# Time's Asymmetry in Counterfactuals

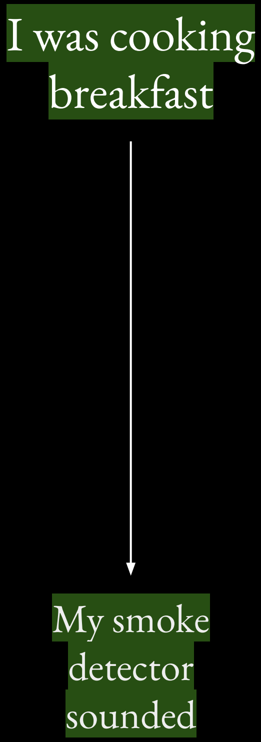
Dominic Le
Work in Collaboration with Ioana Grosu and Patricia Ganea
@ the Language and Learning Lab
University of Toronto

Confluence 2025

Firstly to motivate the research question we’ll discuss the following topics:

1. Definition of counterfactuals as conditionals.
2. Distinction between forward and backward counterfactual constructions.
3. Operationalizing causal reasoning as functional graphs.
4. Psychological literature comparing theories in counterfactual reasoning.

# Imagine this Real Course of Events:

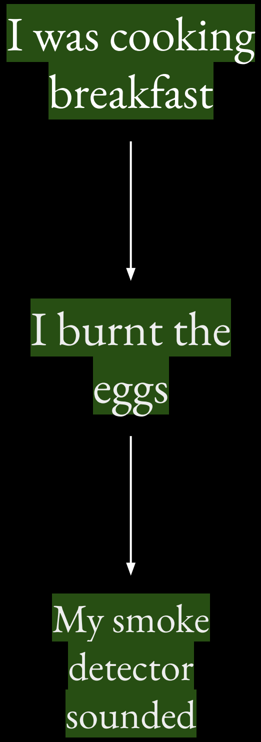


I was cooking breakfast

My smoke detector sounded

You might wonder: “Why did my smoke detector sound?”😅

# Imagine this Real Course of Events:



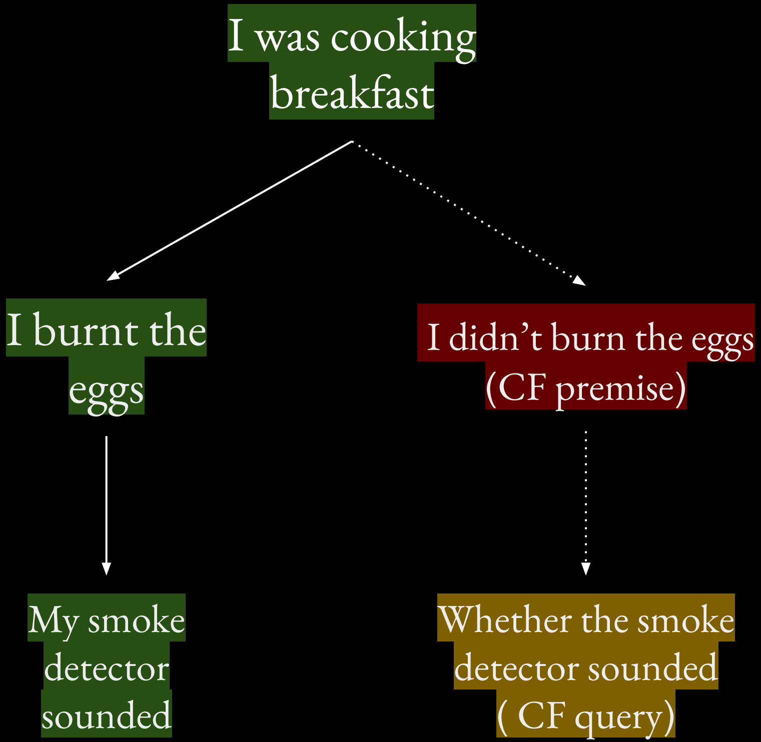
I was cooking breakfast

I burnt the eggs

My smoke detector sounded

Then, consider: “Did burning the eggs cause the smoke detector to sound?” 🤔

# Counterfactuals (CFs)



I was cooking breakfast

I burnt the eggs

I didn’t burn the eggs (CF premise)

My smoke detector sounded

Whether the smoke detector sounded ( CF query)

Definition: a hypothetical situation that includes some false *premise* that diverges from the actual course of events

E.g. “What if I didn’t burn the eggs”

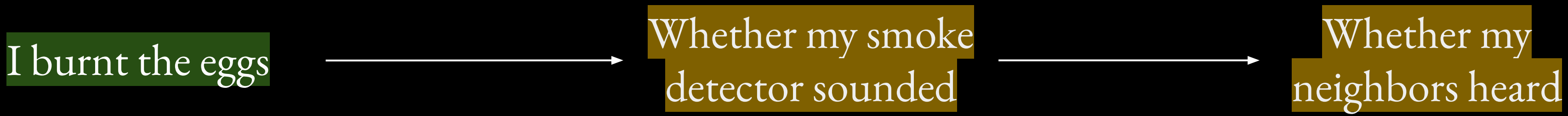
Process: evaluate the correctness of *query* events given the counterfactual premise (CF conditional).

Purpose: to understand the underlying causal mechanism behind events (i.e generate explanations and assign credit, Lucas and Kemp, 2015).

# Direction in Time (Broadbent, 2007)

Forwards

- *Premise* occurs before the *query*
- E.g "If my smoke detector had not sounded, my neighbors would not have heard it"



I burnt the eggs

Whether my smoke detector sounded

Whether my neighbors heard

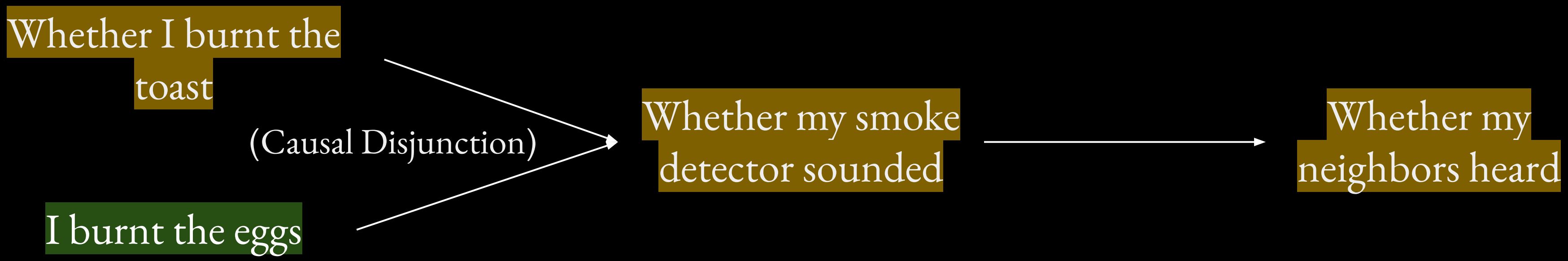
# Direction in Time (Broadbent, 2007)

Forwards

- *Premise* occurs before the *query*
- E.g "If my smoke detector had not sounded, my neighbors would not have heard it."
- Diagnoses *sufficient* causation

Backwards

- *Query* occurs before the *premise*
- E.g "If my smoke detector had not sounded, the toast would not have burnt."
- Diagnoses *necessary* causation



Whether I burnt the toast

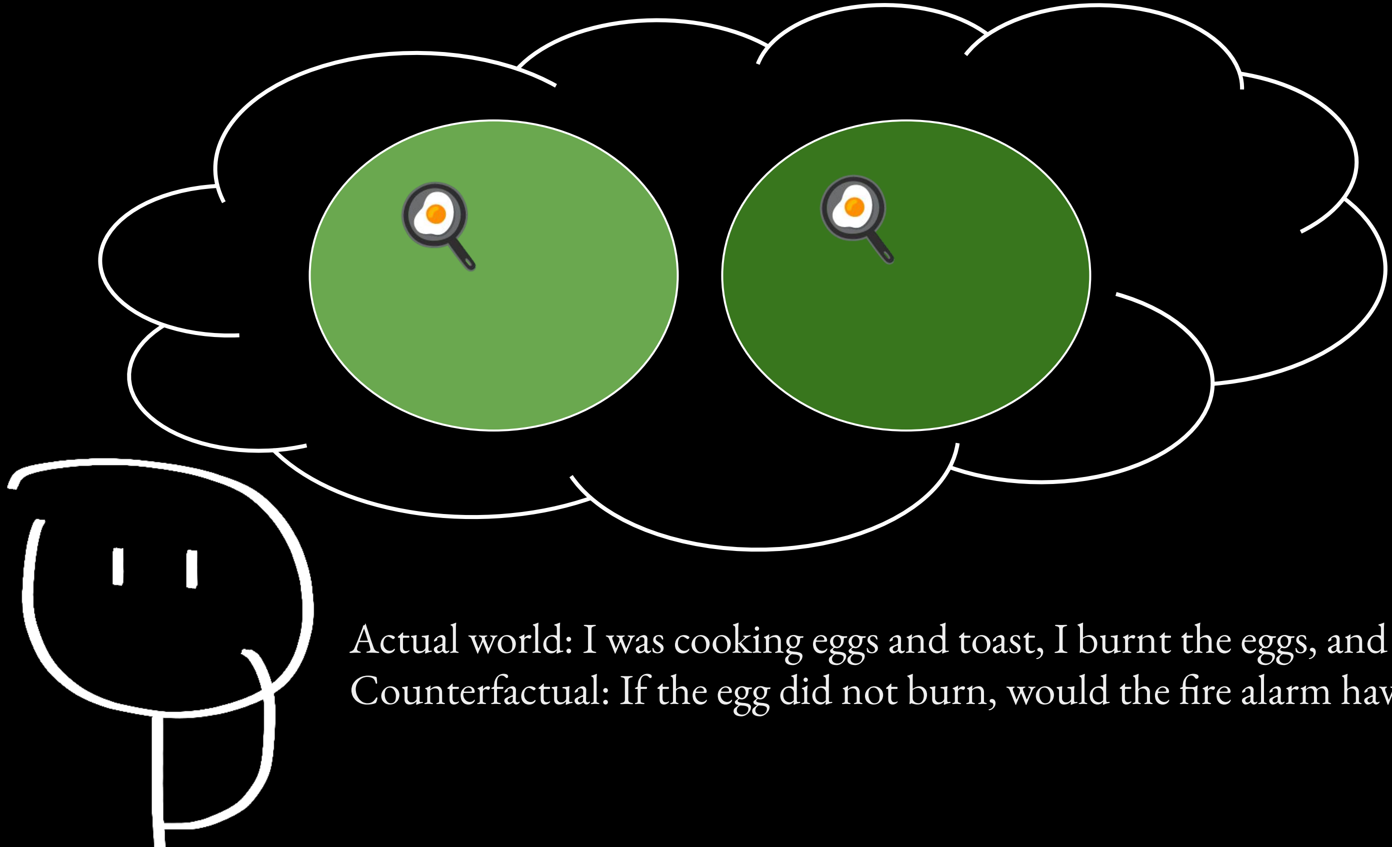
(Causal Disjunction)

I burnt the eggs

Whether my smoke detector sounded

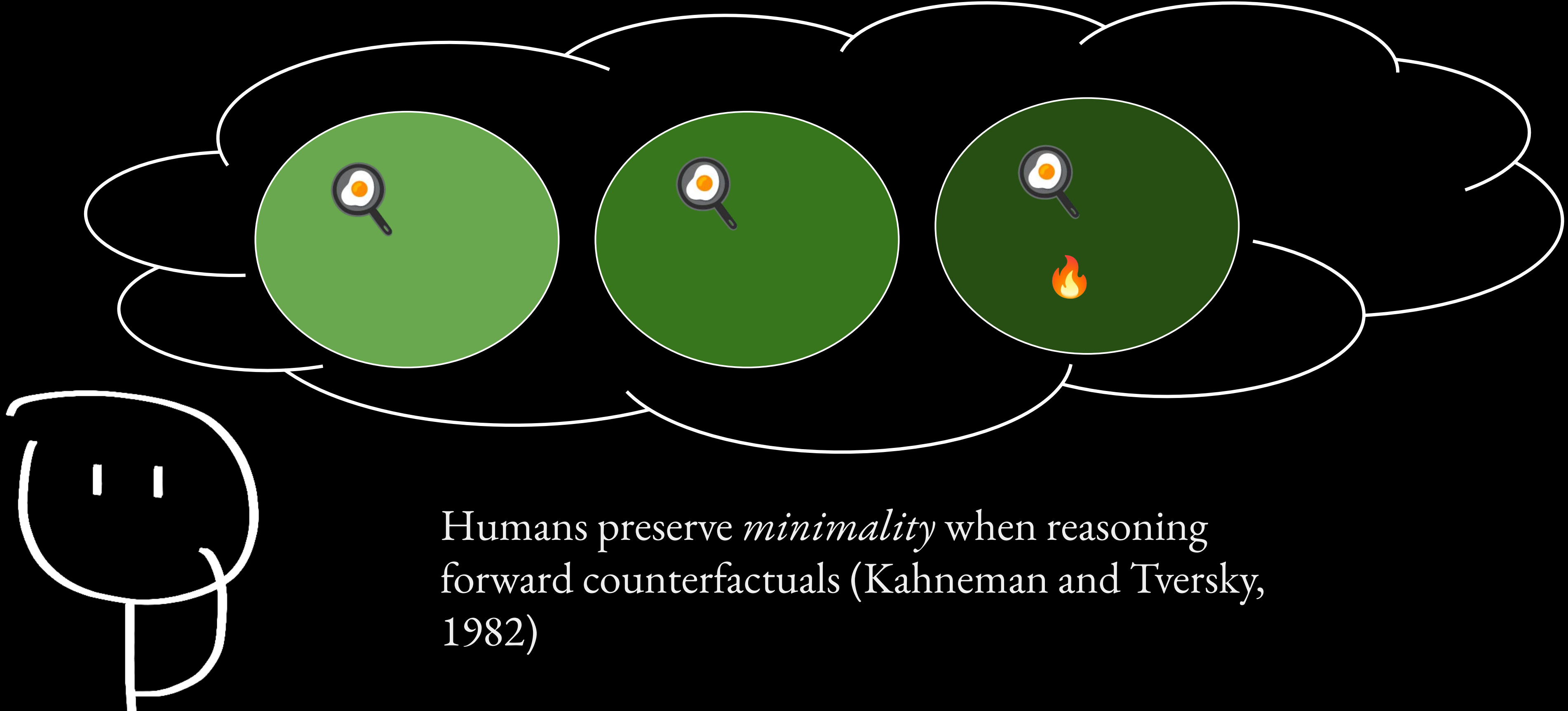
Whether my neighbors heard

# Reasoning Possible Worlds (Stalnaker, 1981; Lewis, 1979)



Actual world: I was cooking eggs and toast, I burnt the eggs, and then the fire alarm sounded!
Counterfactual: If the egg did not burn, would the fire alarm have sounded?

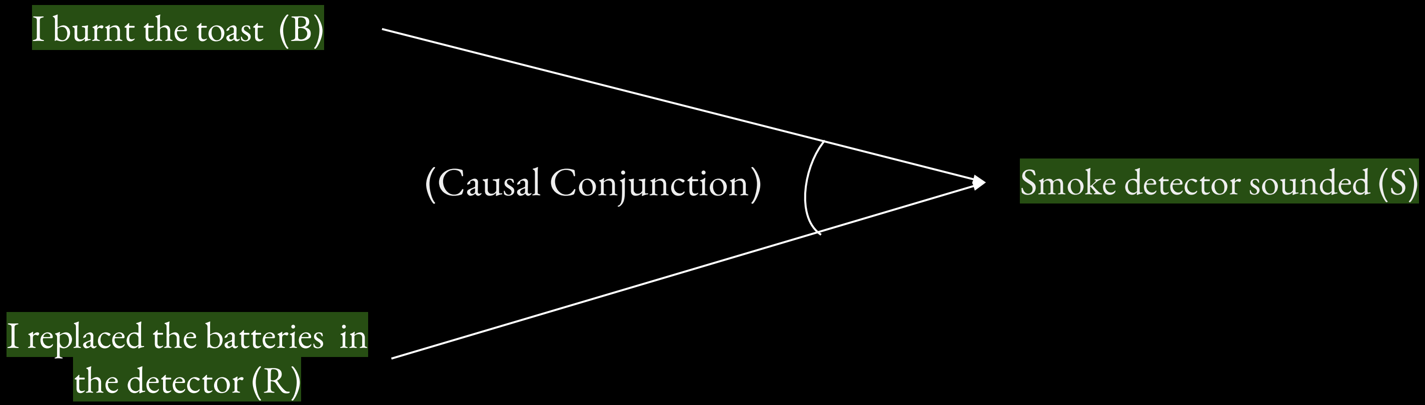
# Reasoning Possible Worlds



Humans preserve *minimality* when reasoning forward counterfactuals (Kahneman and Tversky, 1982)

# Formalizing Causal Reasoning as Functional Graphs (Danks, 2014)

- Events (X) are nodes variable between present or absent (1 or 0)
- Links (->) denote directional causal dependencies (as functions!)



I burnt the toast (B)

(Causal Conjunction)

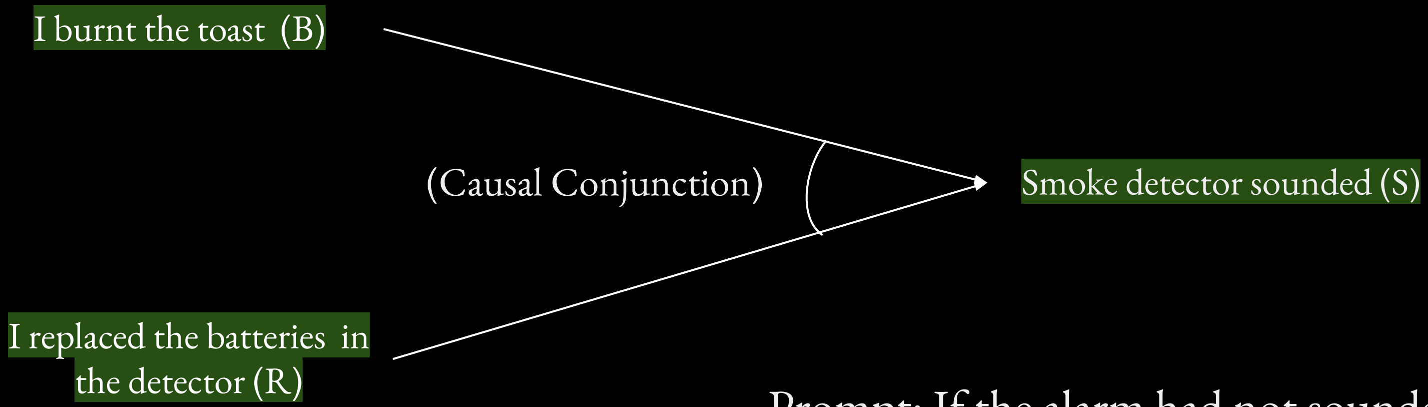
Smoke detector sounded (S)

I replaced the batteries in the detector (R)

^ New* what really happened

# Formalizing Causal Reasoning as Functional Graphs (Danks, 2014)

- Events (X) are nodes variable between present or absent (1 or 0)
- Links (->) denote directional causal dependencies (as functions!)



I burnt the toast (B)

(Causal Conjunction)

Smoke detector sounded (S)

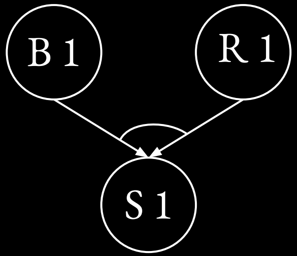
I replaced the batteries in the detector (R)

Prompt: If the alarm had not sounded (S0), would the toast have been burnt (B#)?

# Minimal Networks Theory (Hiddleston, 2005)

1. Represent the situation as a functional graph

Actual Model



S = B, when R = 1
S = R, when B = 1

^ What really happened (in symbol form)

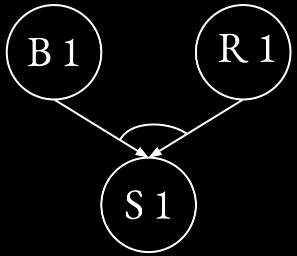
Prompt:

If the alarm had not sounded (S0), would the toast have been burnt (B#)?

# Minimal Networks Theory (Hiddleston, 2005)

1. Represent the situation as a functional graph
2. Generate Possible Worlds

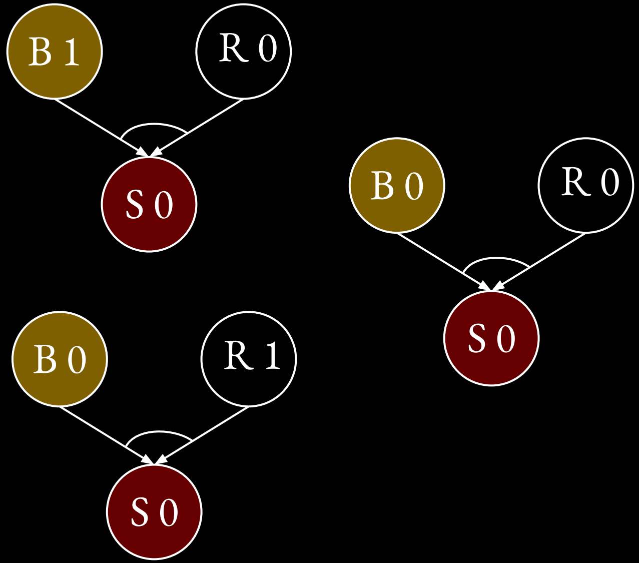
Actual Model



S = B, when R = 1
S = R, when B = 1

^ What really happened (in symbol form)

Possible Worlds



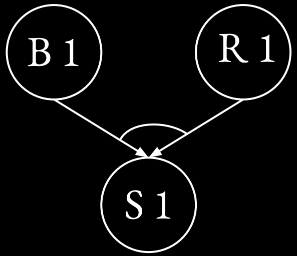
Prompt:

If the alarm had not sounded (S0), would the toast have been burnt (B#)?

# Minimal Networks Theory (Hiddleston, 2005)

1. Represent the situation as a functional graph
2. Generate Possible Worlds
3. Pick minimal world or N/A

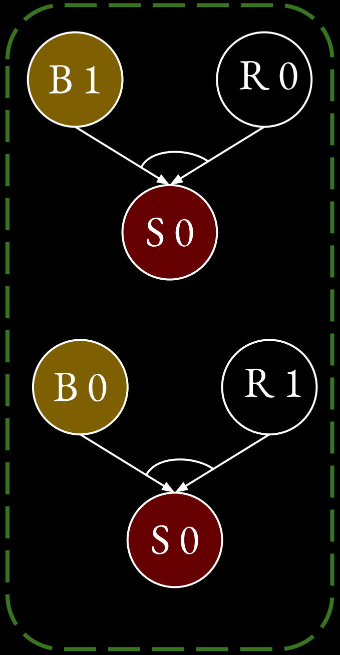
Actual Model

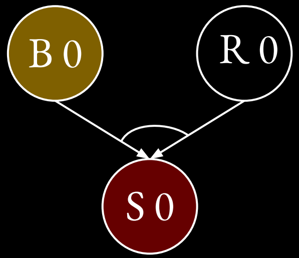


S = B, when R = 1
S = R, when B = 1

There are two minimally different worlds ->

Possible Worlds





Prompt:

If the alarm had not sounded (S0), would the toast have been burnt (B#)?

# Previous Psychological Data

Participants’ mean “yes” responses were at chance, 55.3 ± 7.2 % per 1 SD (Rips, 2010).

So Rips theorized that people sample between the two models.

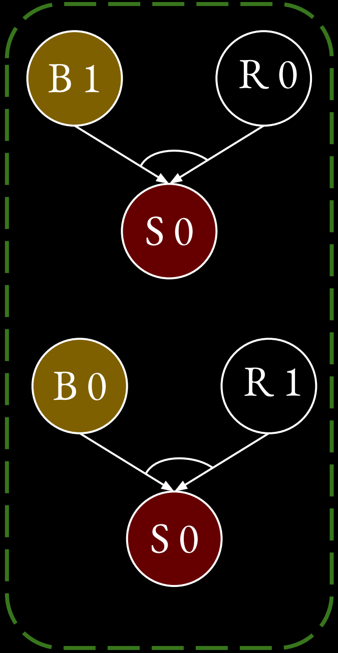
Lucas and Kemp (2015) compared their model that guessed between the two Minimal Worlds in a similar conjunction, and found above chance “yes” (p < .001).

Regardless, this type of question framing does not explain which mental model participant utilize.

Prompt:

If the alarm had not sounded (S0), would the toast have been burnt (B#)?

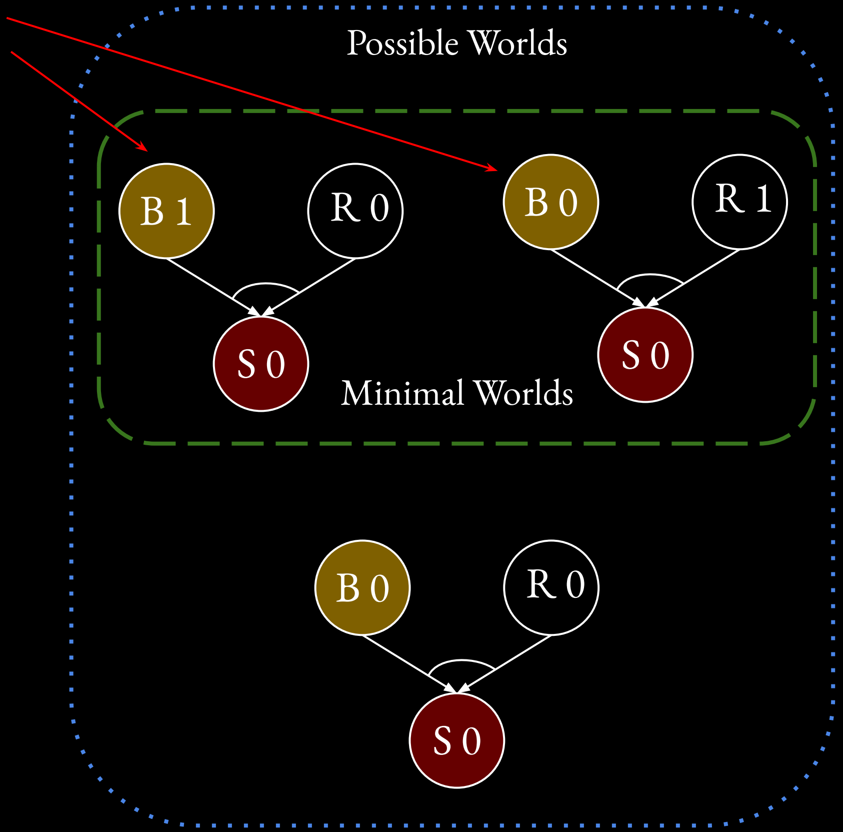
Minimal Worlds



Two minimally contradicting worlds ->

# Problem with State Query in Minimal Networks

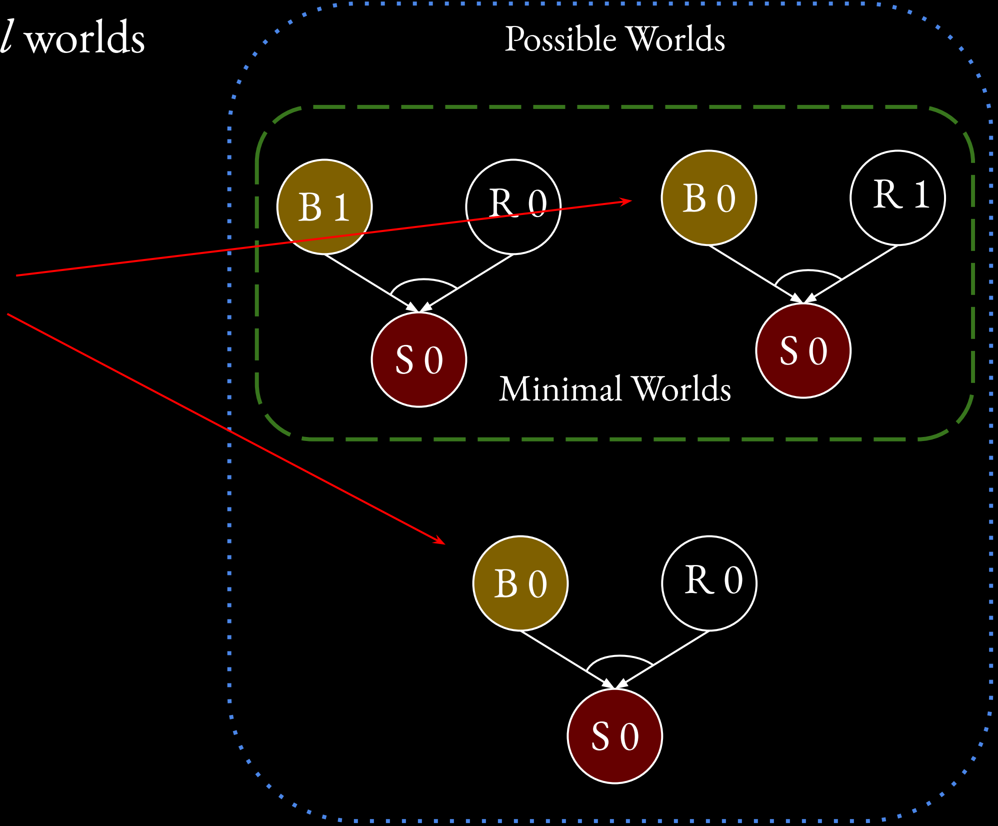
1. The state of B (=0; =1) in both *minimal* worlds contradict each other.



# Problem with State Query in Minimal Networks

1. The state of B (=0; =1) in both *minimal* worlds contradict each other.

2. There are multiple worlds where **B=0**, including *non*-minimal worlds.

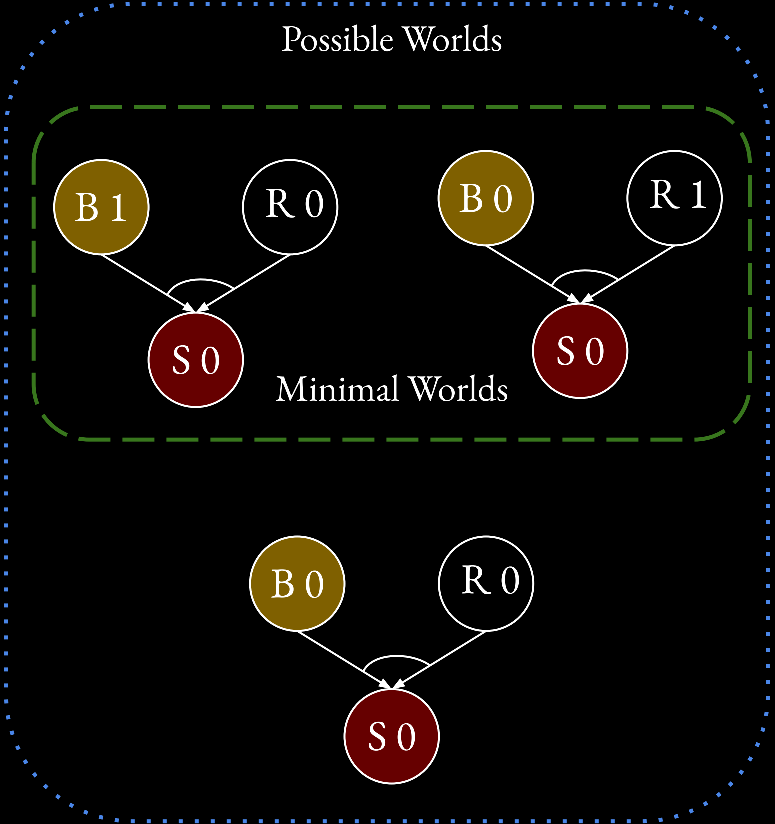


# Problem with State Query in Minimal Networks

1. The state of B (=0; =1) in both *minimal* worlds contradict each other.

2. There are multiple worlds where **B=0**, including *non*-minimal worlds.

3. Whether people believe B=0, does not indicate whether <u>they have a preference for the minimal worlds</u>, which is the central to forward CFs.



# Research Question

How does *minimality* differs in backwards counterfactuals in regards to mental models?

***Why?***

This may give us a deeper understanding what draws people to diagnose backwards CFs.

# Methods

Materials: Blicket Detector Machine,

Design: 3 x 2 repeated-measures

System conditions:

1. Disjunction
2. Conjunction
3. Single-cause with Inert

Direction conditions:

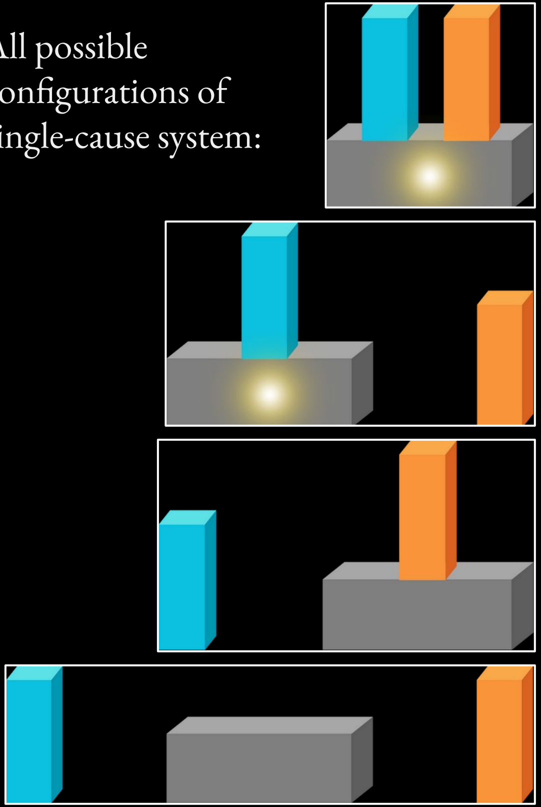
1. Forward
2. Backwards

Procedure: In random order, participants learned about each different causal system, similarly to Nyhout and Ganea (2019).

In each case, they answered control questions, e,g, “the blue block goes on the box, what does the light look?”

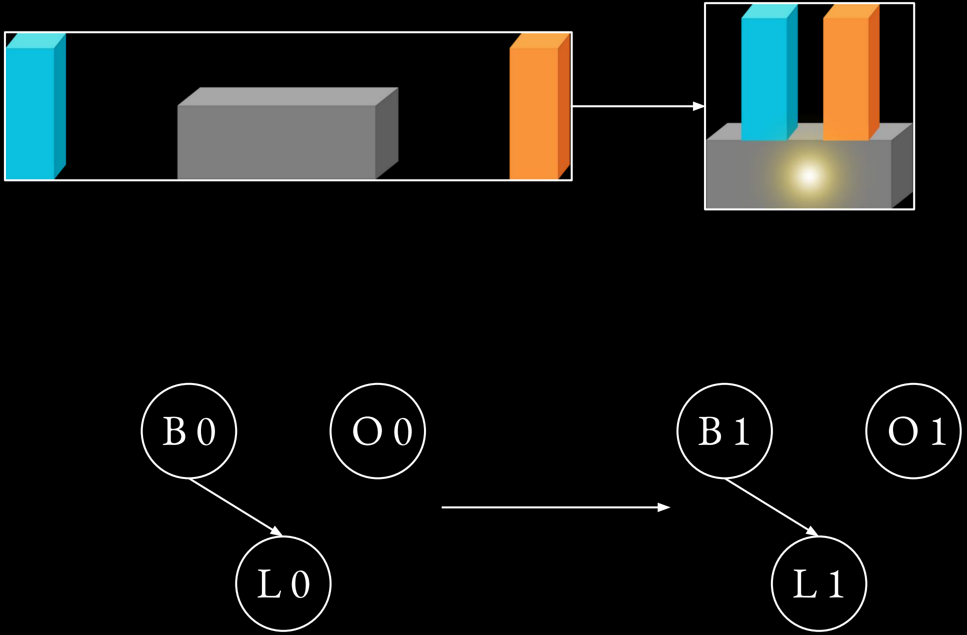
Across CF test trials, we prompted a *forced-choice response* between a minimal or a non-minimal world configuration.

All possible configurations of single-cause system:

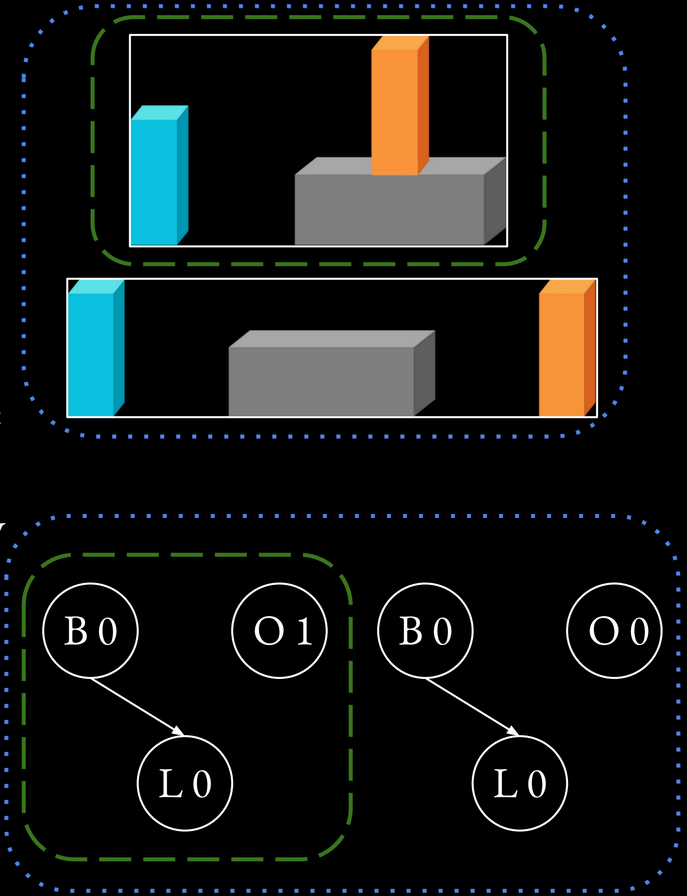


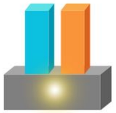
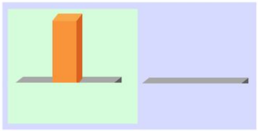
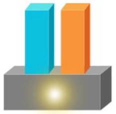
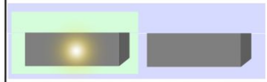
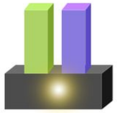
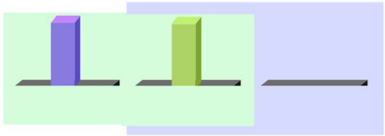
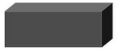
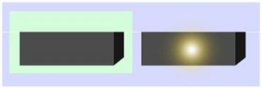
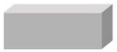
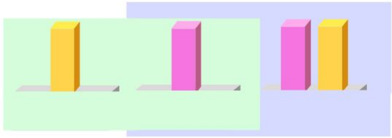
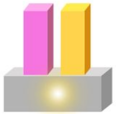
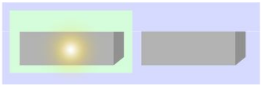
(Design and Collection in Collaboration with Ioana Gorus and Patricia Ganea at the Language and Learning Lab at OISE)

# Test Trial Sequence Example (Single-cause)



"If the light had been off, what would the blocks have likely looked like?"



|  |  | Actual World | Counterfactual | Possible Worlds |
|---|---|---|---|---|
| Single-cause |  |  |  |  |
| with Inert | Backwards |  | "If the light had been off, what would the blocks more likely have looked like?" |  |
|  | Forwards |  | "If the orange block had been off the box, what would the light more likely have looked like?" |  |
| Multi-cause |  |  |  |  |
| Con-junction | Backwards |  | "If the light had been off, what would the blocks more likely have looked like?" |  |
|  | Forwards |  | "If the purple block had been off the box, what would the light more likely have looked like?" |  |
| Dis-junction | Backwards |  | "If the light had been on, what would the blocks more likely have looked like?" |  |
|  | Forwards |  | "If the yellow block had been off the box, what would the light more likely have looked like?" |  |

🟪 = Given choices 🟩 = Closest possible worlds ▣ = Closest possible choice

*Blocks that were not on the box were shown to the side in real materials, not shown to save space.

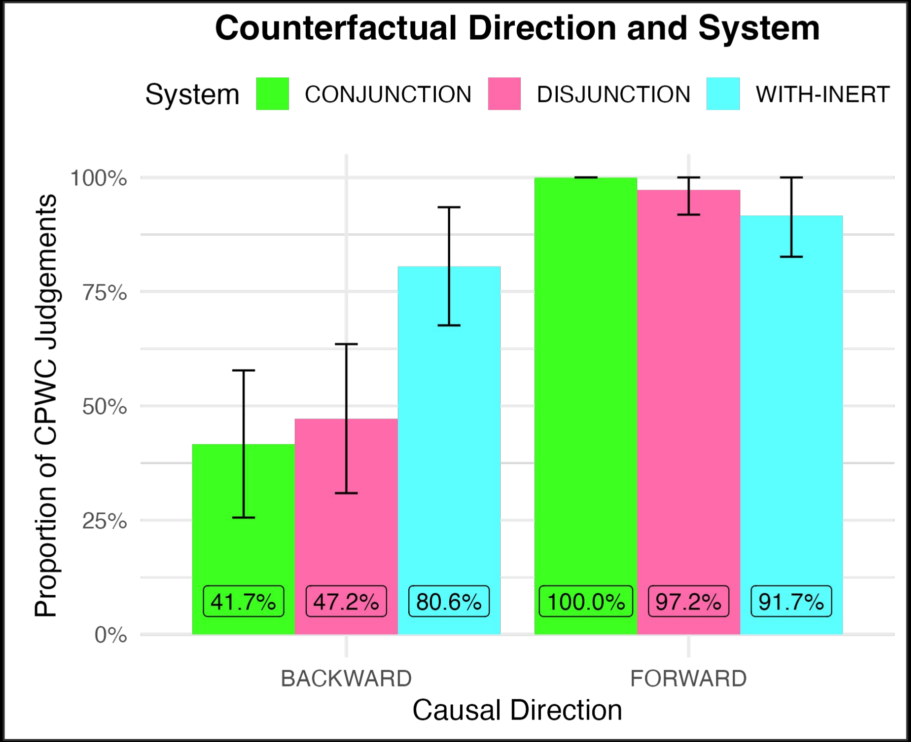
# Results

Participants: 46 adult English speakers from the U.S via Amazon Mechanical Turk, excluded 10 from analysis failing control questions.

In both multi-cause, backward conditions, participants' preference for the minimal world did not differ from chance (p = 1.0).

In the single-cause, backward condition, participants' preference was significantly above chance (p<.001)

Responses across forward conditions were all at ceiling for minimal world answers.



# Discussion & Conclusions

Interpretation of Findings:

- Potential task effects
- Evidence for necessary/sufficient causation

Limitations:

- Low sample size
- Lack of pre-existing data of backward CFs
- Lack of existing data on this methodology
- Lack of Scenarios

Future Directions:

- Ordering effects
- Causal Powers

Overview of Topics:

- Definition of counterfactuals as conditionals.
- Distinction between forward and backward counterfactual constructions.
- Operationalizing causal reasoning as functional graphs.
- Psychological literature comparing theories in counterfactual reasoning.

# References

Broadbent, A. (2007). *A reverse counterfactual analysis of causation*. http://www.dspace.cam.ac.uk/handle/1810/226170

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. the MIT press.

Hiddleston, E. (2005). A Causal Theory of Counterfactuals. *Noûs*, *39*(4), 632–657. https://doi.org/10.1111/j.0029-4624.2005.00542.x

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In A. Tversky, D. Kahneman, & P. Slovic (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 201–208). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.015

Lewis, D. (1979). *Counterfactual Dependence and Time's Arrow on JSTOR*. https://www.jstor.org/stable/2215339

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700–734. https://doi.org/10.1037/a0039655

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition*, *183*, 57–66. https://doi.org/10.1016/j.cognition.2018.10.027

Rips, L. (2010). *Two Causal Theories of Counterfactual Conditionals—Rips—2010—Cognitive Science—Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2009.01080.x

Stalnaker, R. (1968). A Theory of Conditionals. In N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98–112). Blackwell.