

## MAIN QUESTION

**When adults reason backwards over possibilities, what is the nature of the changes they allow to the actual world?**

Do they reason over only the possibilities *closest* to the actual world?

## BACKGROUND

### COUNTERFACTUALS

If the mailman had shown up, Rex would have barked.

### BACKTRACKING

Reasoning *backwards* from effect to cause.

**Effect:** If Rex had barked...

**Cause:** that means he would've seen the mailman.

### CLOSEST POSSIBILITIES

We tend to reason over counterfactual possibilities which are closest to our model of reality (following [1]).

- E.g., "If the mailman had shown up, Rex would have flown." is unlikely.

Similarity can be conceptualized with causal models - changes from actual world result in a less 'minimally-altered' model.

- Minimal networks theory (MNT):** Determine which models are minimally altered, and reason counterfactually over those [2]

### REASONING WHEN BACKTRACKING

**Testing predictions of MNT:** In [3], evidence so far is not conclusive for MNT.

**(1) Different structures:** Responses not consistent with MNT for deterministic jointly-caused structure, more MNT responses for separately-caused structure.

**(2) Forward vs. backtracking:** Similarly inconclusive MNT responses for both in a deterministic system (56%, 63% respectively).

### Some open questions:

- What worlds do participants consider when they give non-MNT responses?
- Are there causal structures for which MNT predictions consistently pan out?

## METHODS

**Materials:** Scenarios involved simple mechanical devices (blicket-detector, e.g., [4]) with three basic causal structures (CONJUNCTION, DISJUNCTION, WITH-INERT), with both FORWARD and BACKWARD conditions for each structure.

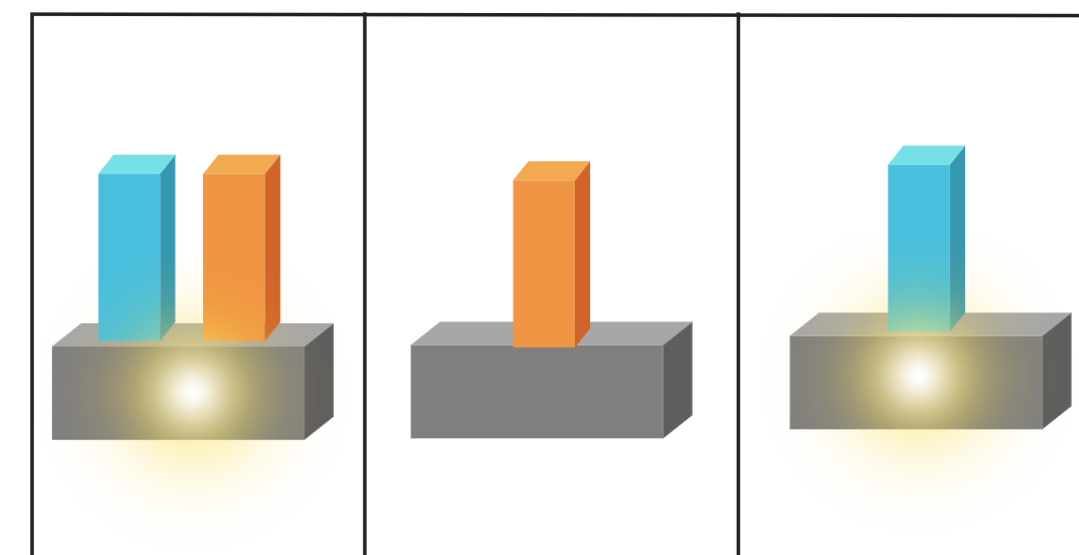
**Participants:** 46 English-speaking adults (36 analyzed, 10 excluded for incorrect responses to at least one control question).

**Design:** Within-subjects for condition.

**Procedure:** Each participant presented with each scenario, with two conditions and three control questions per scenario. Scenarios presented to participants through narrated video, in an asynchronous online survey (Qualtrics, via Amazon Mechanical Turk).

### Phase 1: Setup and sample control question(WITH-INERT)

#### Causal relations:


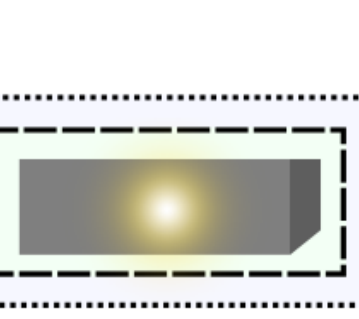
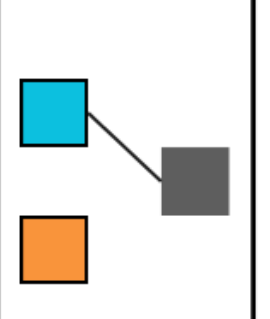
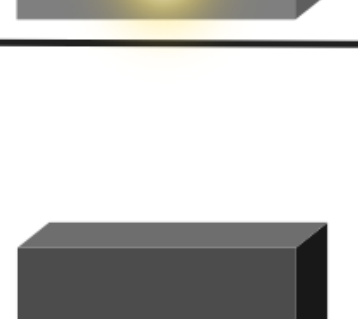
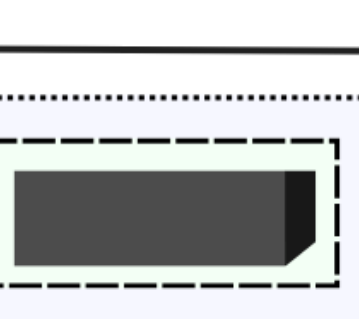
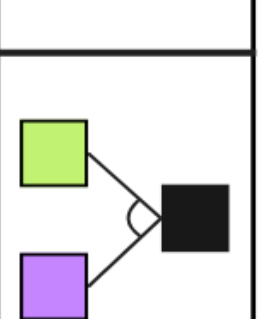
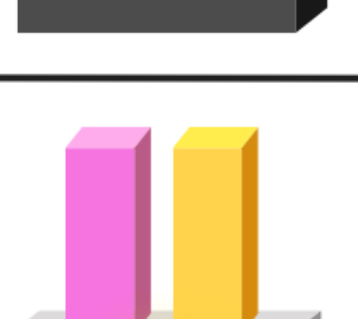
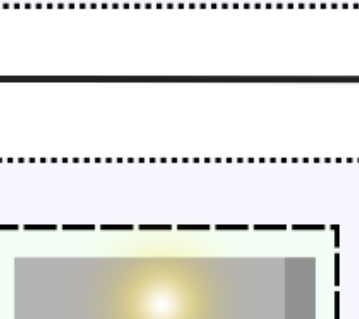
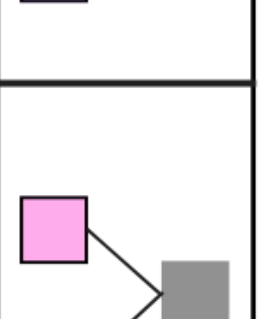
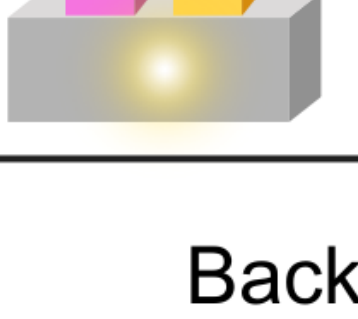
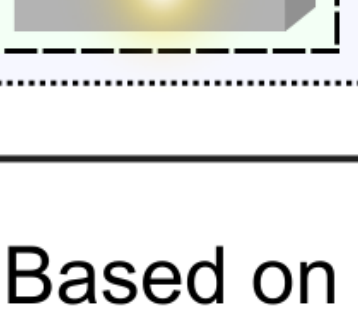
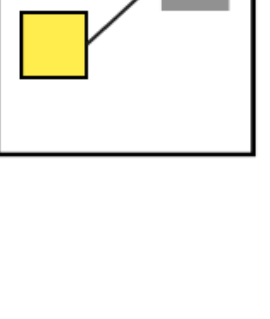


#### Sample control question:

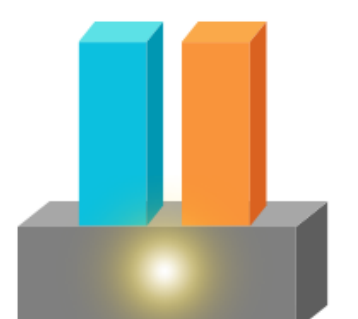
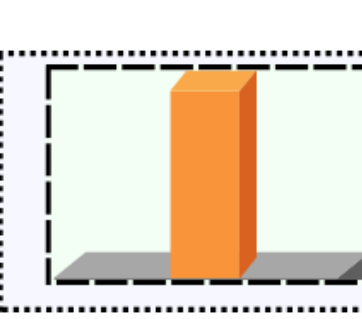
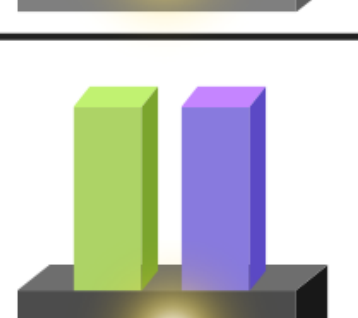



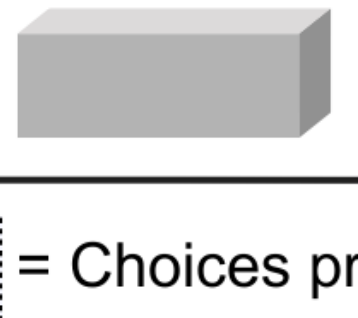



"Now, the blue block goes on the box, and the orange block stays off the box. Will the light turn on?" (yes/no)

### Phase 2: Test questions with causal models

#### Forward Testing Trial Sequences Based on System

	Actual World	Counterfactual Question	Possible Worlds	Model
SINGLE-CAUSE WITH-INERT		"If the <b>orange</b> block had been off, what would the light more likely have looked like?"		
		"If the <b>purple</b> block had been on, what would the light more likely have looked like?"		
MULTI-CAUSE CONJUNCTION		"If the <b>yellow</b> block had been off, what would the light more likely have looked like?"		
		"If the <b>yellow</b> block had been off, what would the light more likely have looked like?"		

#### Backward Testing Trial Sequences Based on System

	Actual World	Counterfactual Question	Possible Worlds
SINGLE-CAUSE WITH-INERT		"If the light had been off, what would the blocks more likely have looked like?"	
		"If the light had been off, what would the blocks more likely have looked like?"	
MULTI-CAUSE CONJUNCTION		"If the light had been off, what would the blocks more likely have looked like?"	
		"If the light had been off, what would the blocks more likely have looked like?"	
MULTI-CAUSE DISJUNCTION		"If the light had been on, what would the blocks more likely have looked like?"	
		"If the light had been on, what would the blocks more likely have looked like?"	

 = Choices presented  = Closest possible worlds  = Closest possible choice

## RESULTS

**Overall:** Participants made more closest possible world judgments in FORWARD vs. BACKWARD. ( $b = 3.138$ ,  $SE = 0.665$ ,  $Wald = 22.27$ ,  $p < .001$ , using GEE model in R).

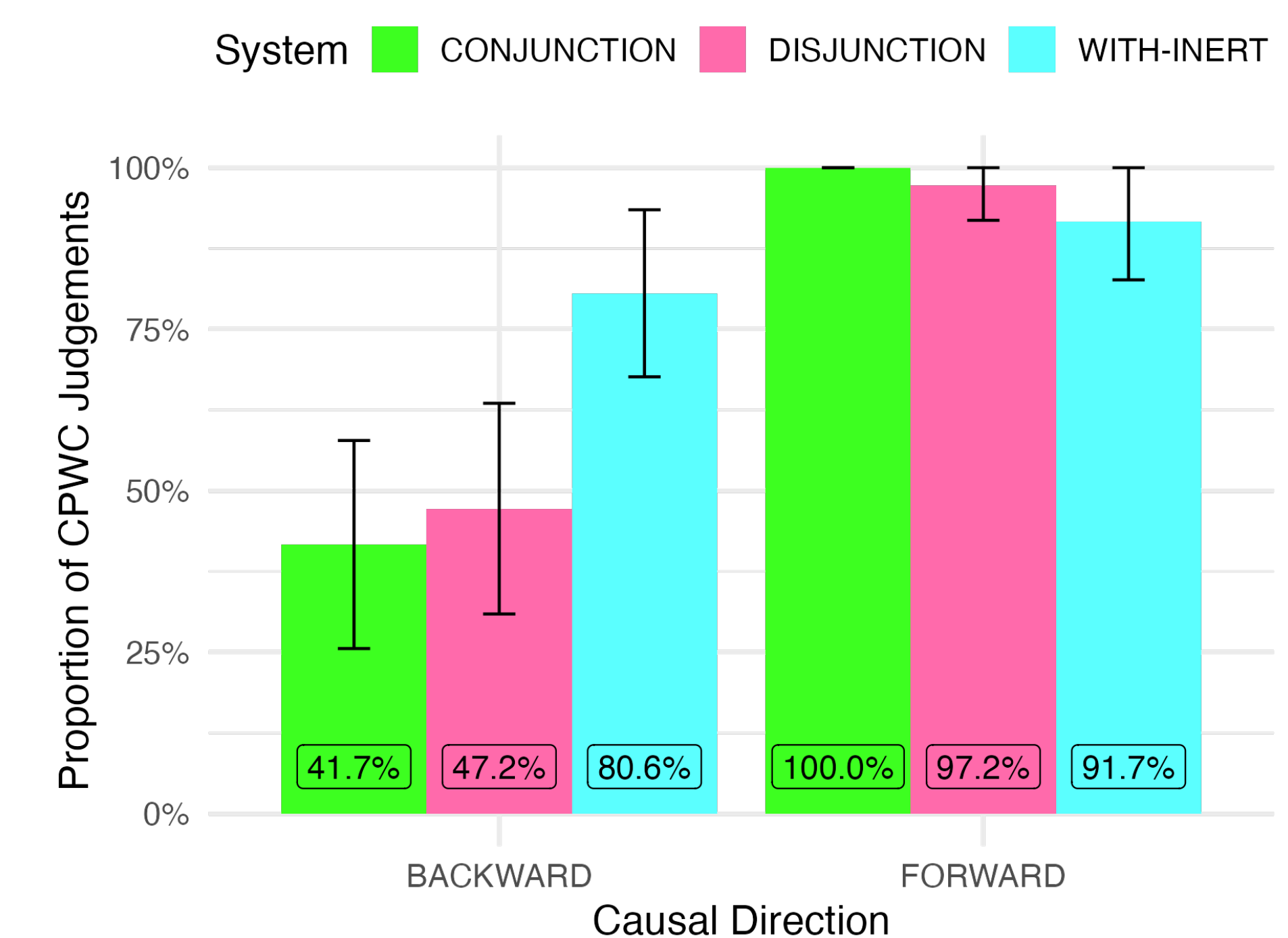
### Forward responses

- All conditions:** participants at (or near) ceiling in choosing the closest possible world.

### Backtracking responses

- WITH-INERT:** Participants consistently preferred the closest possible world (80.6%),  $p_{adjusted} < .001$ , 95%  $CI$  [0.640, 0.918]
- Multi-cause:** Preference for closest possible world did not differ from chance (DISJUNCTION (47.2%),  $p_{adjusted} = 1.00$ , 95%  $CI$  [0.304, 0.645] and CONJUNCTION (41.7%),  $p_{adjusted} = 1.00$ , 95%  $CI$  [0.255, 0.592], significance calculated using binomial tests, Bonferroni corrections).

#### Counterfactual Direction and System



## DISCUSSION

### Principal findings:

- Results align with theoretical predictions in [2] for WITH-INERT
- Difference in counterfactual direction findings unexpected in light of the final experiment in [3], where participants gave similar responses to forward and backward counterfactuals in a deterministic causal system.
- Backtracking results for single-cause scenario are also novel: adults are parsimonious in the changes they allow from the actual world in some backtracking cases.

**Conclusion:** While adults do make use of a notion of similarity when backtracking, they are not consistent in applying this reasoning strategy.

### SELECTED REFERENCES

- [1] Lewis, D. (1973). Counterfactuals and comparative possibility. *Ifs: Conditionals, belief, decision, chance and time*.  
[2] Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*. [3] Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive science*. [4] Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*.

### ACKNOWLEDGMENTS

Research supported by the Social Sciences and Humanities Research Council of Canada (SSHRC).