berries, 3 berries, etc. The ESM solves this problem by specifying a probabilistic simulation process that generates a distribution of possible values for the target variable.

We also find that people spontaneously engage in causal inference when answering counterfactual questions. It is worth clarifying how this finding differs from other cases of "backtracking" (e.g. Rips & Edwards, 2013; Gerstenberg et al., 2013). In the paradigmatic case of counterfactual backtracking, people judge that if the soldier had not shot, the captain would not have given the order to shoot. That is, people make inferences about a causally upstream variable (the captain) from the information in the *counterfactual premise* ("if the soldier had not shot"). Here we show that counterfactual reasoning is also shaped by causal inferences made on the basis of the *actual-world value* of a variable.

The current data complement existing evidence for the ESM as an account of human counterfactual reasoning (Lucas & Kemp, 2015; Quillien & Lucas, 2023). Our main aim was to test *qualitative* predictions of the theory, but we also find a surprisingly high *quantitative* fit to people's judgments. There are however two limitations to our modeling exercise. First, our computational model is not a process-level account. It is unlikely that the judgments of any individual participant give us a full readout of the plausibility distribution induced by his or her causal model. Instead, the ratings made by each participant probably reflect an approximation, perhaps taken by extrapolating from a few samples. Indeed, inspection of the individual-level data (available on the project's OSF page) reveals interesting variability. Second, we had to make some assumptions about how people represented the causal model. In future research, we hope to give participants a fuller description of the system they make judgments about.

Acknowledgments

This research was supported by an EPSRC Grant (EP/T033967/1) to N.B. and C.L.

References

Bernhard, R. M., LeBaron, H., & Phillips, J. (2022). It's not what you did, it's what you could have done. *Cognition*, *228*, 105222.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*(1), 135–157.

Ciardelli, I., Zhang, L., & Champollion, L. (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, *41*(6), 577–621.

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., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy*, 771-791.

Henne, P., & O'Neill, K. (2022). Double prevention, causal judgments, and counterfactuals. *Cognitive Science*, *46*(5), e13127.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

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

Lassiter, D. (2017a). Complex antecedents and probabilities in causal counterfactuals. In *21st amsterdam colloquium* (pp. 45–54).

Lassiter, D. (2017b). Probabilistic language in indicative and counterfactual conditionals. In *Semantics and linguistic theory* (Vol. 27, pp. 525–546).

Lewis, D. (1973). *Counterfactuals*. John Wiley & Sons.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological review*, *122*(4), 700.

Pearl, J. (2000). *Causality*. Cambridge university press.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive science*, *34*(2), 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vandenburgh, J. (2022). Backtracking through interventions: An exogenous intervention model for counterfactual semantics. *Mind & Language*.

Yeung, S., & Whalen, A. (2015). Learning of bimodally distributed quantities. In *Proceedings of the cognitive science society.*