Explaining order effects in counterfactual reasoning

Nicolas Navarre[*1,2], **Tadeg Quillien**[2], **Daniel Lassiter**[2], **Tobias Gerstenberg**[3] **& Neil R. Bramley**[2]
[*]nnavarre@ed.ac.uk
[1]School of Informatics, University of Edinburgh
[2]School of Philosophy, Psychology and Language Sciences, University of Edinburgh
[3]Department of Psychology, Stanford University

Abstract

Human reasoning is often affected by order, such that later judgments depend on earlier ones. For example, order effects have been observed when people answer counterfactual questions. The order in which questions are asked affects whether participants backtrack – imagining how variables that are upstream of a counterfactual change might have been different. Some counterfactual theories predict backtracking while others do not. We build on this prior work and develop a computational model of counterfactual reasoning that captures order effects. We evaluate different versions of this model on existing empirical results, finding that a model which backtracks and produces systematic order effects best explains human judgments. We conclude by discussing implications of the finding that counterfactual reasoning requires a context-sensitive evidence source for both resource-rational and discourse-coherent explanations of reasoning.

Keywords: counterfactuals; backtracking; reasoning; order effects; causality

Introduction

If I had not become a widower, perhaps my life would have been different; I would not be General Bolívar or the Liberator[. . .]

—Simón Bolívar (Perú de Lacroix, 1924)

This reflection from Bolívar's diary illustrates how readily we reason counterfactually. Bolívar's contention that the death of his wife shaped the course of South American history is intuitively plausible. It cannot be empirically validated because the scenario cannot be repeated, yet we can engage with it, agree or disagree with it or be informed by it. Counterfactual reasoning is core to several higher-order cognitive processes including reasoning (Lucas & Kemp, 2015; Quillien et al., 2023), explanation (Lombrozo, 2010; Quillien, 2020), judgment (Gerstenberg, 2024; Icard et al., 2017; Quillien & Lucas, 2023), responsibility attribution (Lagnado et al., 2013; Zultan et al., 2012), as well as the indirect communication of causal beliefs (Kirfel et al., 2022; Navarre et al., 2024) or social evaluations (Davis et al., 2025). Therefore, it is important to understand the cognitive processes involved in the conception and comparison of counterfactual possibilities to better understand their role in facilitating communication and judgment.

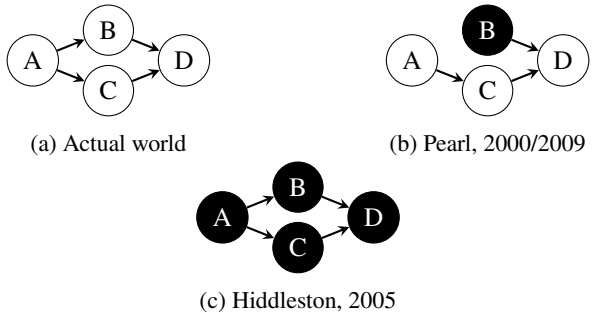
In this paper, we enrich existing theories of counterfactuals by investigating successive counterfactual judgments. We achieve this by re-analyzing previous data on counterfactual reasoning with sequential counterfactual judgments. In this data, responses to a counterfactual query differ systematically based on the previous queries the participant has responded to – an effect we term *counterfactual order effect*. We present an extension of existing theories of counterfactual reasoning to account for these counterfactual order effects. We discuss the implications of this theory for human cognition including resource rational inference and discourse coherent language.

Background

Consider the following scenario (from Pearl, 2000/2009):

> **The firing squad:** A prisoner (D) is to be executed by a firing squad consisting of two marksmen (B and C) under the command of officer (A). The prisoner dies if at least one of the marksmen shoots. The officer gives the order to shoot ($A = 1$). Both marksmen shoot ($B = C = 1$), and the prisoner dies ($D = 1$, Figure 1a).

If marksman B had not shot, would prisoner D still have died? Would marksman C still have shot? Would officer A still have given the order? Two prominent theories make opposing predictions about what the right answer to each question is (see Lucas & Kemp, 2015; Rips, 2010, for a thorough review of both theories). Pearl (2000/2009) treats counterfactuals as surgical interventions affecting their causal consequences but leaving their antecedent causes unaffected. This account predicts that, if marksman B had not shot, the officer would still have given the order and C would still have shot, with the result that the prisoner D would still have died (Figure 1b). Hiddleston's (2005) theory works differently: rather



(a) Actual world (b) Pearl, 2000/2009

(c) Hiddleston, 2005

Figure 1: (a) Firing squad scenario. White = true, black = false. (b–c) *"If marksman B had not shot, would A, C and D have still occurred?"*.

than using the notion of an intervention to resolve the inconsistency between actual and counterfactual scenarios, it relies on *backtracking*, a reasoning process in which one makes inferences from a counterfactual supposition about how causal parents must have been different to explain the change. In the firing-squad scenario, Hiddleston's *minimal-network* account prioritizes preserving the causal laws of the scenario while allowing there to be other counterfactual changes to the events that actually happened. Here, to make sense of marksman B's counterfactual failure to shoot we might backtrack to infer that commander A might not have given the command in that case. As a result of this, it is unlikely that marksman C would have shot (without being ordered to), and so unlikely that the prisoner D would have died (without being shot; Figure 1c).

To backtrack or not to backtrack Several studies have shown that people sometimes treat counterfactuals as interventions and sometimes explain them by backtracking (Lucas & Kemp, 2015; Rips & Edwards, 2013; Sloman & Lagnado, 2005). The degree to which people backtrack depends on the reliability of the causal connections, and the kind of counterfactual query being asked (Dehghani et al., 2012; Gerstenberg et al., 2013; Sloman & Lagnado, 2005).

Gerstenberg et al. (2013) asked participants the three questions in a scenario similar to the firing squad (here adapted to our scenario): 'If marksman B had not shot,'

(i) would prisoner D have died?

(ii) would marksman C have shot?

(iii) would officer A have given the order to shoot?

Between subjects, they varied the order in which participants were asked about each variable, contrasting judgments to queries in the order DCA against the reverse order ACD. The ACD order produced more backtracking than the DCA order. This shows that the extent to which people backtrack depends on the order in which counterfactual queries are made.

Order effects Sequential effects are not unique to counterfactual reasoning. They are also found in choice-reaction tasks, reasoning tasks, and probabilistic inference tasks (Gershman & Goodman, 2014; Hyman, 1953; Prat-Carrabin & Woodford, 2024; Tversky & Kahneman, 1974). In all of these tasks, participants' responses are systematically affected by their previous responses. To many psychologists, this has suggested the existence of errors in human judgment leading to several accounts of biased judgment departing from a normative standard (Benjamin, 2019). Meanwhile, another view has aimed to understand these effects as a feature of making inference and judgments with limited cognitive resources (Bramley et al., 2017; Dasgupta et al., 2020; Griffiths et al., 2015; Lieder & Griffiths, 2020).

From a linguistic perspective, order effects might arise from the need to remain consistent and coherent with a conversational context (Roberts, 1989). In what follows, we develop a model of counterfactual reasoning that builds on this notion of committing to items in a discourse. Before doing so, we describe existing formal frameworks for modeling counterfactual reasoning.

Modeling counterfactuals

Many models of counterfactual reasoning build on Structural Causal Models (SCMs) to represent the causal system (Pearl, 2000/2009). We illustrate how SCMs work via the firing squad scenario.

Representation: In the context of the firing-squad we can define an SCM representation $\mathcal{M} := \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ where $\mathbf{V} = \{A, B, C, D\}$ represents the officer's order, marksman B's shot, marksman C's shot, and the prisoner's death, respectively. $\mathbf{U} = \{\beta_A, \beta_B, \beta_C, \beta_D, \theta_{AB}, \theta_{AC}, \theta_{BD}, \theta_{CD}\}$ represents exogenous variables. $P(\mathbf{U})$ assigns a probability to each exogenous variable. The structural equations $\mathcal{F}$ can be defined as:

$$
\begin{aligned}
A &:= \beta_A \\
B &:= (A \wedge \theta_{AB}) \vee \beta_B \\
C &:= (A \wedge \theta_{AC}) \vee \beta_C \\
D &:= (B \wedge \theta_{BD}) \vee (C \wedge \theta_{CD}) \vee \beta_D .
\end{aligned}
$$

A graphical illustration of $\mathcal{M}$ is shown on the left in Figure 2. This representation contains additional exogenous variables relative to Pearl's model. The expanded representation has two purposes. First, it allows us to introduce a standard noisy-OR parameterization (Cheng, 1997) whereby effects can occur by themselves (due to their base rates) while causes have a strength or propensity to produce their effects. Causes combine independently, such that an effect occurs if caused by its base rate or by any successful active causes. In the structural model $\mathcal{M}$ the base rates of variables are denoted as β_X for any endogenous variable $X \in \mathbf{V}$. Causal powers are denoted as θ_{XY}, stating the causal power of the connection between any parent X and its child $Y \in \mathbf{V}$. Second, the additional exogenous variables allows our model to capture the interpretation of *local* interventions as inference to the state of the exogenous variables in a fully backtracking semantics of counterfactuals, as explained below. That is, a variable like B can be changed from $B = 1$ to $B = 0$ either by imagining that $A = 0$ and $\beta_B = 0$ or that $\theta_{AB} = 0$ (a local intervention).

In all models of counterfactual inference we present here, the counterfactuals are computed by creating a counterfactual "twin-world" and inferring the state of queried variable on the counterfactual side. The structural equations of the twin are identical to those of the original model, but the exogenous variables are connected to their actual world counterparts. This creates a larger structural model which we will denote as $\mathcal{M}^*$ – see Figure 2.

Counterfactual inference: We frame counterfactual inference as the response to a query of the form 'If X had not happened, would Y have happened?', with some actual world observations O. Supposing that we have a full counterfactual representation $\mathcal{M}^*$ with real and counterfactual twins of each variable, we can compute the inferred distribution on Y to

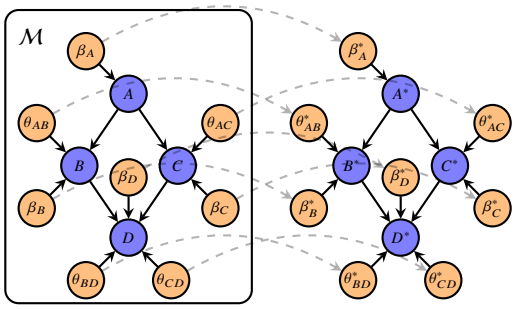


Figure 2: Full causal representation $\mathcal{M}^*$ of the firing-squad scenario with the counterfactual "twin-world" including endogenous variables (blue) and exogenous variables (orange). The subgraph enclosed by the box outlines the original structural model, $\mathcal{M}$, that the twin network copies. All counterfactual variables in $\mathcal{M}^*$ are denoted with a superscript asterisk – e.g. β^*_A.

produce a judgment. We define the counterfactual inference for the query on X and Y on the counterfactual (right hand) side of $\mathcal{M}^*$, so we make an inference based on X^* and Y^*. The benefit here is that counterfactual inference turns into probabilistic inference over $\mathcal{M}^*$, and a counterfactual inference can be defined as:

$$P_{\mathcal{M}^*}(Y^* = y^* \mid X^* = x, O = o) \tag{1}$$

The theories differ primarily in how exogenous variables in the twin-world relate to those of the actual world. Here, we briefly describe how each works.

Structural model: Pearl's (2000/2009) counterfactuals can be computed in a twin-world where the actual and counterfactual variables are connected to the same exogenous variables. Consequently, the counterfactual constraint is treated as an intervention where the counterfactual twin prunes its causal parents.

See Figure 1b as a simplified example of how the structure is modified.

Extended Structural Model (ESM): Lucas and Kemp's (2015) ESM extends Pearl's theory by positing that the counterfactual twin has its own exogenous variables, but each one is connected to its real counterpart (shown with gray dashed lines in Figure 2). In the ESM the counterfactual distribution on exogenous variables is given by:

$$P(U^*_i|U_i) = s\delta(U_i) + (1-s)P(U_i), \tag{2}$$

where $\delta(U_i)$ is the inferred value of the exogenous variable in the actual world, and $0 \leq s \leq 1$, is a stability parameter that mediates how much the counterfactual version either copies that state or returns to the prior $P(U_i)$. In general, the theory treats $\delta(U_i)$ as a posterior inference of U_i in the original model given the observation $P_{\mathcal{M}}(U_i|O = o)$ (Quillien et al., 2023). The ESM also allows counterfactuals to be treated as observations (with no change in structure but a revision of the distribution over exogenous variables) or as interventions (pruning the link to their parents).

Modeling order effects

While some computational models can model backtracking inferences (Lucas & Kemp, 2015; von Kügelgen et al., 2023), no current theory explains why backtracking inferences depend on the order in which questions are posed as seen in Gerstenberg et al. (2013). Here, we will propose a model of counterfactual reasoning with sequential commitment. Note that we consider the case where reasoners make binary judgments ('would the effect have happened or not?'). Our account has two components:

(i) Judgments are sampled from a distribution that depends on the relevant counterfactual probabilities.

(ii) Participants tend to commit to their previous judgments, using them as part of the counterfactual context when answering subsequent queries.

In defining a counterfactual representation $\mathcal{M}^*$, we keep the distribution of actual and counterfactual variables as in Equation 2 to maintain the psychologically interpretable stability parameter used in Lucas and Kemp (2015). The key components of our novel proposal are as follows:

Judgment: We model the process of making a counterfactual judgment, p_{J_1}, from the inferred probability of the query p_{Q_1} using a softmax choice rule with temperature $\tau > 0$ (Luce, 1959):

$$p_{J_1} = \frac{e^{p_{Q_1}/\tau}}{e^{p_{Q_1}/\tau} + e^{(1-p_{Q_1})/\tau}} \tag{3}$$

Sequential commitment: Participants remember their previous judgments and probabilistically incorporate them into the background information for future counterfactual judgments. We assume that previous judgments are treated as observations in future judgments with probability P_{keep}. Figure 3 illustrates an example of sequential commitment with $P_{\text{keep}} = 1$.

Formally, for any query Q_n, $n > 1$, and any set of previous judgments $\{j_i | 1 \leq i < n\}$, Q_n is answered by updating the evidence set in the twin-world model to include previous judgments as counterfactual evidence $O^* := \bigcup_{1 \leq i < n} \{Y^*_i = j_i\}$, with each previous judgment (independently) included with probability P_{keep}. That is, P_{keep} models an individual's propensity to remember or store each judgment for future inferences. The inferred probability for the next query, p_{Q_n}, is then given by:

$$P_{\mathcal{M}^*}(Y^*_n = y_n \mid O = o, X^*_n = x_n, O^* = o^*), \tag{4}$$

with O^* varying according to the distribution induced by the parameter P_{keep}.

Intuitively, the notion of sequential commitment can be seen as modeling a participant's attempt to maintain coherence among successive counterfactual judgments. That is, participants do not view each judgment as being unrelated to the rest: instead they view the task as one of building up a counterfactual scenario, and so previous choices are treated as context for subsequent choices in continuing the discourse. From this perspective, one can view sequential commitment as a species of the well-studied linguistic phenomenon of *modal*

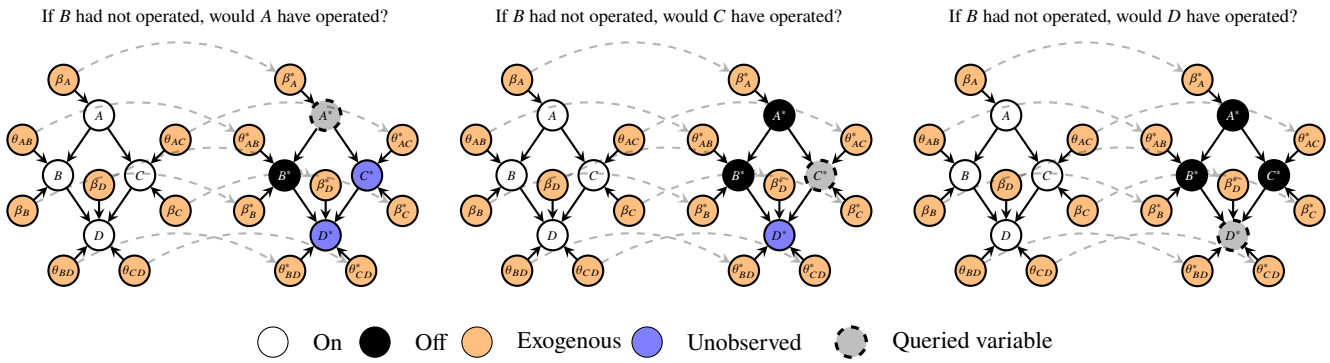


Figure 3: Example of sequential commitment to previous judgments in a sequence of counterfactual queries (from left to right) in the firing-squad scenario. Each query specifies counterfactual observations in the twin model including the actual world observations ($A = B = C = D = 1$), the counterfactual constraint ($B^* = 0$), and previous judgments – if any – at query time. Commitment means adding the previous judgments to the counterfactual observation set for the current query. In this example, the participant answers 'No' to the first two queries in the (ACD) order. The 'No' to the first query about A^* is carried over to the second query, and the 'No' to the query about C^* is carried over to the third query, making it likely that the answer to that query will be 'No' as well.

subordination (Roberts, 1989). Alternatively, sequential commitment might be a form of *amortization*, where reasoners commit to their previous choices to simplify the current inference (Gershman & Goodman, 2014). Our formal model is independent of this interpretive question. Our model was implemented using the probabilistic graphical model inference toolkit 'pgmpy' (Ankan & Textor, 2024). All code and supplementary material is available in our public repository[1].

Special cases

Here, we highlight a few special cases of the model. First, consider the extremes of the commitment-rate parameter. When $P_{\text{keep}} = 1$, then the full set of previous judgments is certain to be included as counterfactual evidence O^*. When $P_{\text{keep}} = 0$, none of the previous judgments are included as counterfactual evidence O^*, which amounts to inferring a counterfactual query as in Equation 1, regardless of any prior judgments. By fitting P_{keep} to human judgments, we get an indication to what extent people really do commit to their previous judgments when making counterfactual judgments in order.

The model setting when $s = 0$ bases all counterfactual inferences on the prior. That is, when $s = 0$ no information from the actual world observations is carried over to the twin network. Counterfactual inference then becomes equivalent to a normal causal conditional inference. When $s = 1$, certain counterfactual queries become undefined in the backtracking model. For example if we observe $A = 1$, this makes it certain that exogenous variable $P(\beta_A) = 1$ and with $s = 1$ we carry this certainty over such that $P(\beta^*_A) = 1$. There is then no "room" to backtrack because the model admits no uncertainty about the value of $P(\beta^*_A)$. Rather than leaving this setting undefined, we treat it as using interventions in the counterfactual twin equivalent to Pearl's theory.

[1]https://github.com/navarrenicolas/backtracking_counterfactuals.git

Comparing theories of backtracking and order effects

In this section, we use our computational framework for modeling order effects to compare several theories of counterfactual reasoning using experimental data from Gerstenberg et al. (2013). Since the theories considered are special cases of the general framework described above, this approach allows us to consider whether incorporating sequential commitments in a theory of backtracking counterfactuals is empirically justified.

We base our model implementations on two broad groups: backtracking vs. interventional and sequential vs. contextless. We use cross-validated model comparisons as an evaluation of which theory best captures participants' judgments in our evaluation dataset.

Methods

Evaluation dataset To evaluate our models we use the order effect data from Experiment 2 in (Gerstenberg et al., 2013) (N=320). This study involves counterfactual judgments in an abstract firing-squad-like scenario with nodes labeled A, B, C, and D as in Figure 1. Each event was an abstract component that is either active or inactive. This experiment featured $2 \times 2 \times 2 = 8$ between-subject conditions, varying the nature of the observation; the combination function governing B and Cs influence on D; and the order in which a series of three counterfactual queries were then posed. Components were observed to be all on or all off (ON/OFF). D was either a disjunctive or conjunctive function of B and C (DISJ/CONJ). Queries always asked what would have happened if B's state were reversed (e.g., OFF if it was actually ON in the observation). Participants were asked about the state of the other three variables in one of two orders (ACD/DCA).

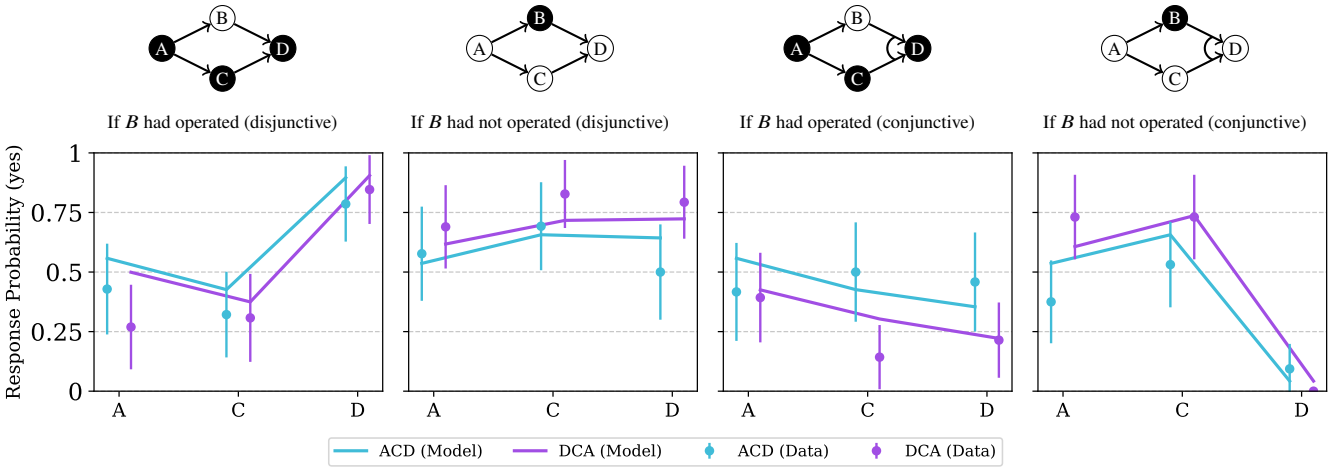


Figure 4: Model predictions of the best fitting model (Backtracking-Sequential) against the empirical data. The structures represent the experimental conditions with their original states with variable B modified according to the counterfactual query. Structures with the joined connections have a conjunctive rule for D, such that both B and C are needed to make D happen.

Evaluation We evaluate our counterfactual reasoning framework with a cross-validated model comparison. We hold out one of the (ON/OFF, DISJ/CONJ, ACD/DCA) conditions as a test set and use the remaining conditions as a training set totaling 8 cross-validation folds. Each cross-validation fold is fit via maximum likelihood estimation (MLE). We score the model predictions of each fold as the negative log-likelihood of the test data given the model fit to the training data. The final cross-validation score is averaged across all 8 folds. We also report AIC and BIC scores for each model's fit.

Models to evaluate Based on our hypotheses we categorize the models into two broad groups: backtracking vs. interventional (Pearl) and sequential vs. contextless. The backtracking and non-backtracking models are distinguished by whether they contain a free stability parameter $0 \leq s < 1$ (backtracking) or a fixed stability of $s = 1$ (non-backtracking). The sequential models assume that the commitment rate is non-zero ($0 < P_{\text{keep}} \leq 1$). The contextless models will have a commitment rate of zero, $P_{\text{keep}} = 0$. For models with context, the temperature parameter $\tau > 0$ interacts with P_{keep} in predicting judgments. We test each of these theories within a structural model $\mathcal{M}$ as in Figure 2, with one constraint where $P(\beta_B) = P(\beta_C)$. For robustness, we also considered several constrained versions of the structural representation and ablations of the inference parameters. We report one implementation, but our results are generally consistent across all ablations. Full results are in the supplementary materials.

Table 1: Evaluation scores.

Model	NLL	CV-NLL	AIC	BIC	k
Backtracking-Sequential	**502.0**	**64.6**	**1024.1**	**1071.1**	10
Pearl-Sequential	574.7	72.2	1167.3	1209.7	9
Backtracking-Contextless	572.9	75.4	1163.7	1206.0	9
Pearl-Contextless	612.6	77.2	1241.3	1278.9	8
Baseline (random)	665.4	83.17	–	–	–

Note: NLL = negative log-likelihood, CV-NLL = mean cross-validation loss, k = number of parameters.

Model comparison results

Table 1 summarizes the results of our model comparisons, and the parameter fits for each of the models are shown in Table 2. Overall, the backtracking model with sequential updating outperforms the other models across all 3 evaluation metrics. Furthermore, this model reproduces the order effects in all conditions of the original data (see Figure 4). For the contextless models, we also see that the backtracking semantics of counterfactuals outperforms the interventional semantics across all three evaluation metrics. Moreover, models with commitment to previous judgments consistently out-perform models with no commitment, regardless of the counterfactual semantics used. Among backtracking models, the sequential commitment model has a significantly better fit than the contextless version ($\chi^2(1) = 141.645$, $p < 0.001$). The same pattern is found when comparing the sequential and contextless intervention models ($\chi^2(1) = 75.95$, $p < 0.001$).

These results provide evidence that people reason with a backtracking semantics of counterfactuals, and also frequently commit to previous judgments when making new judgments.

Discussion

We introduced a model of counterfactual reasoning in a context where reasoners make several judgments about a given scenario. Our model is based on the idea that reasoners engage in *sequential commitment*, adding their previous responses to the current context when making a counterfactual judgment. We find that this can explain why people's counterfactual judgments exhibit order effects (Gerstenberg et al., 2013), and can

Table 2: Parameter fits for compared models.

Model	s	τ	P_{keep}	β_B, β_C	β_A	β_D	θ_{AB}	θ_{AC}	θ_{BD}	θ_{CD}
Backtracking-Sequential	1.00*	0.32	0.66	0.24	0.47	0.00	0.59	0.73	0.78	0.98
Pearl-Sequential	**1**	0.98	0.97	0.09	0.36	0.00	0.70	1.00	1.00	1.00
Backtracking-Contextless	0.29	0.38	**0**	0.55	0.55	0.00	0.36	0.00	0.86	0.94
Pearl-Contextless	**1**	1.33	**0**	0.25	0.67	0.00	0.75	0.75	1.00	0.92

Note: Asterisk marks stability approaching but not exactly 1, which would be undefined in the backtracking model. Boldface marks fixed parameter values.

systematically account for the direction of these effects.

Normative vs. descriptive

A popular approach to interpreting order effects in psychological tasks has been to consider that human inference deviates from 'normative' or 'rational' standards due to simple heuristics (Benjamin, 2019; Tversky & Kahneman, 1974). Others have proposed that order effects are a byproduct of the rational use of limited cognitive resources (Griffiths et al., 2015; Lieder & Griffiths, 2020). Gershman and Goodman (2014) present a resource-rational account of order effects in probabilistic reasoning where they posit that people will reuse computations from previous probabilistic queries to answer subsequent queries, a process termed *amortized inference*. Our model of counterfactual reasoning with commitment to previous judgments is similar to amortized inference as we can treat previous judgments as evidence in the counterfactual twin-network, simplifying the computation over the larger extended structural model. This amortization process also connects to a related literature linking the role of memory for constructing mental simulations of future events (De Brigard, 2014). As noted above, order effects might also stem from a pressure to construct a coherent scenario in sequential judgment (cf. Roberts 1989). Our account of counterfactual order effects is consistent with both perspectives.

Backtracking in counterfactual reasoning

Our model has also implications for debates about counterfactual backtracking. One possible account of backtracking is that participants backtrack when they interpret a counterfactual premise as an *observation* (e.g., *seeing* the soldier shoot provides evidence that the captain gave the order), and do not backtrack when they interpret the counterfactual premise as an *intervention* (e.g., *making* the soldier shoot has no effect on the captain). Our best-fitting model assumes that reasoners interpret counterfactual premises as observations rather than interventions.

In this light, why does our model predict that participants will sometimes backtrack and sometimes not? In essence, it assumes that people always backtrack, but that they frequently backtrack to proximal unobserved variable, leading to no change in the observed variables.

Consider for example a participant who affirms that the captain would not have given the order if soldier B had not shot. Under our account, the participant constructed the counterfactual scenario by inferring from the soldier's non-shooting that the captain decided not to give the order (an observed variable). In this model, the order of questions influences whether participants observably backtrack because it influences whether they change the value of an observed or an unobserved variable when making an inference. Consider a participant who is first asked whether the prisoner would still have died if soldier B had not shot, and that participant judges that the prisoner still would have died, and adds this judgment to the context for the next query. Next, the participant is asked whether the captain would have given the order. Since the participant now assumes the prisoner would have died, this is evidence that the captain did in fact give the order, and so soldier B's refusal to shoot is now best explained by appeal to an unobserved variable $\beta_{AB} = 0$ (soldier B ignored the order) instead of an observed variable ($\beta_A = 0$ the captain not giving the order).

We speculate that many of the factors that influence backtracking inferences could be understood in this way. For example, people's tendency to backtrack depends on the linguistic probe used to elicit a judgment (Rips & Edwards, 2013; Sloman & Lagnado, 2005): people are for instance more likely to backtrack when told to imagine that a device had 'not operated' than 'failed' (Rips & Edwards, 2013). The contrast in these counterfactual judgments could be due to the fact that 'failed' suggests that the explanation for the device's failure is internal to the device (prompting an inference to an unobserved variable), whereas 'not operated' does not, leaving the reasoner to make inferences about observable variables. In both cases the reasoner is making a backtracking inference, but it only shows as such in the latter case.

Where does backtracking end? Under our implementation, backtracking ends when the reasoner attributes the counterfactual contrast to a difference in an exogenous "noise" variable. Our capacity to represent and reason over indirect causes, as well as the value of doing so, is limited and context dependent. The question of how far a given reasoner will backtrack thus limited by what they choose to represent endogenously in their model of the situation. Future work may therefore explore to what extent backtracking phenomena can be subsumed by a general account of ad hoc causal model construction.

Acknowledgments

NN was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. TG was supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and from the Cooperative AI Foundation. DL was supported by the UK Arts and Humanities Research Council [grant AH/Z507490/1]. AI tools on Github Copilot were used to assist with developing code, figures, and diagrams. No AI tools were used in writing the manuscript.

References

Ankan, A., & Textor, J. (2024). Pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, *25*(265), 1–8.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, *2*, 69–186.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Davis, Z. J., Allen, K. R., Kleiman-Weiner, M., Jara-Ettinger, J., & Gerstenberg, T. (2025). Inference from social evaluation. *Journal of Personality and Social Psychology*.

De Brigard, F. (2014). Is memory for remembering? recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155–185.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55–85.

Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, *28*(10), 924–936.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188.

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

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481–1501.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

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

Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.

Navarre, N., Konuk, C., Bramley, N. R., & Mascarenhas, S. (2024). Functional rule inference from causal selection explanations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Pearl, J. (2000/2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Perú de Lacroix, L. (1924). *Diario de bucaramanga: Vida pública y privada del libertador simón bolívar*.

Prat-Carrabin, A., & Woodford, M. (2024). Imprecise probabilistic inference from sequential data. *Psychological Review*.

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*.

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

Quillien, T., Szollosi, A., Bramley, N. R., & Lucas, C. (2023). Causal inference shapes counterfactual plausibility. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

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.

Roberts, C. (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy*, *12*(6), 683–721.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, *29*(1), 5–39.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments re-

veal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

von Kügelgen, J., Mohamed, A., & Beckers, S. (2023). Backtracking counterfactuals.

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