Explaining Order Effects in Counterfactual Reasoning

Nicolas Navarre

May 10, 2026

THE UNIVERSITY
of EDINBURGH

A true story

A true story

A true story

"If I had not become a widow, I would not become the great general Bolívar, nor the Liberator."

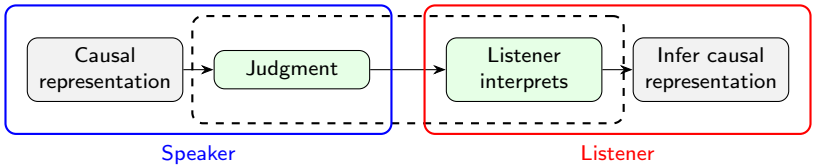
— Simón Bolívar

Causal representations and judgment

Counterfactual and causal judgments communicate causal beliefs!

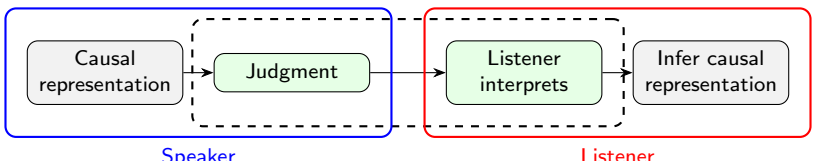
Causal representations and judgment

Counterfactual and causal judgments communicate causal beliefs!



Causal representations and judgment

Counterfactual and causal judgments communicate causal beliefs!



Goal: Understand how counterfactual reasoning affects judgment and interpretation: A semantics for counterfactuals that is *psychologically plausible*.

Other research

- Causal explanations shape probabilistic reasoning
- Causal selection judgments and functional rule inference
- Agent-based model of moral value inference in a network
- Dataset of embedded clause sentences for linguistics research looking at syntax/semantics interface

Other research

- Causal explanations shape probabilistic reasoning
- Causal selection judgments and functional rule inference
- Agent-based model of moral value inference in a network
- Dataset of embedded clause sentences for linguistics research looking at syntax/semantics interface

Today:

- Order effects in counterfactual reasoning with backtracking counterfactuals!

Outline

Order effects in counterfactual reasoning

A paradigmatic example

The firing squad

A paradigmatic example

The firing squad

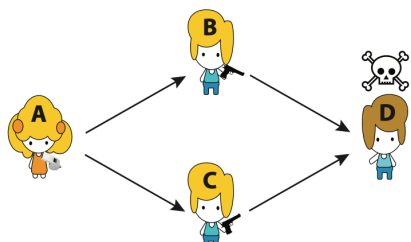
A prisoner (D) is to be executed by a firing squad consisting of two marksmen (B and C) under the command of officer (A). The prisoner dies if *at least one* of the marksmen shoots.

A paradigmatic example

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.

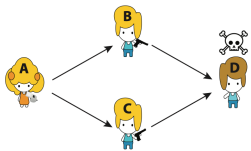
Today, the officer gives the order to shoot ($A = 1$). Both marksmen shoot ($B = C = 1$), and the prisoner dies ($D = 1$).



(a) What actually happened

Two perspectives on counterfactual reasoning

What does a semantics of counterfactuals look like?



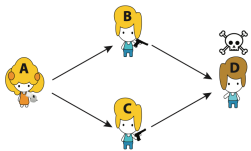
(a) What actually happened

If Marksman B had not shot,

- would prisoner D have died?
- would Marksman C have shot?
- would commander A have sent out the order?

Two perspectives on counterfactual reasoning

What does a semantics of counterfactuals look like?

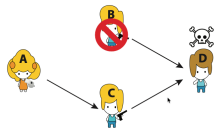


(a) What actually happened

If Marksman B had not shot,

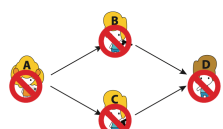
- would prisoner D have died?
- would Marksman C have shot?
- would commander A have sent out the order?

Interventional:



(b) Pearl's (2000) prediction

Backtracking:

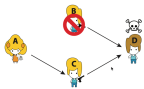


(c) Hiddleston's (2005) prediction

Do we backtrack?

Factors that affect backtracking inferences

Interventional:



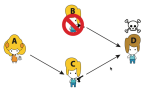
(b) Pearl's (2000) prediction

More likely to happen when:

Do we backtrack?

Factors that affect backtracking inferences

Interventional:



(b) Pearl's (2000) prediction

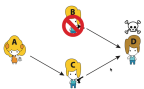
More likely to happen when:

1. Active query: If Marksman *B*'s gun had *failed*

Do we backtrack?

Factors that affect backtracking inferences

Interventional:



(b) Pearl's (2000) prediction

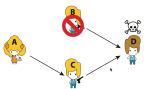
More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak

Do we backtrack?

Factors that affect backtracking inferences

Interventional:



(b) Pearl's (2000) prediction

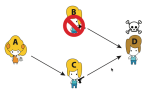
More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Do we backtrack?

Factors that affect backtracking inferences

Interventional:

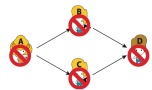


(b) Pearl's (2000) prediction

More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Backtracking:



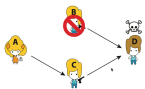
(c) Hiddleston's (2005) prediction

More likely to happen when:

Do we backtrack?

Factors that affect backtracking inferences

Interventional:

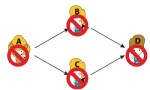


(b) Pearl's (2000) prediction

More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Backtracking:



(c) Hiddleston's (2005) prediction

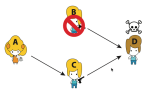
More likely to happen when:

1. Passive query: If Marksman B's gun had *not operated*

Do we backtrack?

Factors that affect backtracking inferences

Interventional:

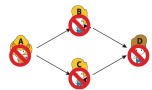


(b) Pearl's (2000) prediction

More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Backtracking:



(c) Hiddleston's (2005) prediction

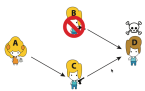
More likely to happen when:

1. Passive query: If Marksman B's gun had *not operated*
2. Causal link $A \rightarrow B$ is strong

Do we backtrack?

Factors that affect backtracking inferences

Interventional:

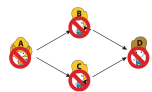


(b) Pearl's (2000) prediction

More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Backtracking:



(c) Hiddleston's (2005) prediction

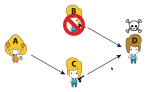
More likely to happen when:

1. Passive query: If Marksman B's gun had *not operated*
2. Causal link $A \rightarrow B$ is strong
3. Query order: would A? would C? would D?

Do we backtrack?

Factors that affect backtracking inferences

Interventional:

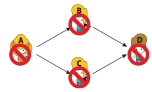


(b) Pearl's (2000) prediction

More likely to happen when:

1. Active query: If Marksman B's gun had *failed*
2. Causal link $A \rightarrow B$ is weak
3. Query order: would D? would C? would A?

Backtracking:



(c) Hiddleston's (2005) prediction

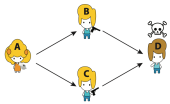
More likely to happen when:

1. Passive query: If Marksman B's gun had *not operated*
2. Causal link $A \rightarrow B$ is strong
3. Query order: would A? would C? would D?

1. Kind of query (Sloman and Lagnado, 2005; Rips and Edwards, 2013):
2. Stength of causal links (Dehghani et al., 2012; Gerstenberg et al., 2013)
3. **Query order (Gerstenberg et al., 2013)**

Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)

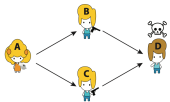


(a) What actually happened

What happens when we change the order of counterfactual queries?

Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)

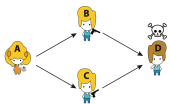


(a) What actually happened

What happens when we change the order of counterfactual queries?
If B had not happened,

Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)



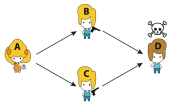
(a) What actually happened

What happens when we change the order of counterfactual queries?
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)



(a) What actually happened

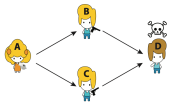
What happens when we change the order of counterfactual queries?
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?

Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)

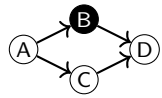


(a) What actually happened

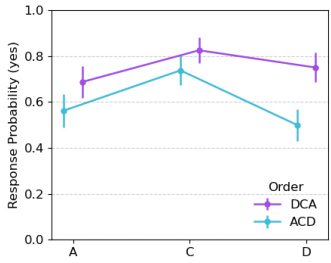
What happens when we change the order of counterfactual queries?
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?

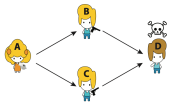


If B had not operated



Order effects in counterfactual reasoning

Gerstenberg et al. (2013) Experiment 2 (N = 320)

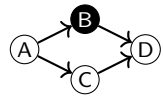


(a) What actually happened

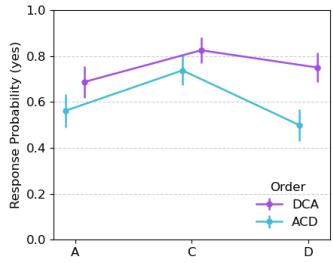
What happens when we change the order of counterfactual queries?
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?



If B had not operated



Why are the judgments systematically different?

No current theory of counterfactuals account for these differences.

Modeling counterfactual order effects

Modeling counterfactual order effects

Model breakdown

Part 1 Representation

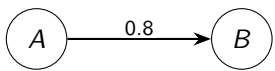
Part 2 Counterfactual semantics

Part 3 Sequential inference

Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

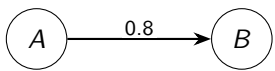
Causal Bayes Net (CBN)



Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

Causal Bayes Net (CBN)

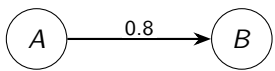


Stochasticity is ‘baked’ into the graph.

Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

Causal Bayes Net (CBN)



Stochasticity is 'baked' into the graph.
Noisy-OR parameterization:

$P(B = 1 \mid A; \beta_B) = 1 - (1 - \beta_B)(1 - 0.8)^A$

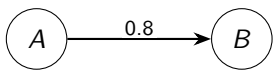
$P(A) = 0.1$

$\beta_B = 0.1$

Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

Causal Bayes Net (CBN)



Stochasticity is 'baked' into the graph.
Noisy-OR parameterization:

$P(B = 1 \mid A; \beta_B) = 1 - (1 - \beta_B)(1 - 0.8)^A$

$P(A) = 0.1$

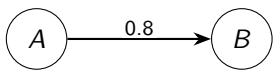
$\beta_B = 0.1$

This is mathematically opaque!

Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

Causal Bayes Net (CBN)



Stochasticity is 'baked' into the graph.
Noisy-OR parameterization:

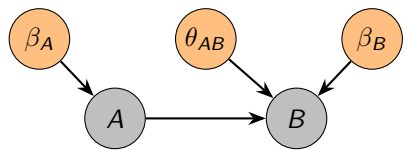
$P(B = 1 \mid A; \beta_B) = 1 - (1 - \beta_B)(1 - 0.8)^A$

$P(A) = 0.1$

$\beta_B = 0.1$

This is mathematically opaque!

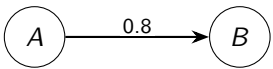
Structural Causal Model (SCM)



Modeling counterfactual order effects

Causal Models for $A \rightarrow B$

Causal Bayes Net (CBN)



Stochasticity is 'baked' into the graph.
Noisy-OR parameterization:

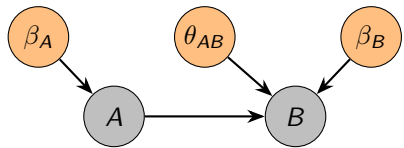
$P(B = 1 \mid A; \beta_B) = 1 - (1 - \beta_B)(1 - 0.8)^A$

$P(A) = 0.1$

$\beta_B = 0.1$

This is mathematically opaque!

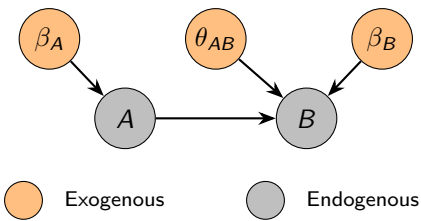
Structural Causal Model (SCM)



Stochasticity in the exogenous variables
Exogenous (stochastic priors)
Endogenous (deterministic equations)
Much more modular and explicit about the source of the randomness!

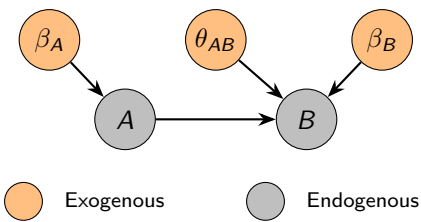
Modeling Counterfactual Order Effects

Defining a SCM



Modeling Counterfactual Order Effects

Defining a SCM



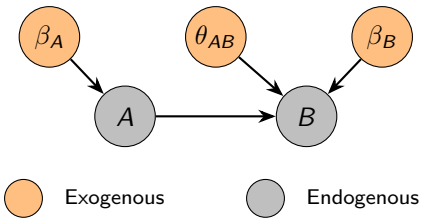
Exogenous:

- β_i: **Base rate** — probability that variable i activates spontaneously (without any cause)
- θ_{ij}: **Causal power** — probability that cause i successfully produces effect j when i is active

$$\beta_A, \beta_B \sim \text{Bernoulli}(0.1)$$
$$\theta_{AB} \sim \text{Bernoulli}(0.8)$$

Modeling Counterfactual Order Effects

Defining a SCM



Exogenous:

- β_i: **Base rate** — probability that variable i activates spontaneously (without any cause)
- θ_{ij}: **Causal power** — probability that cause i successfully produces effect j when i is active

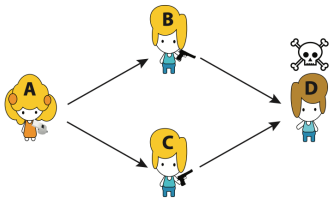
$$\beta_A, \beta_B \sim \text{Bernoulli}(0.1)$$
$$\theta_{AB} \sim \text{Bernoulli}(0.8)$$

Endogenous:

$$A := \beta_A$$
$$B := \beta_B \vee (A \wedge \theta_{AB})$$

Modeling Counterfactual Order Effects

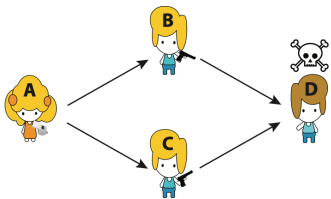
Representing the firing squad



(a) What actually happened

Modeling Counterfactual Order Effects

Representing the firing squad

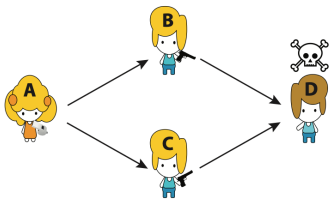


(a) What actually happened

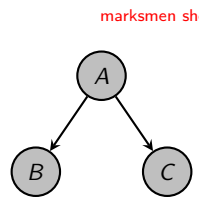
commander sends order

Modeling Counterfactual Order Effects

Representing the firing squad

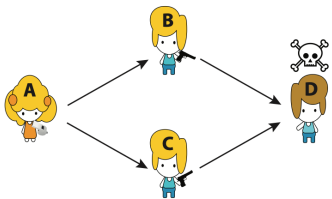


(a) What actually happened



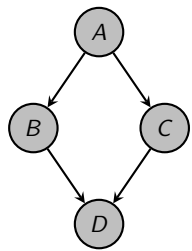
Modeling Counterfactual Order Effects

Representing the firing squad



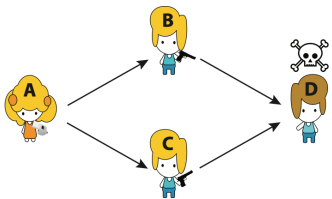
(a) What actually happened

prisoner dies

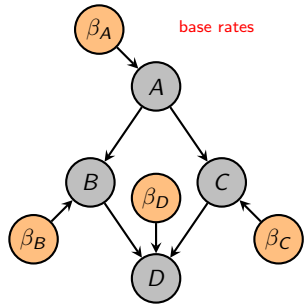


Modeling Counterfactual Order Effects

Representing the firing squad

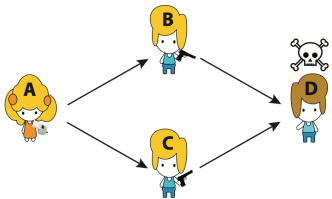


(a) What actually happened

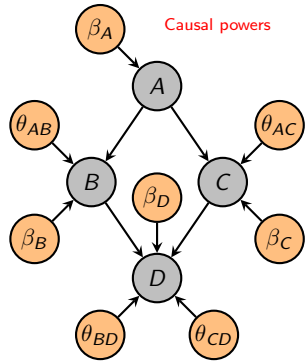


Modeling Counterfactual Order Effects

Representing the firing squad



(a) What actually happened



Modeling counterfactual order effects

Model breakdown

Part 1 Representation ✓

Part 2 Counterfactual semantics

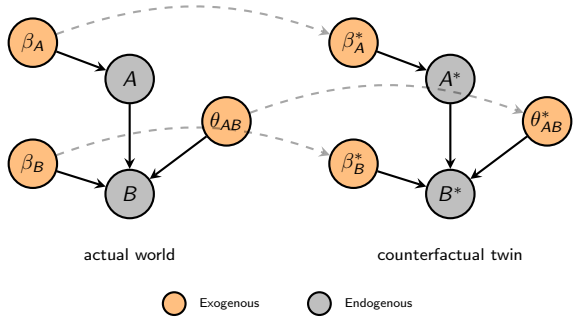
Part 3 Sequential inference

Counterfactual inference as probabilistic inference over a counterfactual twin world:

Modeling Counterfactual Order Effects

Counterfactual semantics

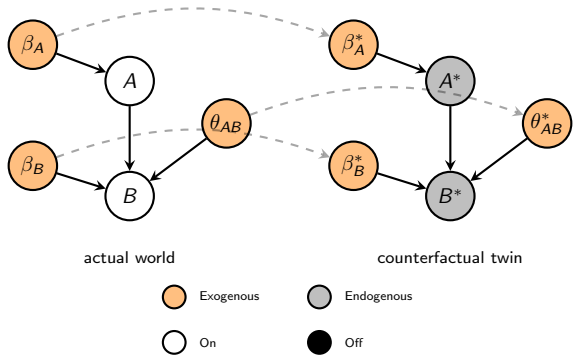
Counterfactual inference as probabilistic inference over a counterfactual twin world:



Modeling Counterfactual Order Effects

Counterfactual semantics

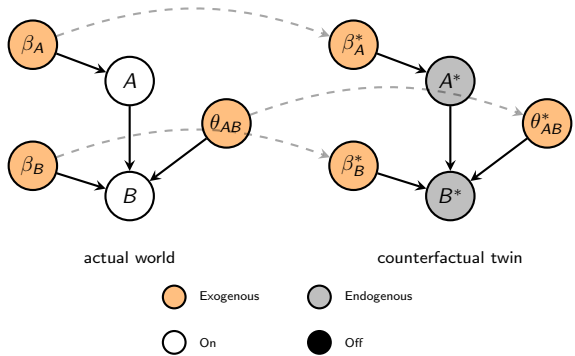
Suppose we observe A and B to be on:



Modeling Counterfactual Order Effects

Counterfactual semantics

Suppose we observe A and B to be on:

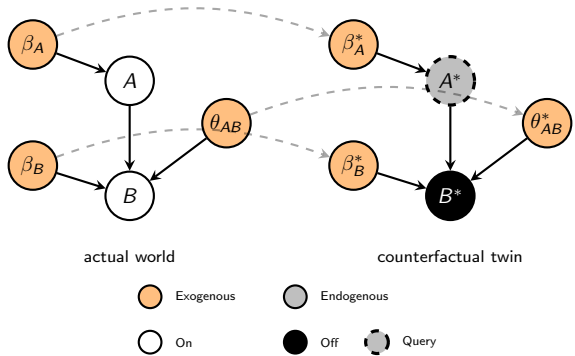


If B had been off, would A be on?

Modeling Counterfactual Order Effects

Counterfactual semantics

Suppose we observe A and B to be on:

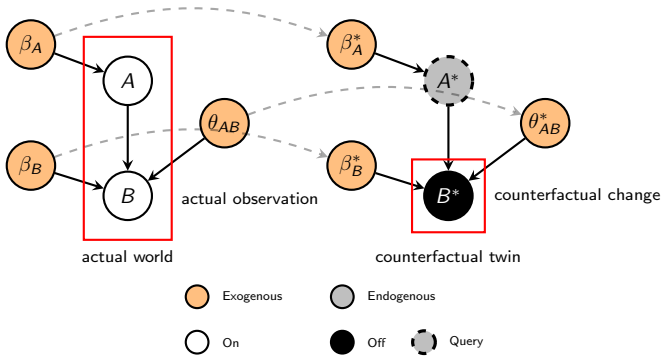


If B had been off, would A be on?

Modeling Counterfactual Order Effects

Counterfactual semantics

Suppose we observe A and B to be on:



If B had been off, would A be on?

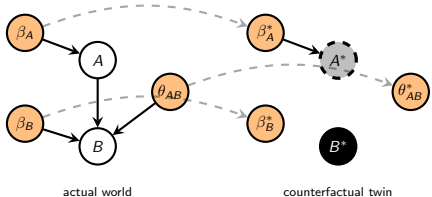
Modeling Counterfactual Order Effects

To backtrack or not to backtrack?

Modeling Counterfactual Order Effects

To backtrack or not to backtrack?

Interventional

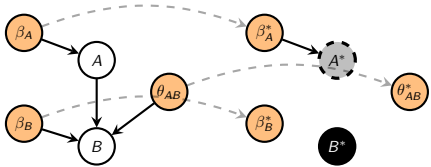


A^* determined uniquely by β_A^*.

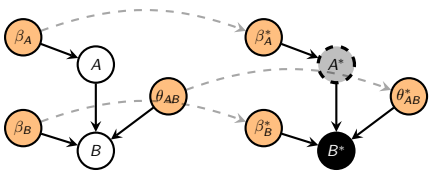
Modeling Counterfactual Order Effects

To backtrack or not to backtrack?

Interventional



Backtracking

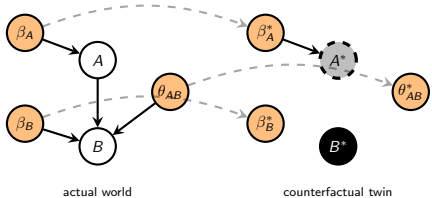


A^* determined uniquely by β^*_A.

Modeling Counterfactual Order Effects

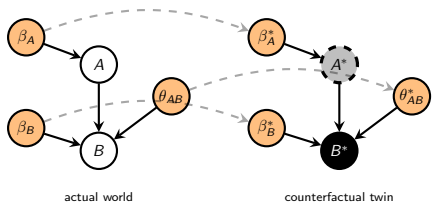
To backtrack or not to backtrack?

Interventional



A^* determined uniquely by β_A^*.

Backtracking

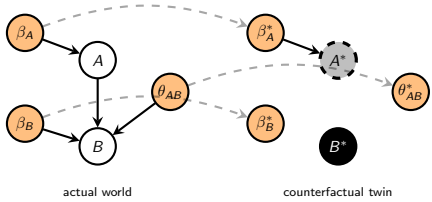


A^* inferred via several paths.

Modeling Counterfactual Order Effects

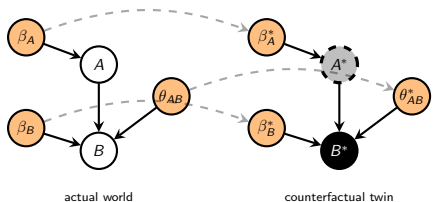
To backtrack or not to backtrack?

Interventional



A^* determined uniquely by β_A^*.

Backtracking



A^* inferred via several paths.

Exogenous distribution by $P(\mathbf{U}, \mathbf{U}^*) = P(\mathbf{U}^*|\mathbf{U})P(\mathbf{U})$.

$$P(U_i^*|U_i) = s\delta(U_i) + (1-s)P(U_i)$$

Stability s mediates the actual world observation from the priors (Lucas and Kemp, 2015; von Kügelgen et al., 2023).

Modeling counterfactual order effects

Model breakdown

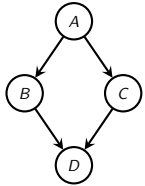
Part 1 Representation ✓

Part 2 Counterfactual semantics ✓

Part 3 Sequential inference

Modeling counterfactual order effects

Ordering queries



actual world

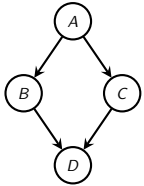
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?

Modeling counterfactual order effects

Ordering queries

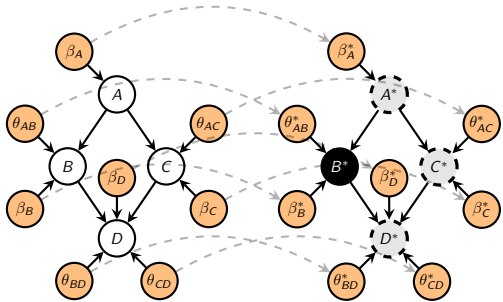


actual world

If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

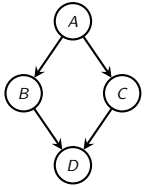
ACD would A have happened? would C have happened? would D have happened?



actual world counterfactual twin

Modeling counterfactual order effects

Ordering queries

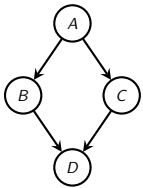


actual world

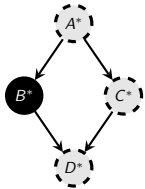
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?



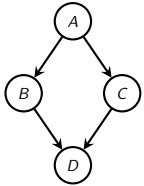
actual world



counterfactual twin

Modeling counterfactual order effects

Ordering queries

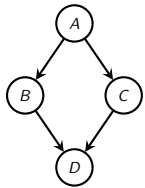


actual world

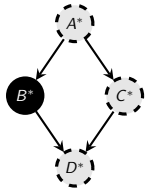
If B had not happened,

DCA would D have happened? would C have happened? would A have happened?

ACD would A have happened? would C have happened? would D have happened?



actual world

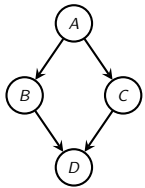


counterfactual twin

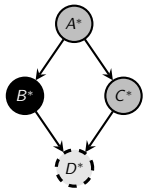
Let's consider possible responses in either order!

Modeling counterfactual order effects

Sequential inference – DCA order



actual world

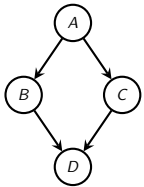


counterfactual twin

Query 1 If B had not happened, would D have happened?

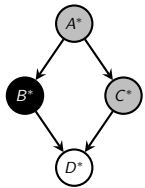
Modeling counterfactual order effects

Sequential inference – DCA order



actual world

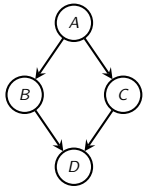
Query 1 If B had not happened, would D have happened?
Response: Yes



counterfactual twin

Modeling counterfactual order effects

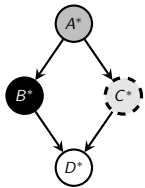
Sequential inference – DCA order



actual world

Query 1 If B had not happened, would D have happened?
Response: Yes

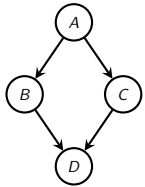
Query 2 If B had not happened, would C have happened?



counterfactual twin

Modeling counterfactual order effects

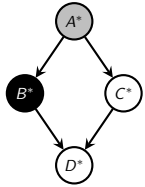
Sequential inference – DCA order



actual world

Query 1 If B had not happened, would D have happened?
Response: Yes

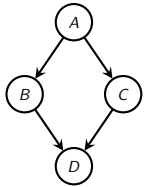
Query 2 If B had not happened, would C have happened?
Response: Yes



counterfactual twin

Modeling counterfactual order effects

Sequential inference – DCA order

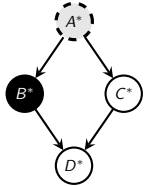


actual world

Query 1 If B had not happened, would D have happened?
Response: Yes

Query 2 If B had not happened, would C have happened?
Response: Yes

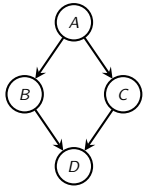
Query 3 If B had not happened, would A have happened?



counterfactual twin

Modeling counterfactual order effects

Sequential inference – DCA order

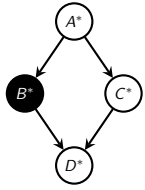


actual world

Query 1 If B had not happened, would D have happened?
Response: Yes

Query 2 If B had not happened, would C have happened?
Response: Yes

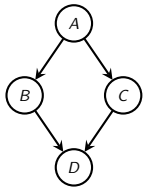
Query 3 If B had not happened, would A have happened?
Response: Yes



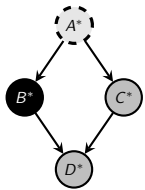
counterfactual twin

Modeling counterfactual order effects

Sequential inference – ACD order



actual world

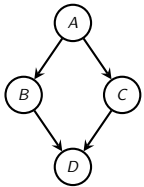


counterfactual twin

Query 1 If B had not happened, would A have happened?

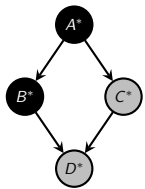
Modeling counterfactual order effects

Sequential inference – ACD order



actual world

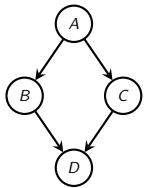
Query 1 If B had not happened, would A have happened?
Response: No



counterfactual twin

Modeling counterfactual order effects

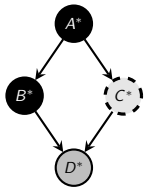
Sequential inference – ACD order



actual world

Query 1 If B had not happened, would A have happened?
Response: No

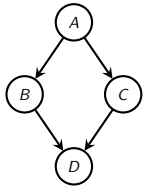
Query 2 If B had not happened, would C have happened?



counterfactual twin

Modeling counterfactual order effects

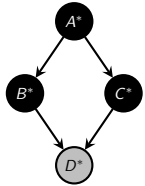
Sequential inference – ACD order



actual world

Query 1 If B had not happened, would A have happened?
Response: No

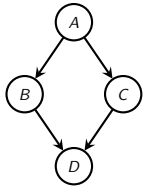
Query 2 If B had not happened, would C have happened?
Response: No



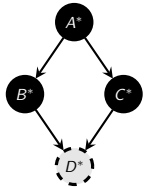
counterfactual twin

Modeling counterfactual order effects

Sequential inference – ACD order



actual world



counterfactual twin

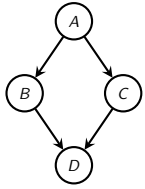
Query 1 If B had not happened, would A have happened?
Response: No

Query 2 If B had not happened, would C have happened?
Response: No

Query 3 If B had not happened, would D have happened?

Modeling counterfactual order effects

Sequential inference – ACD order

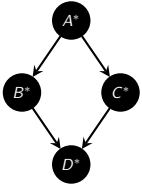


actual world

Query 1 If B had not happened, would A have happened?
Response: No

Query 2 If B had not happened, would C have happened?
Response: No

Query 3 If B had not happened, would D have happened?
Response: No

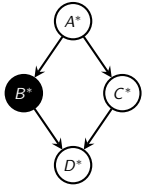


counterfactual twin

Modeling counterfactual order effects

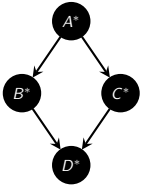
Different outcomes based on order

DCA order:



counterfactual twin

ACD order:

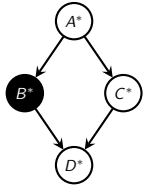


counterfactual twin

Modeling counterfactual order effects

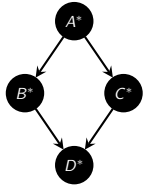
Different outcomes based on order

DCA order:



counterfactual twin

ACD order:



counterfactual twin

Why are these outcomes so different?

Modeling counterfactual order effects

The role context on counterfactual inference

Conversational Background: You observed $A = B = C = D = 1$.

- If B had been off, would D be on? Yes!
- If B had been off, would C be on? Yes!
- **If B had been off, would A be on?**

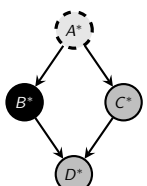
Modeling counterfactual order effects

The role context on counterfactual inference

Conversational Background: You observed $A = B = C = D = 1$.

- If B had been off, would D be on? Yes!
- If B had been off, would C be on? Yes!
- **If B had been off, would A be on?**

Contextless inference



counterfactual twin

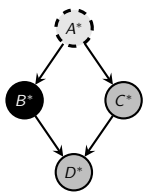
Modeling counterfactual order effects

The role context on counterfactual inference

Conversational Background: You observed $A = B = C = D = 1$.

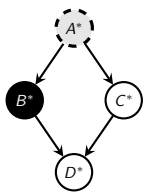
- If B had been off, would D be on? Yes!
- If B had been off, would C be on? Yes!
- **If B had been off, would A be on?**

Contextless inference



counterfactual twin

Sequential inference:



counterfactual twin

C^* and D^* left unobserved. Inference does not assume values to them.

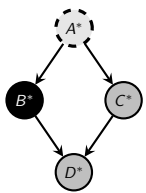
Modeling counterfactual order effects

The role context on counterfactual inference

Conversational Background: You observed $A = B = C = D = 1$.

- If B had been off, would D be on? Yes!
- If B had been off, would C be on? Yes!
- **If B had been off, would A be on?**

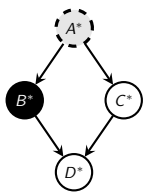
Contextless inference



counterfactual twin

C^* and D^* left unobserved. Inference does not assume values to them.

Sequential inference:



counterfactual twin

C^* and D^* are determined from the background. Inference fixes both 'on'.

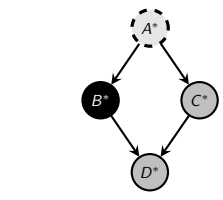
Modeling counterfactual order effects

The role context on counterfactual inference

Conversational Background: You observed $A = B = C = D = 1$.

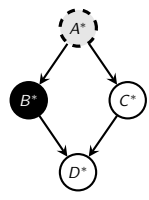
- If B had been off, would D be on? Yes!
- If B had been off, would C be on? Yes!
- **If B had been off, would A be on?**

Contextless inference



C^* and D^* left unobserved. Inference does not assume values to them.

Sequential inference:



C^* and D^* are determined from the background. Inference fixes both 'on'.

Parameter $0 \leq p_{keep} \leq 1$ mediates the extent to which previous judgments are kept.

Modeling counterfactual order effects

Model breakdown

Part 1 Representation ✓

Part 2 Counterfactual semantics ✓

Part 3 Sequential inference ✓

Model comparisons

Models to evaluate

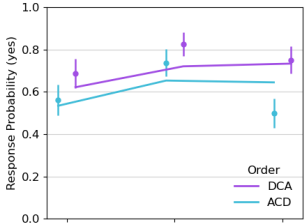
	Sequential	**No Context**
Backtrack	Backtrack-sequential	Backtrack-contextless
Interventional	Pearl-sequential	Pearl-contextless

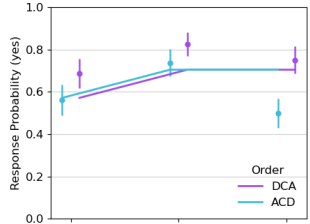
Results

Model fits to Gerstenberg et al. (2013)

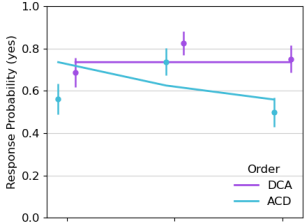
Sequential | **No Context**

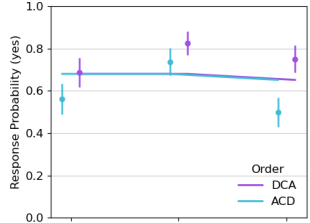
Backtrack





Interventional



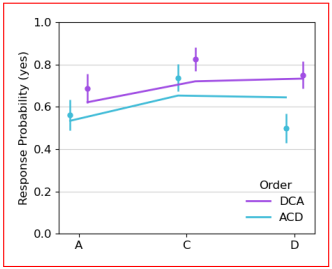


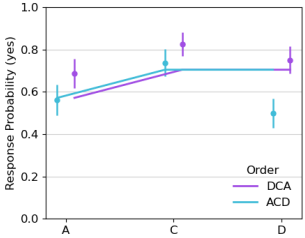
Results

Model fits to Gerstenberg et al. (2013)

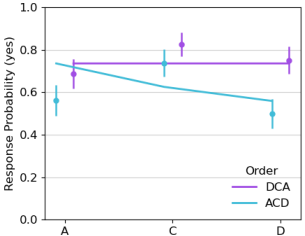
Sequential | **No Context**

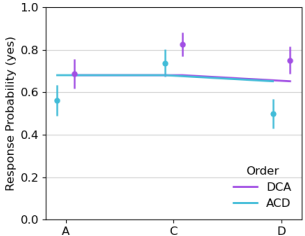
Backtrack





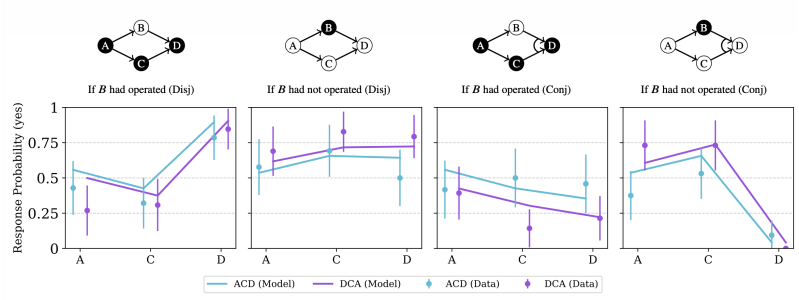
Interventional





Model comparisons

Backtracking-Sequential best fits data Gerstenberg et al. (2013)



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	–	–	–

Conclusion

Order effects in counterfactual reasoning

Two possible explanations

Order effects in counterfactual reasoning

Two possible explanations

Amortized inference:
(Gershman and Goodman, 2014)
Previous counterfactual judgments are reused in future inferences rather than recomputing states during inference.

Order effects in counterfactual reasoning

Two possible explanations

Amortized inference:
(Gershman and Goodman, 2014)
Previous counterfactual judgments are reused in future inferences rather than recomputing states during inference.

Discourse coherence:
(Roberts, 1989)
Sequential judgments try to preserve the previous judgments in the context of evaluating future utterances as a way to preserve discourse coherence.

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments
- We can rethink counterfactual interventions as inference to local latent variables

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments
- We can rethink counterfactual interventions as inference to local latent variables
- The backtracking counterfactual judgments are highly sensitive to causal beliefs

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments
- We can rethink counterfactual interventions as inference to local latent variables
- The backtracking counterfactual judgments are highly sensitive to causal beliefs

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments
- We can rethink counterfactual interventions as inference to local latent variables
- The backtracking counterfactual judgments are highly sensitive to causal beliefs

Open questions:

- Does backtracking counterfactual reasoning assist with causal structure induction?
- Is the reliance on previous judgments driven by discourse coherence or resource-rational constraints?

Conclusion

Implications:

- Psychologically plausible counterfactual changes will depend on previous judgments
- We can rethink counterfactual interventions as inference to local latent variables
- The backtracking counterfactual judgments are highly sensitive to causal beliefs

Open questions:

- Does backtracking counterfactual reasoning assist with causal structure induction?
- Is the reliance on previous judgments driven by discourse coherence or resource-rational constraints?

Thank you!

Special thanks to:

Neil Bramley

Dan Lassiter

Tobi Gerstenberg

Tadeg Quillien

References I

Morteza Dehghani, Rumen Iliev, and Stefan Kaufmann. Causal explanation and fact mutability in counterfactual reasoning. *Mind & language*, 27(1):55–85, 2012. ISSN 0268-1064.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Tobias Gerstenberg, Christos Bechlivanidis, and David A. Lagnado. Back on track: Backtracking in counterfactual reasoning. *Cognitive Science*, 35, 2013. URL `https://api.semanticscholar.org/CorpusID:6383502`.

Eric Hiddleston. A causal theory of counterfactuals. *Noûs*, 39(4):632–657, 2005. doi: 10.1111/j.0029-4624.2005.00542.x.

Christopher G Lucas and Charles Kemp. An improved probabilistic account of counterfactual reasoning. *Psychological review*, 122(4):700, 2015.

References II

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000/2009. ISBN 0-521-77362-8.

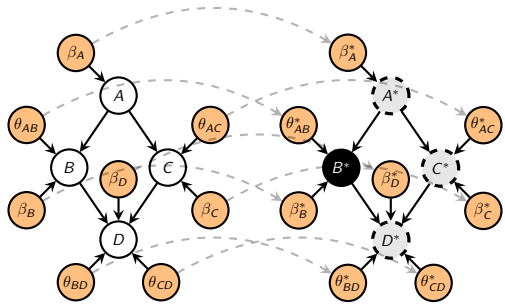
Lance J. Rips and Brian J. Edwards. Inference and explanation in counterfactual reasoning. *Cognitive science*, 37(6):1107–1135, 2013. ISSN 0364-0213.

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

Steven A. Sloman and David A. Lagnado. Do we "do"? *Cognitive science*, 29(1):5–39, 2005. ISSN 0364-0213.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals, 2023. URL `https://arxiv.org/abs/2211.00472`.

Parameter fits



Model	s	τ	P_{keep}	$\beta_B, \beta_C \beta_A$		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