Assessing Alternative Models of Counterfactual Reasoning

Bob Rehder (bob.rehder@nyu.edu)[1], **Selena Zheng (yz9015@nyu.edu)**[1], **Angela Cao (ayc8004@nyu.edu)**[1]
Neil R. Bramley (neil.bramley@ed.ac.uk)[2], **Tadeg Quillien (tadeg.quillien@ed.ac.uk)**[2]
[1]Department of Psychology, New York University, New York, United States
[2]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom

Abstract

When reasoning about events that occur in the world, two key types of judgments are *counterfactual inferences*—judging how the event might have turned out differently under different conditions—and *causal selection*—identifying which of multiple potential causes is responsible for an effect of interest. We conducted a novel experimental test that asked subjects to make *both* types of judgments about a realistic situation involving probabilistic causal relations and both proximal and distal causes. We found that a new model— the *Exogenous Sampler* (EXS) performed as well as the leading models of these judgment types but also specified the cognitive processes via which such judgments are computed. Yet, that all models failed to predict the full range of subjects' judgments points to the need of a new generation of models accounting for how people reason about events that arise from complex probabilistic causal structures.

Keywords: Causal reasoning, functional causal models, counterfactual reasoning, causal selection judgments

Introduction

In daily life, people often reason about states of the world (e.g., missing a flight), either by considering counterfactual alternatives (e.g., "What if I had woken up on time?") or by selecting a cause for the effect (e.g., "I missed the flight because I woke up late."). Previous research suggests that, when making such inferences, people use a probability distribution that reflects a compromise between the actual world and their prior expectations about events. Models based on this assumption have provided impressively accurate accounts of human judgments (Lucas & Kemp, 2015; Quillien & Lucas, 2024). However, despite their success at the computational level, these models do not specify how such calculations are carried out by human reasoners. To provide a process-level explanation, we introduce two new models, the *Counterfactual Mutation Sampler* (MS/CF) and the *Exogenous Sampler* (EXS), where the compromise between actual-world and prior is explained by finite sampling under cognitive resource constraints. We evaluate these models against leading accounts of counterfactual and causal selection judgments using a novel experiment in which subjects made both inferences across various states of affairs within a complex causal structure representative of everyday reasoning.

Models of Counterfactual Reasoning

The Extended Structural Model Leading models of counterfactual and causal selection judgments are based on the *Extended Structural Model* (ESM) proposed by Lucas & Kemp (2015). The ESM is itself an extension of the *Structural Model* proposed by Pearl (2000, 2013). A key innovation of these models is their use of *functional causal models*, models that stipulate that causal relations are deterministic but that allow for the introduction of *exogenous variables* that carry probabilistic information. Consider Fig. 1A which shows a functional causal model depicting the example first presented in Lucas & Kemp (2015), namely, the causal model where cooking bacon (B) causes the activation of a smoke detector (S) that in turn annoys the neighbors (N). The exogenous variables of this functional model are depicted as black squares and diamonds. The diamonds U_B, U_S, and U_N are independent causes of B, S, and N, respectively. For example, U_S represents causes of the smoke detector going off other than cooking bacon (e.g., a roommate smoking in their bedroom). The squares U_{BS} and U_{SN} represent enabling conditions for the causal relations $B \rightarrow S$ and $S \rightarrow N$, respectively. For example, U_{BS} represents the preconditions that must hold for cooking bacon to set off the smoke detector (e.g., the absence of a strong draft in the kitchen that blows the smoke away from the detector). In a functional causal model the endogenous variables are a deterministic function of the exogenous variables. For example, the smoke detector will activate if bacon is cooked (*and* the preconditions represented by U_{BS} are met) *or* one of alternative causes of the smoke detector activation represented by U_S is present, a situation represented by the structural equation $S = (B \wedge U_{BS}) \vee U_S$. Analogously, N is a deterministic function of S, U_{SN}, and U_N, namely, $N = (S \wedge U_{SN}) \vee U_N$. As the root cause of the causal chain, B is a simple function of U_B, $B = U_B$.

Given the functional causal model in Fig. 1A, counterfactual reasoning proceeds in two steps. First, an actual outcome (i.e., values for the endogenous variables B, S, and N) is observed, which in turn allows inferences to be drawn regarding the states of the exogenous variables. The red text Fig. 1A presents a quantitative example: The independent cause of B (the likelihood we cook bacon today), U_B is present with probability .50 and the independent alternative causes of S and N, U_S and U_N, are present with probability .20. The enablers for $B \rightarrow S$ and $S \rightarrow N$ are each present with probability .67. After observing that B, S, N each occurred, denoted $b^1 s^1 n^1$, we can infer the posterior probability of each exogenous variable: $P(u_B^1 | b^1 s^1 n^1) = 1$, $P(u_S^1 | b^1 s^1 n^1) = P(u_N^1 | b^1 s^1 n^1) = .27$ and $P(u_{BS}^1 | b^1 s^1 n^1) = P(u_{SN}^1 | b^1 s^1 n^1) =$

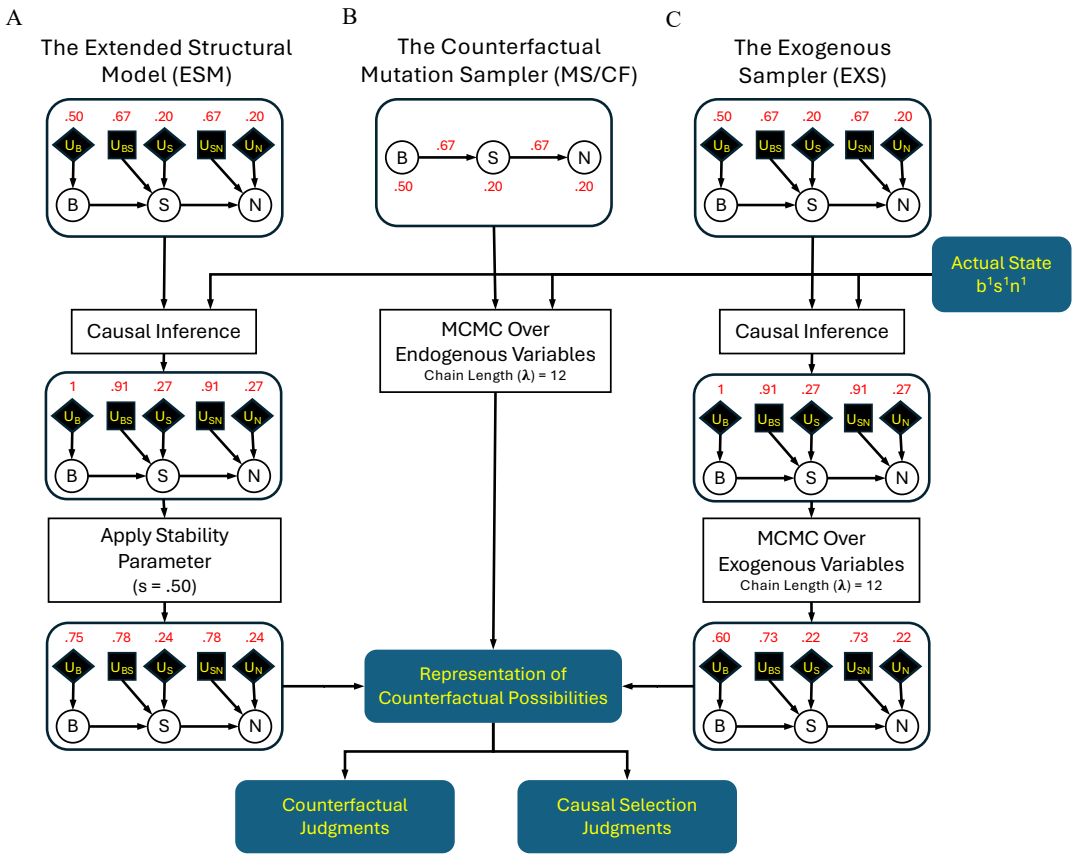

Figure 1: Three models of counterfactual reasoning. (A) The extended structural model (ESM). (B) The counterfactual mutation sampler, MS/CF. (C) The exogenous sampler, EXS. The black squares and diamonds represent exogenous variables where the squares represent enabling conditions and the diamonds represent alternative causes. Red text denotes a specific parameterization. Predictions are derived from an actual state in which B, S, and N are each observed present, denoted $b^1s^1n^1$.

.91. These posterior probabilities define a new functional causal model from which counterfactual inferences can be drawn (e.g., one can compute the probability that the neighbors would have been annoyed *if the smoke detector hadn't gone off*, $P_{b^1s^1n^1}(n^1|s^0)$).

The extension to the Structural Model proposed by Lucas & Kemp (2015) is that the exogenous variables' posterior probabilities should be adjusted by a *stability parameter* s whereby they are a weighted average of their prior and posterior values. For example, given stability $s = .5$ Fig. 1A shows that the effective posterior probability of U_{BS} should be, $.67 \times .50 + .91 \times .50 = .78$. The functional causal model defined by these stability-adjusted probabilities is, according to the ESM, the representation of counterfactual possibilities upon which counterfactual judgments are computed.

Another type of judgment believed to be related to counterfactual reasoning is *causal selection*, that is, identifying how responsible causes are for an outcome (Gerstenberg & Icard 2020; Henne & O'Neill 2022; Icard et al. 2017). For example, according to Quillien (2020)'s *Counterfactual Effect Size Model* (CESM), causal selection judgments are carried out by sampling over a distribution of counterfactual possibilities and then computing the resulting correlation between the cause and effect of interest. Quillien & Lucas (2024) extended the CESM by assuming that that sampling is carried out with the same kind of stability-adjusted functional causal model proposed by Lucas & Kemp (2015). To continue our running example, measures of how responsible bacon being cooked and the smoke detector going off, respectively, are for the neighbors being annoyed are, according to the CESM, computed from the stability-adjusted functional causal model shown in Fig. 1A. Hereafter we assume that the ESM (and all models tested in this article) include the CESM as a submodel that computes causal selection judgments.

Table 1 presents example inferences derived from the ESM for the causal model in Fig. 1A. The table first presents normal (i.e. non-counterfactual) inferences. For example, under the parameterization stipulated in Fig. 1A, if the smoke detector doesn't go off then the probability that the neighbors are annoyed, denoted $P(n^1|s^0)$, is .200, and the probability that bacon was cooked, $P(b^1|s^0)$, is .250. However, the counterfactual version of these inferences in which B, S, and N are each observed to be present—denoted $P_{b^1s^1n^1}(n^1|s^0)$ and $P_{b^1s^1n^1}(b^1|s^0)$—are .236 and .389.[1] This is so because

[1]We restrict discussion to *observational* as opposed to *interventional* counterfactuals, e.g., that the computation of $P_{b^1s^1n^1}(b^1|s^0)$ involves *backtracking*, the possibility that S was (counterfactually) absent because its cause B was too. The models herein support interventional counterfactuals, but, because our experiment only asked for observational counterfactuals, we do not discuss them further.

Table 1: Example causal, counterfactual, and causal selection judgments for four models.

Judgment	ESM	MS/CF	EXS	Baseline
Causal				
$P(n^1\|s^0)$	.200	.174	.200	.200
$P(b^1\|s^0)$	.250	.217	.250	.250
$P(b^1\|s^0n^1)$	.250	.255	.250	.250
Counterfactual				
$P_{b^1s^1n^1}(n^1\|s^0)$	.236	.297	.223	.200
$P_{b^1s^1n^1}(b^1\|s^0)$	.389	.360	.289	.250
$P_{b^1s^1n^1}(b^1\|s^0n^1)$	.389	.336	.289	.250
Selection				
$S_{b^1s^1n^1}(B,N)$	.329	.253	.318	.286
$S_{b^1s^1n^1}(S,N)$	.585	.499	.567	.535

Note: Predictions are based on the parameters shown in Fig. 1. The counterfactual and causal selection judgments are derived assuming the observed actual state $b^1s^1n^1$.

the stability-adjusted posterior probabilities of the exogenous variables computed from the actual state $b^1s^1n^1$ entail that B and N are more likely to be present in a counterfactual world. Table 1 also presents the ESM's predicted causal selection judgments regarding how responsible bacon and smoke are for the annoyed neighbors, denoted $S_{b^1s^1n^1}(B,N)$ and $S_{b^1s^1n^1}(S,N)$, respectively. These predictions are computed by calculating a measure of correlation (specifically, the phi coefficient) between N and both B and S in the ESM's representation of counterfactual possibilities.

Note that Table 1 includes the predictions of what we call a Baseline model that treats all counterfactual judgments as normal causal inferences, that is, ignoring the actual outcome. As its stability parameter s approaches 0, the predictions of the ESM converge to those of the Baseline model.

The Counterfactual Mutation Sampler. The first new model we consider is the *Counterfactual Mutation Sampler*, or the MS/CF (Fig. 1B). The mutation sampler (MS) has been proposed as a model of (non-counterfactual) causal reasoning (Davis & Rehder 2020). The mutation sampler carries out inference on causal Bayesian networks by assuming that when drawing causal inferences, people think about concrete system states—states of a causal system in which each variable of the causal system is instantiated with a value—in order to estimate the system's joint probability distribution, that is, the probability that it will be in any one state. Consider again the cooking bacon example. The MS assumes that reasoners sequentially sample the eight states formed by the three binary variables B, S, and N using Markov chain Monte Carlo (MCMC) methods. Use of MCMC ensures that the generated samples will, after normalization, converge to the true joint distribution as the length of the chain of samples grows large. But as a psychological model the MS assumes that people's capacity for sampling is limited and thus they take relatively few samples. It also assumes that there is a bias in the starting point for sampling: Sampling always starts from one of the causal network's *prototype states*. For example, when the causal relations are all generative, the prototype states are those in which nodes are either all present ($b^1s^1n^1$) or all absent ($b^0s^0n^0$). Together, these assumptions explain some well-known causal reasoning phenomena, such as that reasoners commit *Markov violations* and exhibit *explaining away* that is too weak (Davis & Rehder 2020; Kolvoort et al. 2023; Rehder 2024, 2025; cf., Chaigneau et al. 2026).

The top section of Table 1 presents the predictions of the MS for (non-counterfactual) causal inferences assuming a sampling chain length $\lambda = 12$. The table confirms that the MS yields numerically different conditional probabilities than those predicted by the ESM. The fact that $P(b^1|s^0) \neq P(b^1|s^0n^1)$ illustrates a violation of the Markov condition that states that, for a causal model $B \rightarrow S \rightarrow N$, B and N are independent conditioned on S. In contrast, the ESM and the Baseline model are obligated to predict that $P(b^1|s^0) = P(b^1|s^0n^1)$.

The MS/CF does not use the functional causal model representation of the ESM. Instead, it extends the spirit of the MS to counterfactual reasoning by stipulating that sampling begins at the actual observed state rather than the prototype states. That is, the MS/CF treats counterfactuals just like normal causal inferences except for the change in where sampling starts. Table 1 presents the distinct counterfactual inferences implied by the MS/CF (again, with $\lambda = 12$) as compared to the ESM. Note that ESM's stability parameter s and MS/CF's λ are conceptually related, because as s approaches 0 and λ approaches ∞, both models ignore the actual state and so make the same predictions as the Baseline model.

The Exogenous Sampler. The second new model we consider is the *The Exogenous Sampler*, or the EXS (Fig. 1C). The EXS has elements of both the ESM and the MS/CF. It retains the ESM's functional causal model representation. And, like the ESM, given an actual outcome it draws inferences regarding the states of the exogenous variables. But then, rather than applying a stability parameter, it samples counterfactual possibilities. Specifically, it begins MCMC sampling at the distribution over the exogenous variables entailed by the actual world. The distribution that holds after sampling is complete defines a new functional causal model from which counterfactual inferences are computed.

Table 1 reveals that the EXS makes the same predictions for the causal inferences (where no sampling occurs) as the ESM. But their predictions diverge for the counterfactual and selection judgments. Just as with MS/CF, as λ approaches ∞, EXS's predictions converge to those of the Baseline model.

Overview of Experiment

The main contribution of this article is the introduction and empirical assessment of several alternative models of counterfactual reasoning. But we believe that our experimental approach incorporates other innovations. One is that each subject will be asked *all three* of the judgment types shown in Table. 1, namely, causal, counterfactual, and causal selection judgments (cf. Skovgaard-Olsen et al. 2021). Another is that whereas most previous research has evaluated group performance, the larger number of judgments presented will allow models to be fit to each subject, allowing for the assessment

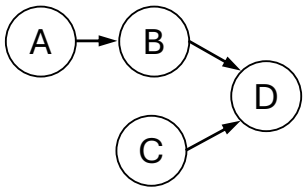

Figure 2: The causal model taught to subjects relating the hormones in mice.

of individual differences. A third will be our testing of causal models that we believe correspond more closely to those that people reason with daily, namely, those that are (a) more complex than those usually tested (they have more variables and variables that are both a cause and effect of other variables), and (b) involve probabilistic causal relations.

Methods

Materials & Design

Fig. 2 presents the causal model that subjects were provided with in the current experiment. Following Lucas & Kemp (2015), the cover story involved the causal relations among hormones in mice. To convey the probabilistic nature of, say, the $A \rightarrow B$ causal relation, subjects were told that if A is present, it "usually" (but not always) causes hormone B and that, because of other factors that can cause it, B is "rarely" present even when A is absent. Subjects were also told that hormones A and C were each present in "some" mice and that B and C were disjunctive causes of D.

Subjects were asked to draw both causal inferences (e.g., $P(c^1|d^0)$) and, given a specific mouse and the observed values of its hormones (e.g., $a^0b^1c^1d^1$), to reason counterfactually (e.g., $P_{a^0b^1c^1d^1}(d^1|b^0)$) and to make causal selection judgments (e.g., $S_{a^0b^1c^1d^1}(B,D)$). Table 2 shows the five mice (named Harold, Albert, Nellie, Ronald, and Martha) that were presented to each subject and their corresponding counterfactual and selection judgments (chosen as those that maximized the differences in the predictions of the models in Table 1).

Procedure

After providing informed consent, subjects carried out the first phase of the experiment, which consisted of reading several screen information about Fig. 2 and then answering 11 multiple-choice questions. The questions asked them to identify the causes(s) and effect(s) of each variable (6 questions) and to indicate how often and why some variables were caused by others (5 questions). If any question was answered incorrectly, the subject was returned to the information screens for re-study. Subjects were required to answer all questions correctly before continuing with the experiment.

In the second phase, subjects drew (non-counterfactual) causal inferences on the basis of the causal model they just learned (i.e., Fig. 2). Fig. 3A shows an example of a causal inference in which a subject judges the probability that a mouse has hormone A given the absence of hormone B (i.e, $P(a^1|b^0)$). There were a total of 23 causal inferences whose presentation order was randomized for each participant.

The final phase presented the five mice (in a random order for each subject). For each mouse, subjects were first presented with the observed values of its hormones and then presented with a block of counterfactual inferences and a block of causal selection judgments about that mouse.

For counterfactual inferences, each trial presented one or more counterfactual conditions and asked subjects to judge the probability of another hormone. For example, in Fig. 3B the subject is asked to judge, for a mouse with all hormones present, the probability of hormone A if hormone B had been observed to be absent (i.e., $P_{a^1b^1c^1d^1}(a^1|b^0)$). The presentation order of each mouse's counterfactual inferences was randomized. Note that each of the counterfactual inferences was also presented in its non-counterfactual form during the earlier causal inference phase, allowing for their comparison.

For the selection inferences, following Quillien & Lucas (2024), subjects were asked about the causes of hormone D. For example, in Fig. 3C the subject is asked how strongly they agree that hormone D was present because of hormone A (i.e., $S_{a^1b^1c^1d^1}(A,D)$). This judgment was presented only for hormones in the same state as D (present if D was present and absent if D was absent). These judgments were presented in a fixed order (A always before B always before C).

The presentation order of the counterfactual and selection blocks was chosen randomly (but was fixed for each subject). For all inferences, subjects responded by moving a tick on a 0%–100% rating scale, with labels at 10% intervals. To ensure that subjects did not have to rely on memory, the information screens were accessible at any time via a help button.

Table 2: The actual hormones of five mice and the corresponding counterfactual and causal selection judgments.

Actual State	Counterfactual Inferences	Causal Selection Judgments
$a^0b^0c^1d^0$	$P(a^1\|c^0)$, $P(b^1\|c^0)$, $P(d^1\|c^0)$	$S(A,D)$, $S(B,D)$
$a^0b^1c^0d^0$	$P(a^1\|c^1)$, $P(b^1\|c^1)$, $P(d^1\|c^1)$	$S(A,D)$, $S(C,D)$
$a^0b^1c^1d^1$	$P(a^1\|c^0)$, $P(b^1\|c^0)$, $P(d^1\|c^0)$, $P(a^1\|d^0)$, $P(b^1\|d^0)$, $P(d^1\|d^0)$	$S(B,D)$, $S(C,D)$
$a^1b^0c^1d^1$	$P(a^1\|c^0)$, $P(b^1\|c^0)$, $P(d^1\|c^0)$	$S(A,D)$, $S(C,D)$
$a^1b^1c^1d^1$	$P(a^1\|b^0)$, $P(c^1\|b^0)$, $P(d^1\|b^0)$, $P(c^1\|a^0b^0)$, $P(d^1\|a^0b^0)$, $P(a^1\|b^0d^0)$, $P(c^1\|b^0d^0)$	$S(A,D)$, $S(B,D)$, $S(C,D)$

Participants

47 New York University undergraduates participated in the experiment for course credit.

Results

Each model was simultaneously fit to each subject's 56 causal, counterfactual, and selection judgments. Model fitting involved estimating free parameters representing the probability of the hormones A and C (parameter c), the strengths of the causal relations (m), the strength of the exogenous causes of B and D (b). The ESM included the stability parameter s and MS/CF and EXS included the chain length parameter λ. To fit the causal selection judgments, the Ss associated with

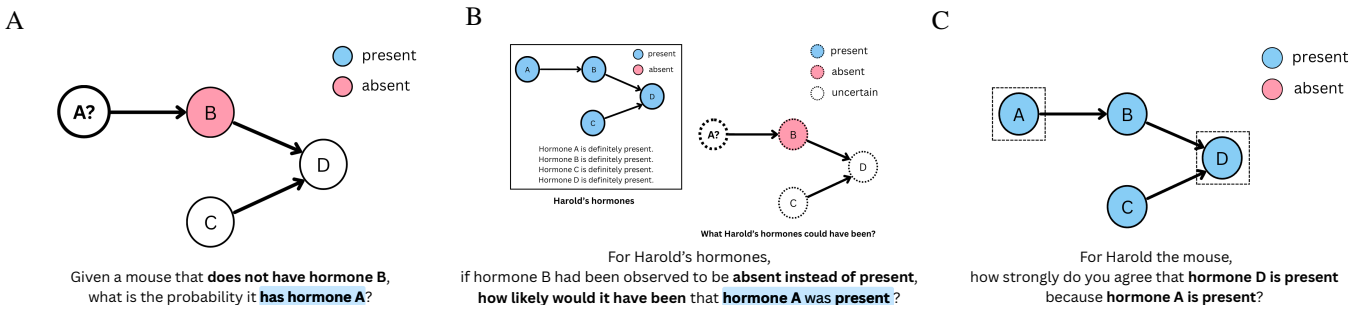

Figure 3: Examples of the three types of inferences. (A) A causal inference in which subjects judged the probability of Hormone A given the absence of Hormone B, i.e., $P(a^1|b^0)$. (B) A counterfactual inference in which, given a mouse with all four hormones ($a^1b^1c^1d^1$), subjects judged the probability of Hormone A if Hormone B had instead been absent, i.e., $P_{a^1b^1c^1d^1}(a^1|b^0)$. (C) A selection inference question in which, given a mouse with all four hormones, subjects rated whether Hormone D was present because of Hormone A, i.e., $S_{a^1b^1c^1d^1}(A,D)$.

each eligible cause in an actual world were scaled by parameter K. The best fitting parameters for each subject were those that minimized squared error and were identified by an initial grid of the parameter space followed by optimization applied to the top three best grid facets.

Table 3 presents each model's best fitting parameters and a number of measures of fit, including the percentage of subjects best fit by each model according to *AIC*, which corrects for a model's number of parameters. Table 3 reveal that the Baseline model yielded the worst fit, indicating that most subjects were indeed reasoning counterfactually. The ESM and EXS yielded the best fits and ones that were very similar to one another. About 70% of the subjects were best fit by either ESM or EXS. The MS/CF yielded a fit substantially worse than ESM or EXS, albeit marginally better than the Baseline model. Yet, only 10% of subjects were best fit by the MS/CF.

Fig. 4 presents an example of subjects' counterfactual inferences for the mouse in which all hormones were observed present ($a^1b^1c^1d^1$) and the corresponding (non-counterfactual) causal inferences. For example, Fig. 4's first sub-panel compares judgments regarding $P(a^1|b^0)$ for a mouse in which all hormones are observed present (and thus b^0 is a counterfactual conditional) and for one where there is no information about the mouse's hormones (and thus $P(a^1|b^0)$ is a normal causal inference). As predicted by the counterfactual models, for most of the inferences shown in Fig. 4 a hormone is judged to be present with greater probability in the counterfactual as compared to the causal inference, reflecting the influence of the actual observed state $a^1b^1c^1d^1$. In contrast, the Baseline predictions are constrained to be the same for the two inference types. As compared to the ESM or EXS, MS/CF often underpredicts subject's rating (e.g., for both the causal and counterfactual versions of $P(c^1|a^0b^0)$).

Table 3: Fits of four models to subject's judgments.

	Median Parameter Estimates					
Model	c	m	b	s	λ	K
ESM	0.508	0.724	0.125	0.264		177
MS/CF	0.523	0.673	0.172		53.9	189
EXS	0.517	0.728	0.120		20.2	174
Baseline	0.522	0.688	0.172			185

	Measures of Fit				
Model	Avg. R	Avg. R^2	Avg. *Loss*	*AIC*	*Pct.S*
ESM	0.676	0.494	16.1	16717	39.6
MS/CF	0.612	0.405	18.0	17242	10.4
EXS	0.675	0.492	16.1	16729	29.2
Baseline	0.566	0.359	18.6	17399	20.8

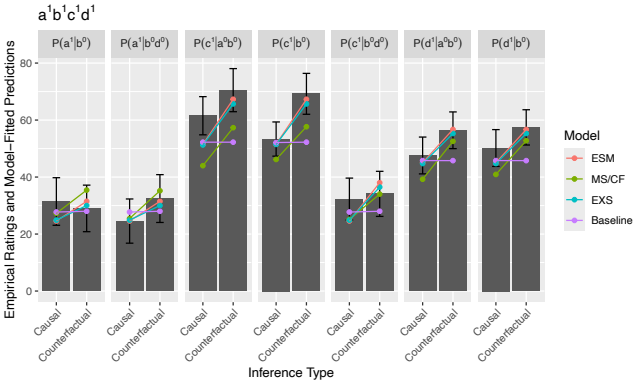

Figure 4: The counterfactual inferences for the mouse in which all hormones are observed present ($a^1b^1c^1d^1$). The corresponding causal inferences are presented side-by-side for comparison. Gray bars are the empirical data and the colored lines are the fitted model predictions. Error bars are 95% CIs.

Overall, the ESM and EXS had considerable success reproducing subjects' average judgments. For example, Fig. 5, which presents the average causal selection judgments for each mouse (actual world), shows they reproduced the overall pattern of those judgments (the correlation between the averaged fitted and observed judgments was about .95). Yet, Fig. 5 shows that they also mis-predicted some items: They under-predicted the judgments for actual state $a^1b^1c^1d^1$ (especially $S_{a^1b^1c^1d^1}(A,D)$) and over-predicted $S_{a^0b^1c^0d^0}(C,D)$. In addition, when computed for each subject and then averaged, the correlation between fitted and observed judgments drops to .73, indicating the presence of substantial unexplained variance at the subject level.

Fig. 6, which presents the counterfactual judgments for three of the five mice, similarly shows successful predictions (correlation between the averaged fitted and observed judgments was again about .95) alongside a few mis-predictions. For example, ESM and EXS over-predict both $P_{a^0b^0c^1d^0}(d^1|c^0)$ and $P_{a^1b^0c^1d^1}(d^1|c^0)$, suggesting that subjects are less likely than the models to think that D might be brought about by causes other than C. Whereas in both $a^0b^0c^1d^0$ and $a^1b^0c^1d^1$ the models judge that there is a sub-

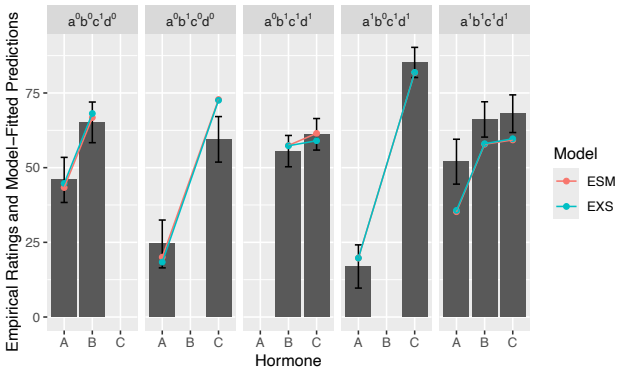

Figure 5: The average ratings for the causal selection judgments. Gray bars are the empirical data and the colored lines are the fitted predictions of the ESM and EXS. Error bars are 95% CIs.

stantial chance that B is (counterfactually) present, subjects seem to be assuming that B is definitely absent, and so judge that D is less likely to be present. This "anchoring" of B to the actual world might have arisen because, for the unbalanced collider structure in Fig. 2, B was viewed as being largely independent of the antecedent C in these judgments. As for the causal selection judgments, the correlation between fitted and observed counterfactual judgments drops considerably (to .64) when averaged over subjects.

To assess individual differences, we conducted a principled component analysis, which revealed three main dimensions of variability. Two of those were well captured by, for the ESM, subjects' causal strength parameter m and the stability parameter s. Fig. 7A confirms that these parameters were distributed widely over subjects. Whereas some subjects treated the causal relations as nearly deterministic ($m > .75$), others treated them as more intermediate in strength. And, although the influence of the actual state on the counterfactual judgments of a few subjects was either quite strong ($s > .75$) or non-existent ($s \approx 0$), the majority were influenced by a mix of the actual state and the underlying causal system. (The third source of variability corresponded to parameter c, the probability of the root causes A and C.)

Finally, Fig. 7B demonstrates the aforementioned relationship between the ESM's stability parameter s and the EXS's

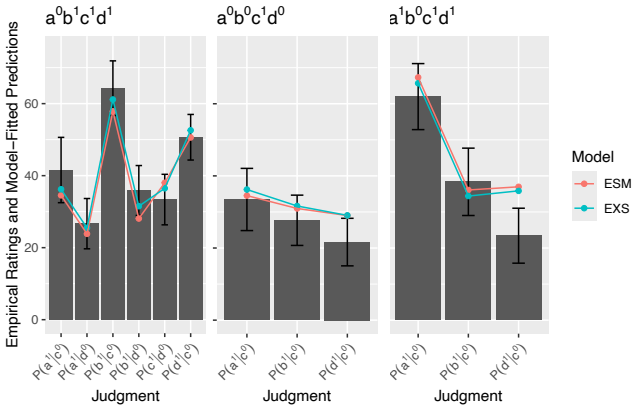

Figure 6: The average counterfactual inferences for three of the five mice. Gray bars are the empirical data and the colored lines are the fitted predictions of the ESM and EXS. Error bars are 95% CIs.

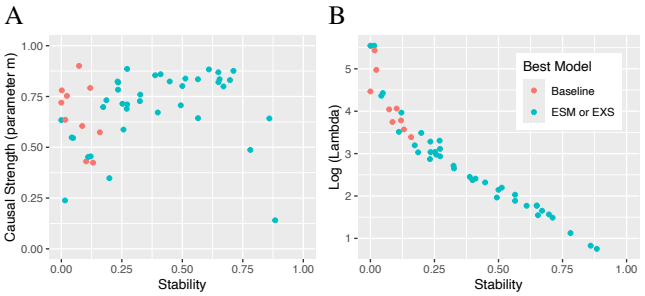

Figure 7: (A) The ESM's fitted values of parameters m and s. (B) The relationship between the ESM's parameter s and the EXS's chain length λ. Red plot points are subjects best fit by the Baseline model and blue ones are those best fit by the ESM or the EXS.

chain length parameter λ. Those participants that were best fit by the ESM with a high/low s were best fit by the EXS with a low/high λ. Not surprisingly, those subjects (21% of the total) who exhibited a low s and a high λ (i.e., who barely distinguished the causal and counterfactual inferences), were fit best by the Baseline model, which makes identical predictions for the causal and counterfactual inferences.

Discussion

We introduced two new models of counterfactual reasoning and tested those models and the ESM in a complex causal scenario that is representative of everyday reasoning (i.e., exhibits a deep causal structure and probabilistic causal relations). We found, first, that subjects indeed reasoned counterfactually, that is, their counterfactual judgments were influenced by the actual state of affairs. Nevertheless, it is notable that the models' parameter that represent the weight given to the actual state (s for ESM and λ for MS/CF and EXS) exhibited large variability over subjects.

The ESM and the EXS yielded the best fits and fits that were quantitatively very similar to one another. This similarity arises because both models draw inferences based on a representation that essentially interpolates between the actual world and the underlying system. Nevertheless, that the EXS appears to be competitive with the leading model of counterfactual and causal selection (the ESM) is important because, as a sampling model, the EXS specifies the cognitive operations that bring rise to such inferences. Thus, unlike the ESM, it is able to make predictions about process-level variables such as RTs, time pressure, within-subject response variability, etc. Still, it is also important to note that although these models reproduced most of subjects' average judgments, their predictions diverged for a number of items and that there remained substantial unexplained variance at the level of individual subject. Thus, there is great opportunity for the introduction of improved models in the future.

This work introduced a methodological innovation, which was to solicit a large number of judgments of different types (i.e., causal, counterfactual, and selection) from the same subject. Because it enables the testing of models across a wide array of conditions and the characterization of individual differences, we hope this method helps establish a new higher bar for the testing of models in this domain.

References

Chaigneau, S. E., Marchant, N., & Rehder, B. (2026). Breaking the chains of independence: A Bayesian uncertainty model of normative violations in human causal probabilistic reasoning. *New Ideas in Psychology*, *81*, 101231. doi: https://doi.org/10.1016/j.newideapsych.2025.101231

Davis, Z., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*, e12839.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599-607.

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

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

Kolvoort, I., Temme, N., & van Maanen, L. (2023). The Bayesian Mutation Sampler explains distributions of causal judgments. *Open Mind: Discoveries in Cognitive Science*, *7*, 318-349.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700–734.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science*, *37*, 977–985.

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

Quillien, T., & Lucas, C. G. (2024). Counterfactuals and the logic of causal selection. *Psychological Review*, *131*(5), 1208–1234.

Rehder, B. (2024). Extending a rational process model of causal reasoning: Assessing Markov violations and explaining away with inhibitory causal relations. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *50*, 1463–1488.

Rehder, B. (2025). A magic act in causal reasoning: Making Markov violations disappear. *Entropy*, *27*, 548.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *150*(12), 2472-2505.