TOPIC 3:
COUNTERFACTUALS & CAUSAL MODELS

Dan Lassiter
Stanford Linguistics

Paris VII
December 11, 2019

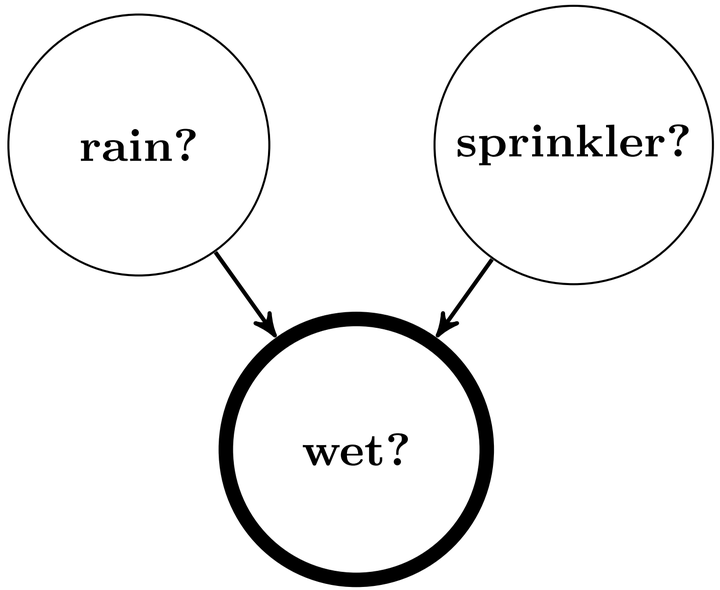
overview

- semantics for counterfactuals built on causal models
- the problem of complex antecedents
- intervention choice as explanatory reasoning
- some experimental evidence

COUNTERFACTUAL REASONING AS INTERVENTION

Causal models

Pearl, Causality (2000); Book of Why (2018)



Obs: grass is wet

Think of causal models as a general framework for knowledge representation

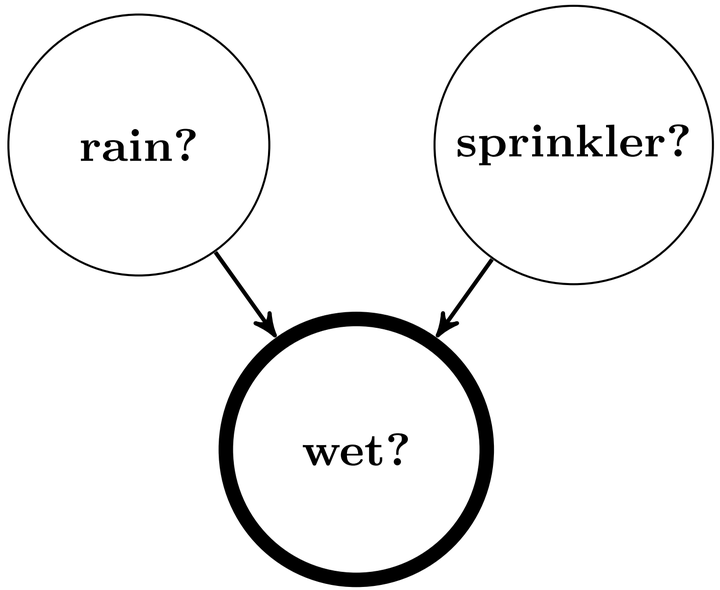
- formalization of ‘theory theory’

(Keil, Gopnik, etc)

Causal, counterfactual reasoning depend on structure of our generative models of the world

Causal models

Pearl, Causality (2000); Book of Why (2018)



Think of causal models as a general framework for knowledge representation

- formalization of ‘theory theory’ (Keil, Gopnik, etc)

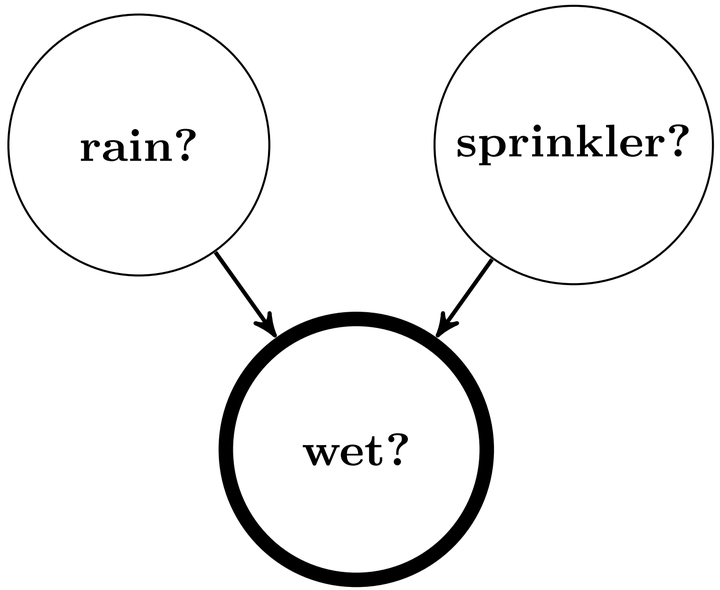
Causal, counterfactual reasoning depend on structure of our generative models of the world

Obs: grass is wet

✓ ‘If the sprinkler is on, it didn’t rain’ (conditioning)

Causal models

Pearl, Causality (2000); Book of Why (2018)



Think of causal models as a general framework for knowledge representation

- formalization of ‘theory theory’ (Keil, Gopnik, etc)

Causal, counterfactual reasoning depend on structure of our generative models of the world

Obs: grass is wet

✓ ‘If the sprinkler is on, it didn’t rain’ (conditioning)

✗ ‘If the sprinkler were on, it wouldn’t have rained’ (intervention)

Cognitive applications of causal models

many!

Intuitive physics

Intuitive theory of mind

- many more domains, e.g., mathematical learning (next lecture)

Rehder: Concepts are causal models

- Gerstenberg et al: Concepts are probabilistic programs

Danks: Causal models are the common language that allow components of a modular mind to share information

Counterfactual evaluation as intervention

Pearl, 2000

To evaluate

If A were the case, C would be the case

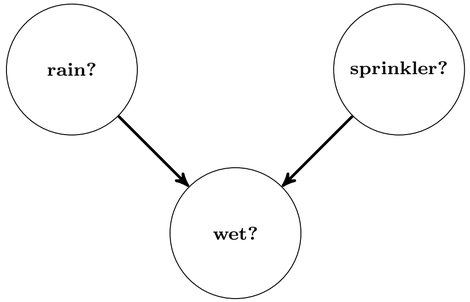
relative to causal model M, construct a model M' by intervening on M to make A true, and then check whether C is the case in M'.

Can be construed as a more explanatory version of Stalnaker 1968:

- A => C is true at w if C is true at f(w, A) for a 'selection function' f

Intervening in causal Bayes nets

Meek & Glamour 1994

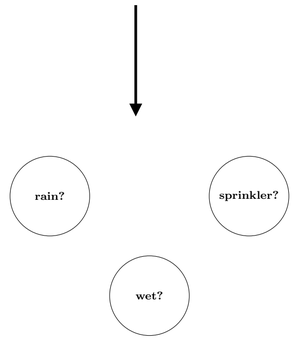


Rain ~ (rain)
Sprinkler ~ P(Sprinkler)
Wet = Rain or Sprinkler

‘If the ground were wet, …’

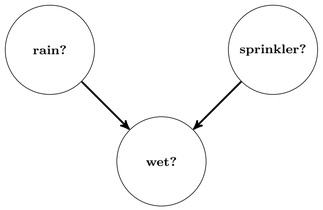
Rain ~ P(rain)
Sprinkler ~ P(Sprinkler)
Wet = True

break dependency
on parents



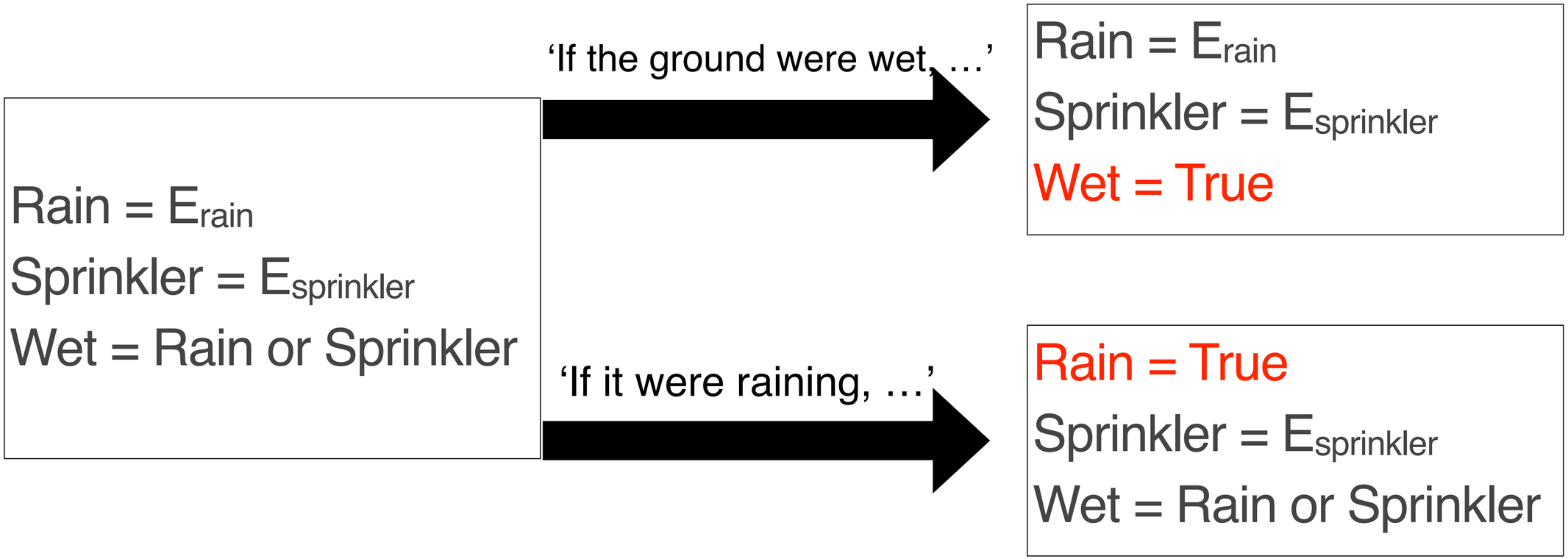
‘If it were raining, …’

Rain = True
Sprinkler ~ P(Sprinkler)
Wet = Rain or Sprinkler



Intervening in structural equation models

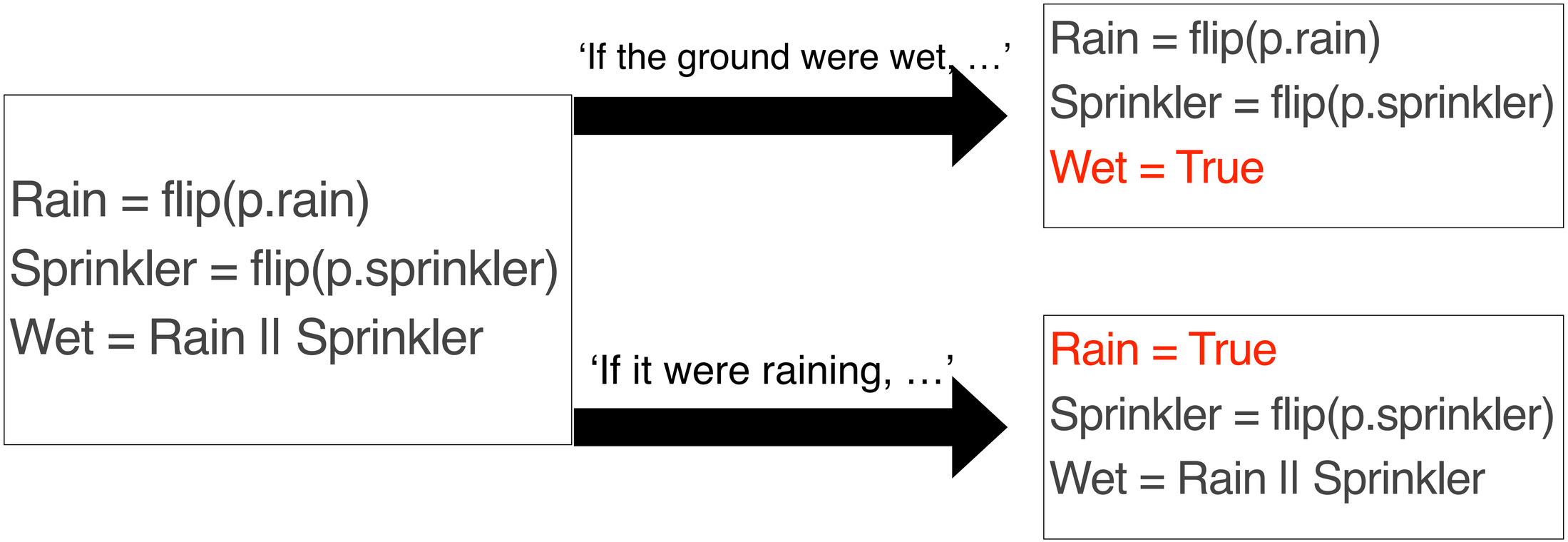
Pearl, Causality (2000)



Intervening in programs

Chater & Oaksford 2013, Icard 2017

intervention = program transformation!



A key gap: Non-binary antecedents (also disjunctions)

Negated conjunctions	If (you and I and Paolo and Dave and …) weren't all here, there would still be someone here
Negated universals	If not everyone here had come, there would still be enough people for a good conference
Certain indefinites	If I had a different kind of dog, I'd probably have a pug (but I might not)
Other non-binary	If I hadn't studied linguistics, I probably would've done philosophy

What operation are we supposed to perform?

NEGATED CONJUNCTIONS IN THE ANTECEDENT

Linguist and Philos (2018) 41:577–621
https://doi.org/10.1007/s10988-018-9232-4

Two switches in the theory of counterfactuals

A study of truth conditionality and minimal change

Ivano Ciardelli[1,3] · **Linmin Zhang**[2,4] · **Lucas Champollion**[2]

Published online: 15 June 2018

Abstract Based on a crowdsourced truth value judgment experiment, we provide empirical evidence challenging two classical views in semantics, and we develop a novel account of counterfactuals that combines ideas from inquisitive semantics and causal reasoning. First, we show that two truth-conditionally equivalent clauses can make different semantic contributions when embedded in a counterfactual antecedent. Assuming compositionality, this means that the meaning of these clauses is not fully determined by their truth conditions. This finding has a clear explanation in inquisitive semantics: truth-conditionally equivalent clauses may be associated with different propositional alternatives, each of which counts as a separate counterfactual assumption. Second, we show that our results contradict the common idea that

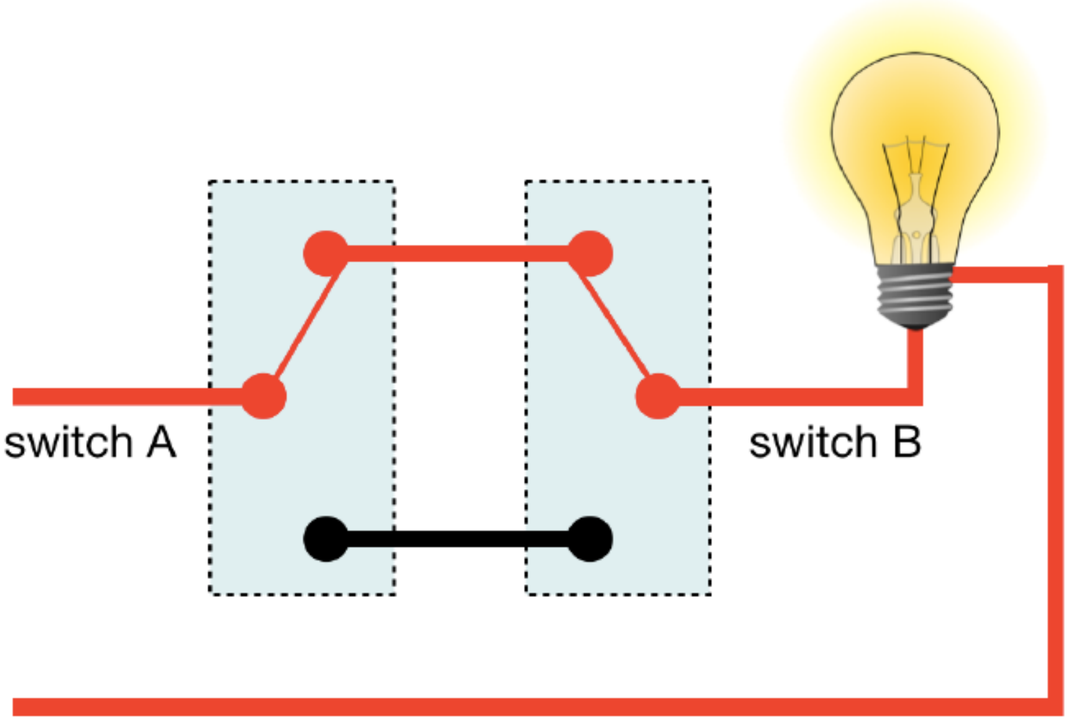


Figure 1: A multiway switch

a. If switch A was down, the light would be off. $\overline{A} > \text{OFF}$

b. If switch B was down, the light would be off. $\overline{B} > \text{OFF}$

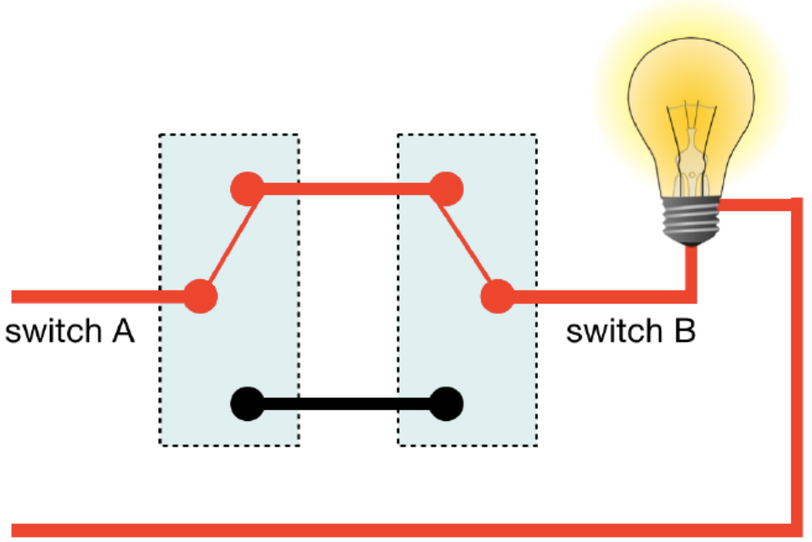
c. If switch A or switch B was down, the light would be off. $\overline{A} \vee \overline{B} > \text{OFF}$

d. If switch A and switch B were not both up, the light would be off.

$\neg(A \wedge B) > \text{OFF}$

e. If switch A and switch B were not both up, the light would be on.

$\neg(A \wedge B) > \text{ON}$



The problem

CZC ‘17

Virtually all possible-worlds theories predict this inference is valid:

- If A were not the case, C would be the case
- If B were not the case, C would be the case
- So, If A and B were not both the case, C would be the case

In CZC’s experiment, participants were much less likely to endorse the conclusion than the premises

- 22% vs. 65/66%

Table 3: Results of the main experiment

Sentence	Number	True	(%)	False	(%)	Indet.	(%)
$\overline{A} >$ OFF	256	169	66.02%	6	2.34%	81	31.64%
$\overline{B} >$ OFF	235	153	65.11%	7	2.98%	75	31.91%
$\overline{A} \vee \overline{B} >$ OFF	362	251	69.33%	14	3.87%	97	26.80%
$\neg(A \wedge B) >$ OFF	372	82	22.04%	136	36.56%	154	41.40%
$\neg(A \wedge B) >$ ON	200	43	21.50%	63	31.50%	94	47.00%

(4)
a. If switch A was down, the light would be off. $\overline{A} >$ OFF
b. If switch B was down, the light would be off. $\overline{B} >$ OFF
c. If switch A or switch B was down, the light would be off. $\overline{A} \vee \overline{B} >$ OFF
d. If switch A and switch B were not both up, the light would be off. $\neg(A \wedge B) >$ OFF
e. If switch A and switch B were not both up, the light would be on. $\neg(A \wedge B) >$ ON

Ciardelli, Zhang & Champollion 2017

Interventions in ‘Background semantics’

Ciardelli, Zhang & Champollion 2017

- Start with a causal model
- Remove contingent facts that contribute to the falsity of the antecedent, or depend causally on facts that do
- Intervene: Force the antecedent to be true
- Consider what follows logically

Effect:

What is true of all ways of making the antecedent true in the revised causal model?

cf.: Briggs 2012, Santorio 2017

Background semantics

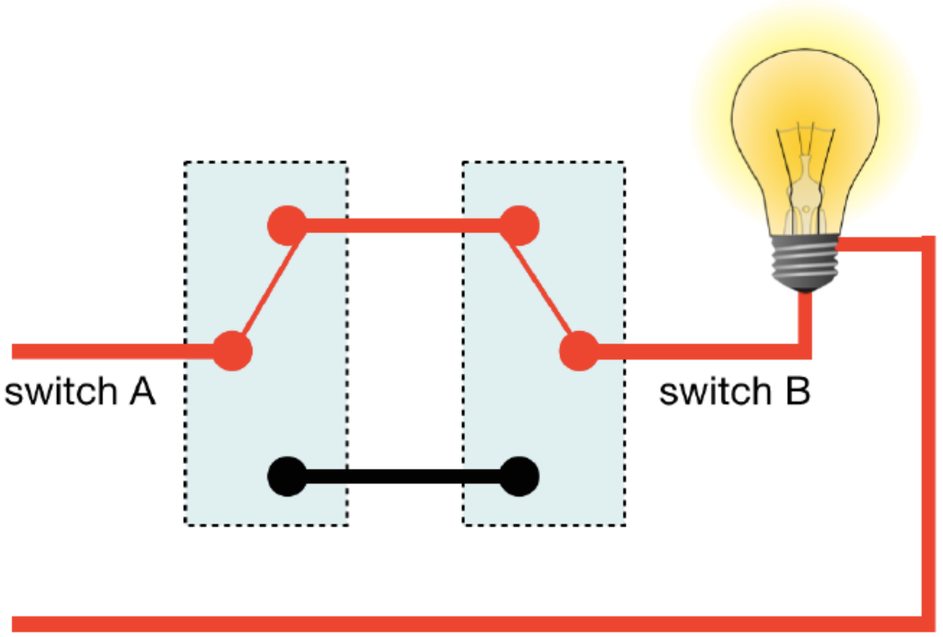
Ciardelli, Zhang & Champollion 2017

Facts:

- A is up
- B is up

Laws:

- Light is on iff A and B agree



Background semantics

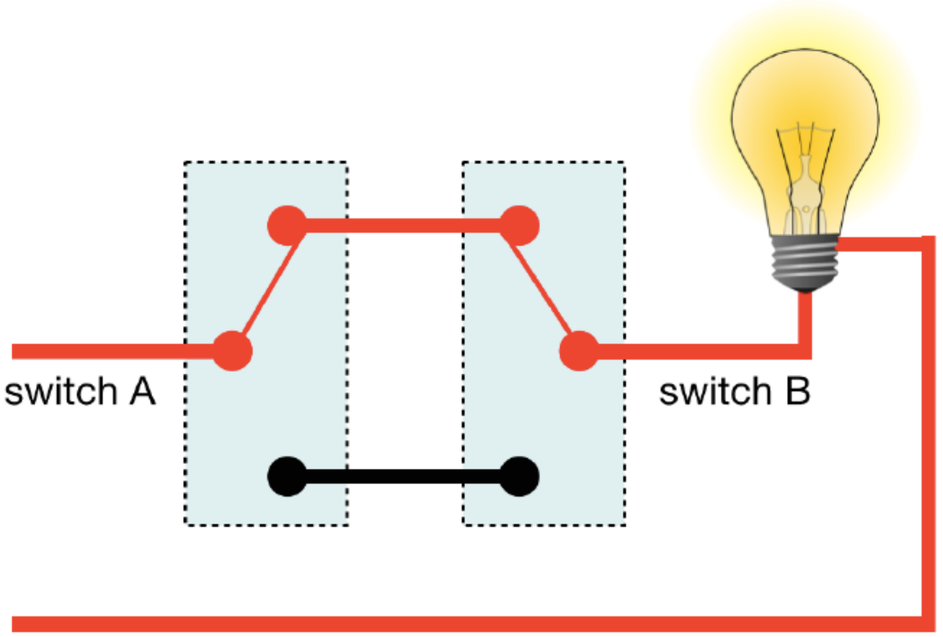
Ciardelli, Zhang & Champollion 2017

Facts:

- A is up
- B is up

Laws:

- Light is on iff A and B agree



‘If A were down, the light would be off’

Background semantics

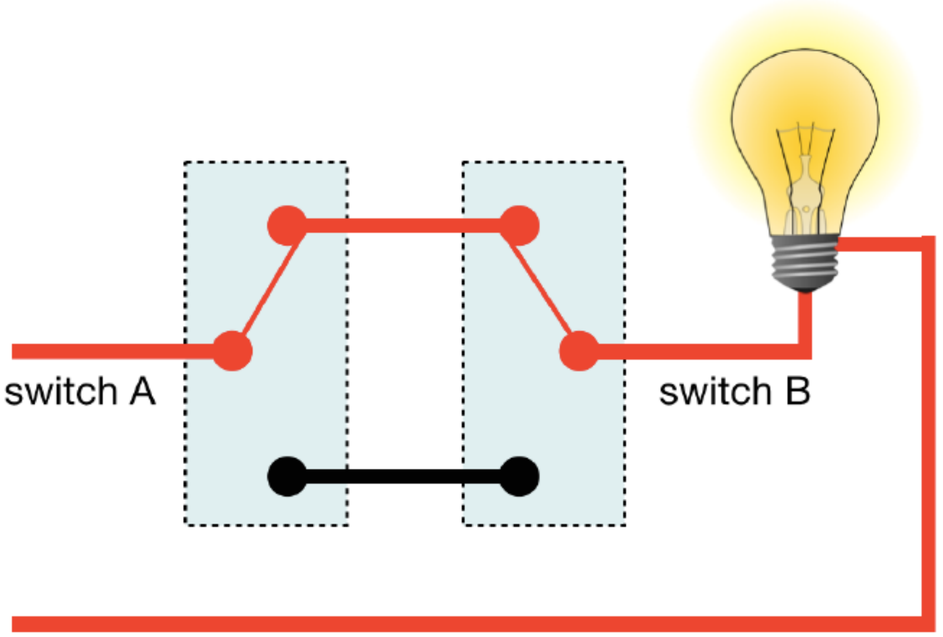
Ciardelli, Zhang & Champollion 2017

Facts:

- A is up
- B is up

Laws:

- Light is on iff A and B agree



‘If B were down, the light would be off’

Background semantics

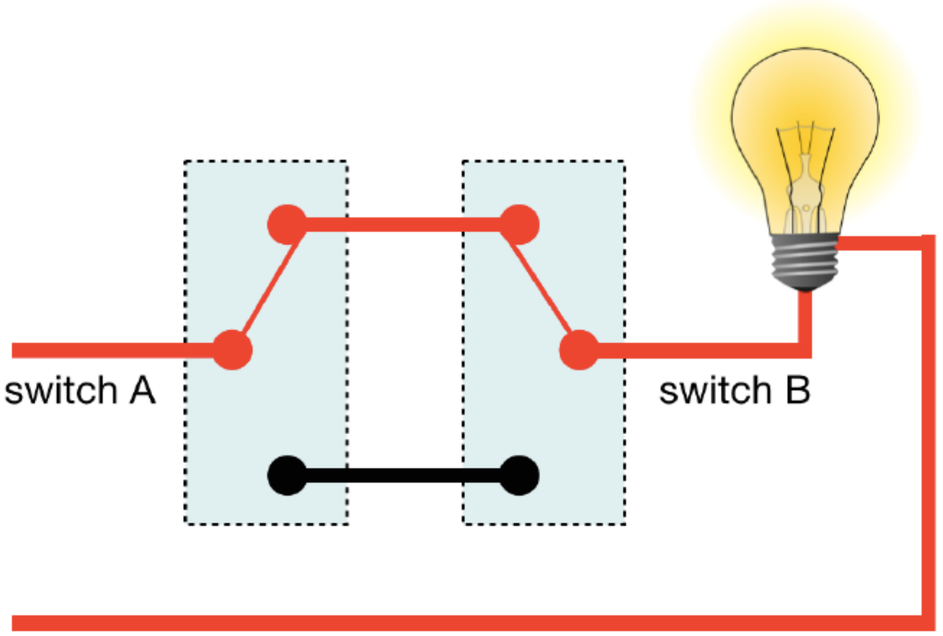
Ciardelli, Zhang & Champollion 2017

Facts:

- A is up
- B is up

Laws:

- Light is on iff A and B agree



‘If A and B were both down, the light would be off’

SOME PUZZLES

Other negated conjunctions, negated universals, indefinites

- If riflemen A and B had not both fired, the prisoner would still have died
  (why? b/c it would be extraordinary if both were to independently …)

- If riflemen A, B, C, D, …, Y and Z had not all fired, the prisoner would still have died

- If not all of these 90,000 fans had come to the concert, there would still be a lot of people here

- If I had a different kind of dog, I'd have a pug

Lesson: 'All models' is too strong

Probability operators in the consequent

If I had a different kind of dog, …

I'd probably have a pug

but I might have a Labrador

(every intervention provides me with a single, fixed dog breed …)

If I were not a physicist,
I would probably be a musician.
I often think in music.
I live my daydreams in music.
I see my life in terms of music.

—Albert Einstein

Failures of Simplification of Disjunctive Antecedents

If it were raining or snowing in Santa Fe, it would be raining

If it were raining or snowing in Santa Fe, it would probably be raining

Failures of SDA

If it were raining or snowing in Santa Fe, it would be raining

≠

If it were raining in Santa Fe, it would be raining

and

if it were snowing in Santa Fe, it would be raining

OK, make it a classical disjunction

If it were raining or snowing in Santa Fe, it would be raining

interpreted classically:

If it were not both not-raining and not-snowing in Santa Fe, …

OK, let's flatten

If it were raining or snowing in Santa Fe, it would be raining

interpreted classically:

If it were not both not-raining and not-snowing in Santa Fe, …

- CZC: remove facts 'no rain' and 'no snow'
- 'rain' can't be true in all consistent models!

CZC predict both readings to be contradictions

Diagnosis

Instantiations of antecedent have very different prior likelihoods

- If the soldiers hadn’t both fired, …
- If not all 26 soldiers had fired, …
- If I had a different kind of dog, …
- If not all of these 90,000 people had come, …
- If it were raining or snowing in Santa Fe, …

HOW TO CHOOSE AN INTERVENTION

Choosing interventions

General inspiration: Dehghani, Iliev & Kaufmann 2012

Idea: we choose interventions using explanatory reasoning

How likely is it that antecedent would have come about this way vs. that way, given the causal model?

Requires probabilistic information

as encoded e.g in causal Bayes nets or structural equation models

Choosing interventions

‘If Mary didn’t have a poodle, she’d have a labrador’

1. Break down complex intervention into simpler components
   - not a poodle => {no dog, labrador, beagle, chihuahua, …}

2. Weight components by probability given pruned facts F*

$$W(I_X) \propto P(X \mid F*)$$

   - Intuitions are graded, tracking probabilities computed in this way
   - NOT the all-or-nothing question: ‘Is consequent is true in all models?’

Structural equation model for Two Switches

$$M_A \sim P(M_A)$$
$$M_B \sim P(M_B)$$
$$A = \neg M_A$$
$$B = \neg M_B$$
$$L = A \Leftrightarrow B$$

$$F = \{A, B, L\}$$

If A were not up …

$$M_A \sim P(M_A)$$
$$M_B \sim P(M_B)$$
$$A = F$$
$$B = \neg M_B$$
$$L = A \Leftrightarrow B$$

$$F = \{B\}$$

Two Switches

$$M_A \sim P(M_A)$$
$$M_B \sim P(M_B)$$
$$A = \neg M_A$$
$$B = \neg M_B$$
$$L = A \Leftrightarrow B$$

$$F = \{A, B, L\}$$

‘If A and B were not both up …’

3 models:

$$M_A = T$$
$$M_B = F$$
$$A = F$$
$$B = T$$
$$L = F$$

Weight: $P(M_A) * (1 - P(M_B))$

(~ ‘If just A were down …’)

$$M_A = F$$
$$M_B = T$$
$$A = T$$
$$B = F$$
$$L = F$$

Weight: $(1 - P(M_A) * P(M_B)$

(~ ‘If just B were down …’)

$$M_A = T$$
$$M_B = T$$
$$A = F$$
$$B = F$$
$$L = T$$

Weight: $P(M_A) * P(M_B)$

(~ ‘If A and B were both down …’)

Two Switches

$$M_A \sim P(M_A)$$
$$M_B \sim P(M_B)$$
$$A = \neg M_A$$
$$B = \neg M_B$$
$$L = A \Leftrightarrow B$$

$$F = \{A, B, L\}$$

If $P(M_A) = P(M_B) = .5$,
what is
P_{CF}(L if not [A & B])?

$$M_A = T$$
$$M_B = F$$
$$A = F$$
$$B = T$$
$$L = F$$

Weight: $P(M_A) * (1 - P(M_B))$

(~ 'If just A were down …')

$$M_A = F$$
$$M_B = T$$
$$A = T$$
$$B = F$$
$$L = F$$

Weight: $(1 - P(M_A) * P(M_B)$

(~ 'If just B were down …')

Weight: $P(M_A) * P(M_B)$

(~ 'If A and B were both down …')

$$M_A = T$$
$$M_B = T$$
$$A = F$$
$$B = F$$
$$L = T$$

Concert example

If not all of these 90,000 fans had come, …

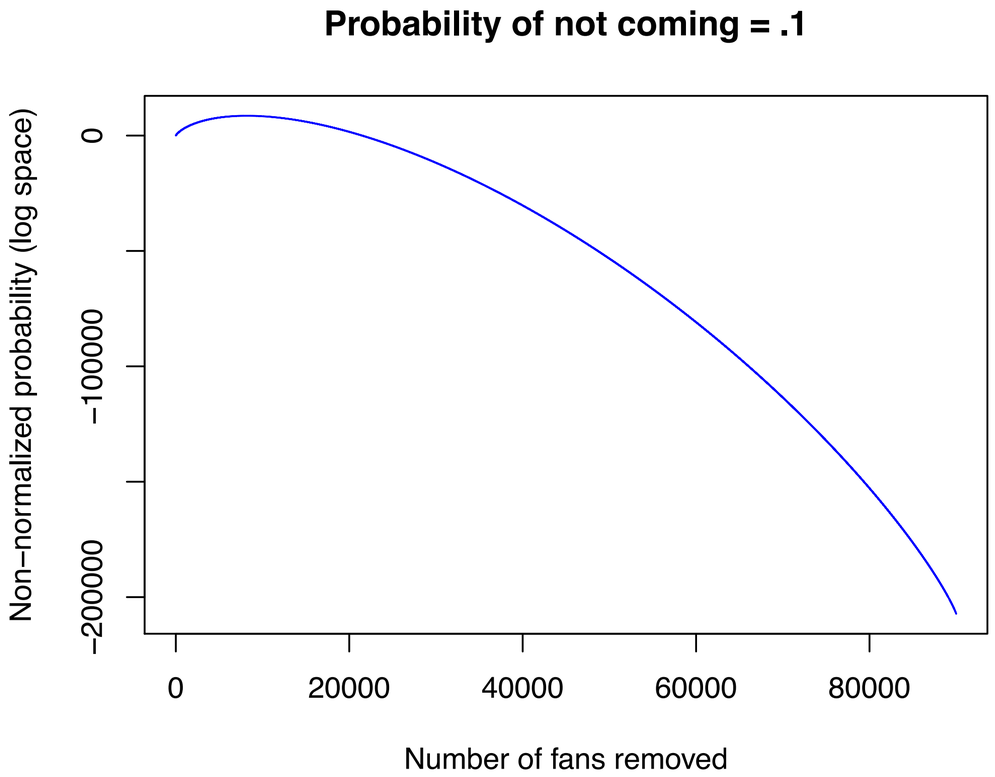
1. Break down complex antecedent
   - not all => {fan 1 doesn't come, … only fans 1000-2000 come, … none one comes}

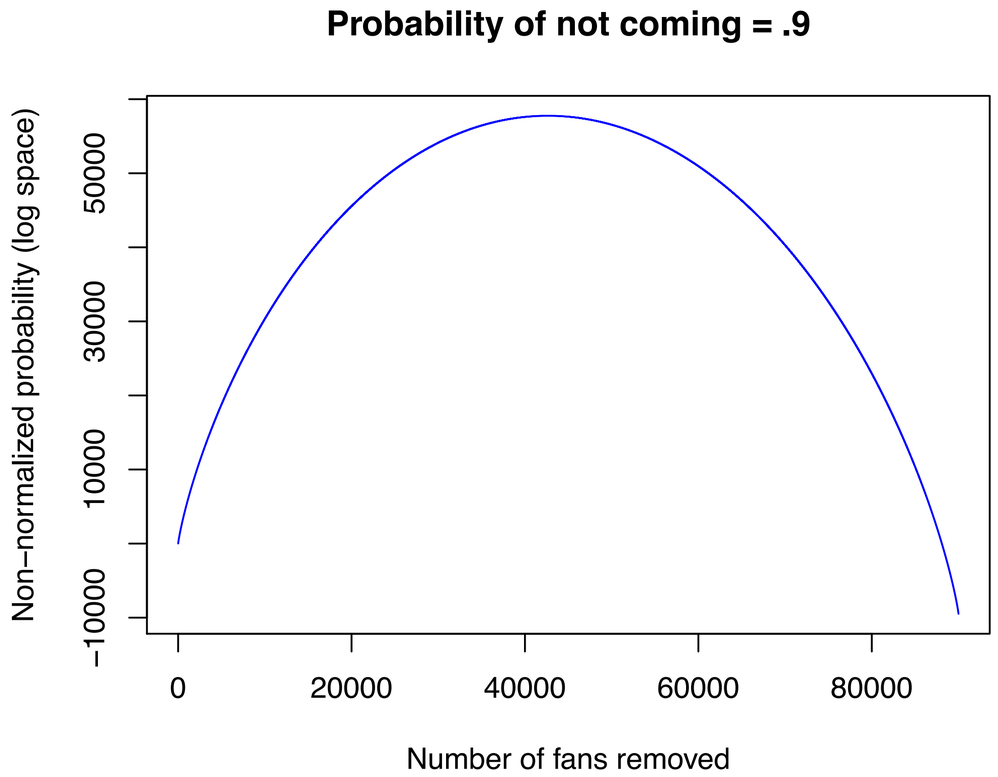
2. Weight components by probability given pruned facts
   - say P(person i does not come) = q, all independent
   - P(person 1 doesn't come) proportional to q
   - …
   - P(no one comes) proportional to q^{90000}

requires a stunning coincidence

Predictions

If not all of these 90,000 fans had come, …





soft preference for more ‘minimal’ revisions

Non-SDA disjunctions

(1) If it were raining or snowing in Santa Fe, it would be raining

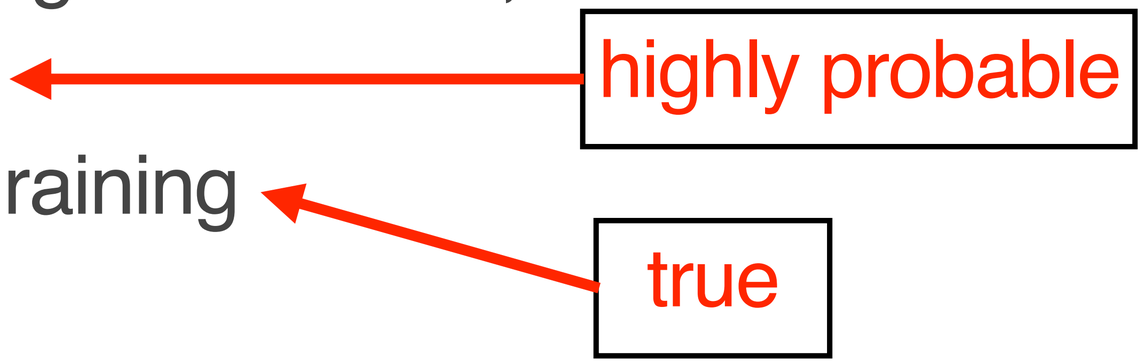
$$P(1) = \frac{W(I_{\mathbf{rain}})}{W(I_{\mathbf{rain}}) + W(I_{\mathbf{snow}})}$$
$$= \frac{P(\mathbf{rain} \mid \mathbf{rain} \cup \mathbf{snow})}{P(\mathbf{rain} \cup \mathbf{snow})}$$

E.g., if P([sun, rain, snow]) = [.9, .09, .01], then P(1) = .9

Probability operators

Immediate:

2) If it were raining or snowing in Santa Fe,
   a) it would be raining
   b) it would probably be raining



(2b) is true iff $P(2a) > \theta_{probable}$

Compositional derivation: small extension of Lassiter 2017

Implications

Counterfactuals do not have determinate truth-conditions unless the antecedent picks out a unique intervention

Explanatory reasoning is invisible for

- simple antecedents (weight is w/w = 1)
- antecedents where all instantiations affect consequent the same way

‘Mary loves dogs but is indifferent among breeds.
If she had a different kind of dog, she’d still be happy’

PILOT 1:
CONCEPTUAL REPLICATION OF CZC, MANIPULATING CORRELATION

Example scenario

Logical structure: minimal and non-minimal revisions disagree on consequent
Manipulate correlation: positive, negative, none
N = 600, 2 scenarios

In 19th-century Bavaria, King Ludwig and Queen Maria lived in Neuschwanstein castle.

After many years of marriage, they hated each other. The king and queen disliked being together and so spent most of their time in different places.

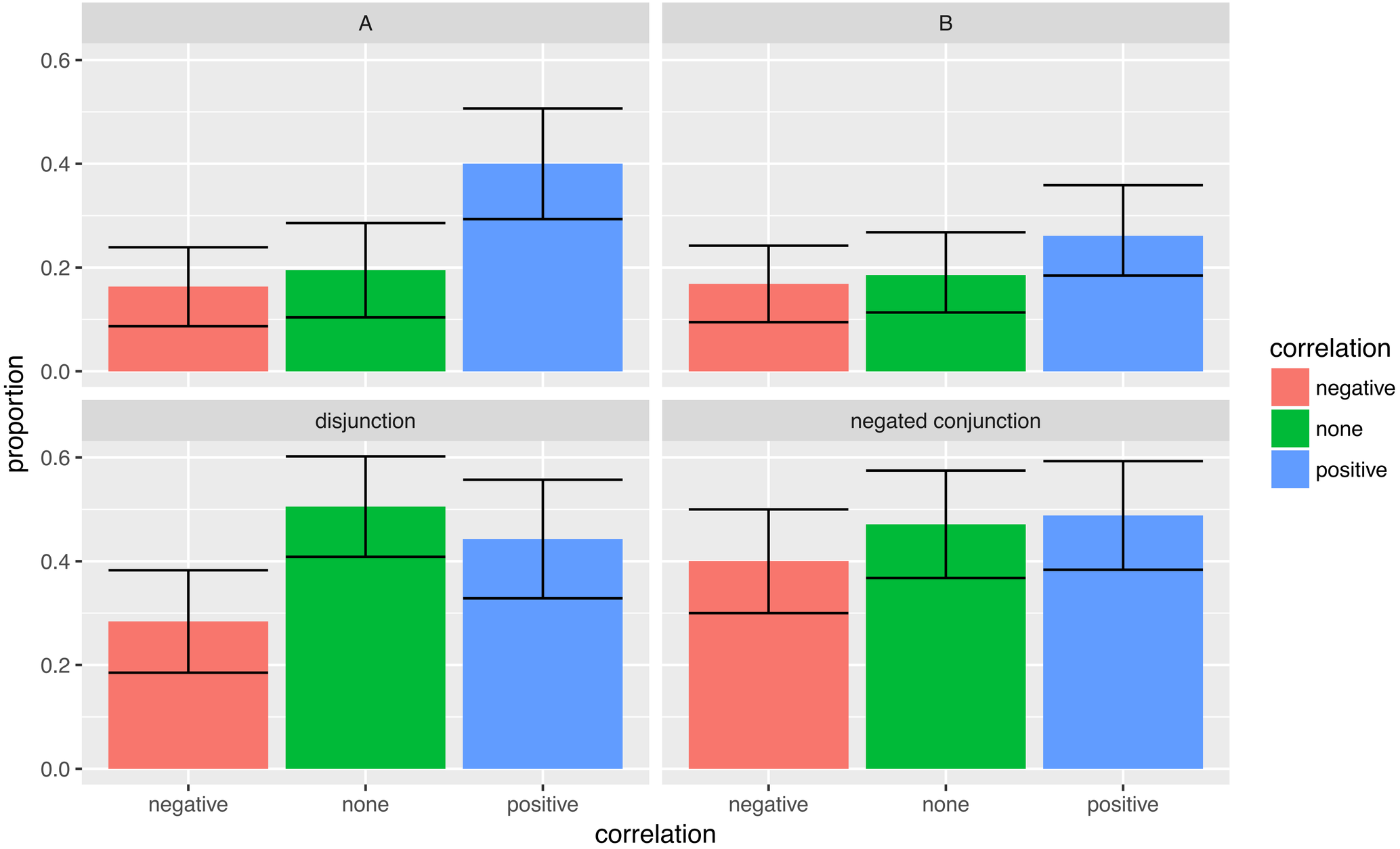
In the castle there was an official flagkeeper, whose job was to ensure that the royal banner was flying whenever the King and Queen were **both** in the castle. However, whenever one or both of them was away from the castle, the flag was always taken down.

One day, the King and the Queen were both away from the castle, and the flag was down.

Is the following statement true or false?

If the King and the Queen were not both away, the flag would be up.

Aggregate results



Experiment 1: Discussion

Key result: Correlation matters (likelihood ratio test: $p < 10^{-4}$)

CZC's disjunction vs. negated conjunction contrast may be present, but it's small

- could be a problem with the theory … or the experiment

hot off the press: Romoli et al. (AC, 2019) find a similar contrast with arguably more natural materials

- but attribute the effect to overt negation
- predicted by exhaustification-based approach of Bar-Lev & Fox, 2019

PILOT 2:
NON-CATEGORICAL SHIFTS IN PROBABILITY MATTER FOR INTERVENTION CHOICE

Example scenario

cf. Petrocelli, Percy, Sherman & Tormala 2011, J.Pers.Soc.Psych.

One day Mary is on the game show "Let's Make a Deal". Mary's options include picking Door #1, Door #2, Door #3, and Door #4. Behind three of the doors there is nothing. Behind one of the doors is $50,000. If Mary picks the correct door, she will win $50,000. Otherwise, Mary will get nothing.

She immediately eliminates Door #4, since it's her unlucky number. She takes her time choosing among doors #1, #2, and #3, thinking hard. She agonizes over the decision.

Mary thinks about Door #2 as well as Door #3. She goes back and forth in her mind. The game show host pressures Mary for an answer. At the last moment, Mary picks Door #3.

Sadly, the $50,000 is actually behind door #2, and Mary wins nothing.

Is the following description true or false here?

If Mary hadn't chosen Door #3, she'd have won.

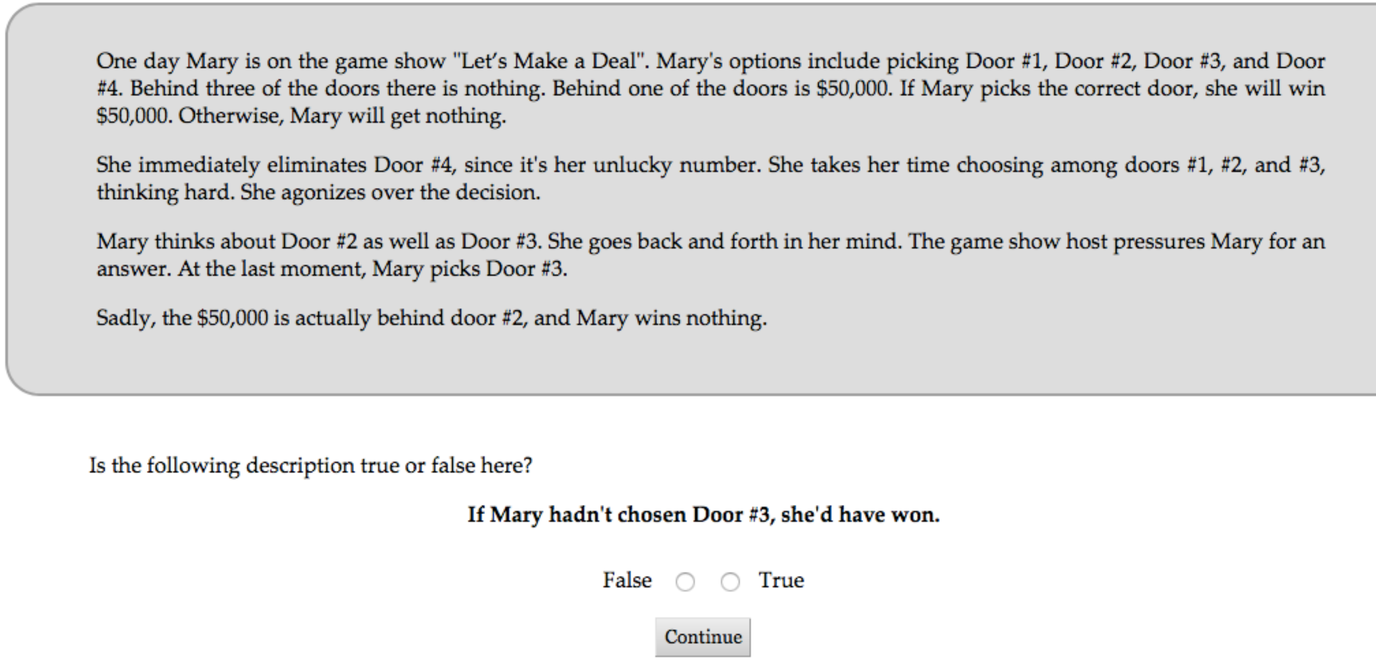
False ○ ○ True

Continue

MTurk experiment

3 scenarios, always 4 choices:

- one ruled out early (impossible)
- either:
  - no info about initial preferences
  - one initially preferred
- one eventually selected



Manipulate whether agent was leaning toward or away from eventual choice

- Forced-choice True/False judgment, N=300
- Prompts: negation (‘If Mary hadn’t chosen door 3’)
  ‘a different’ (‘If Mary had chosen a different door’)

Predictions

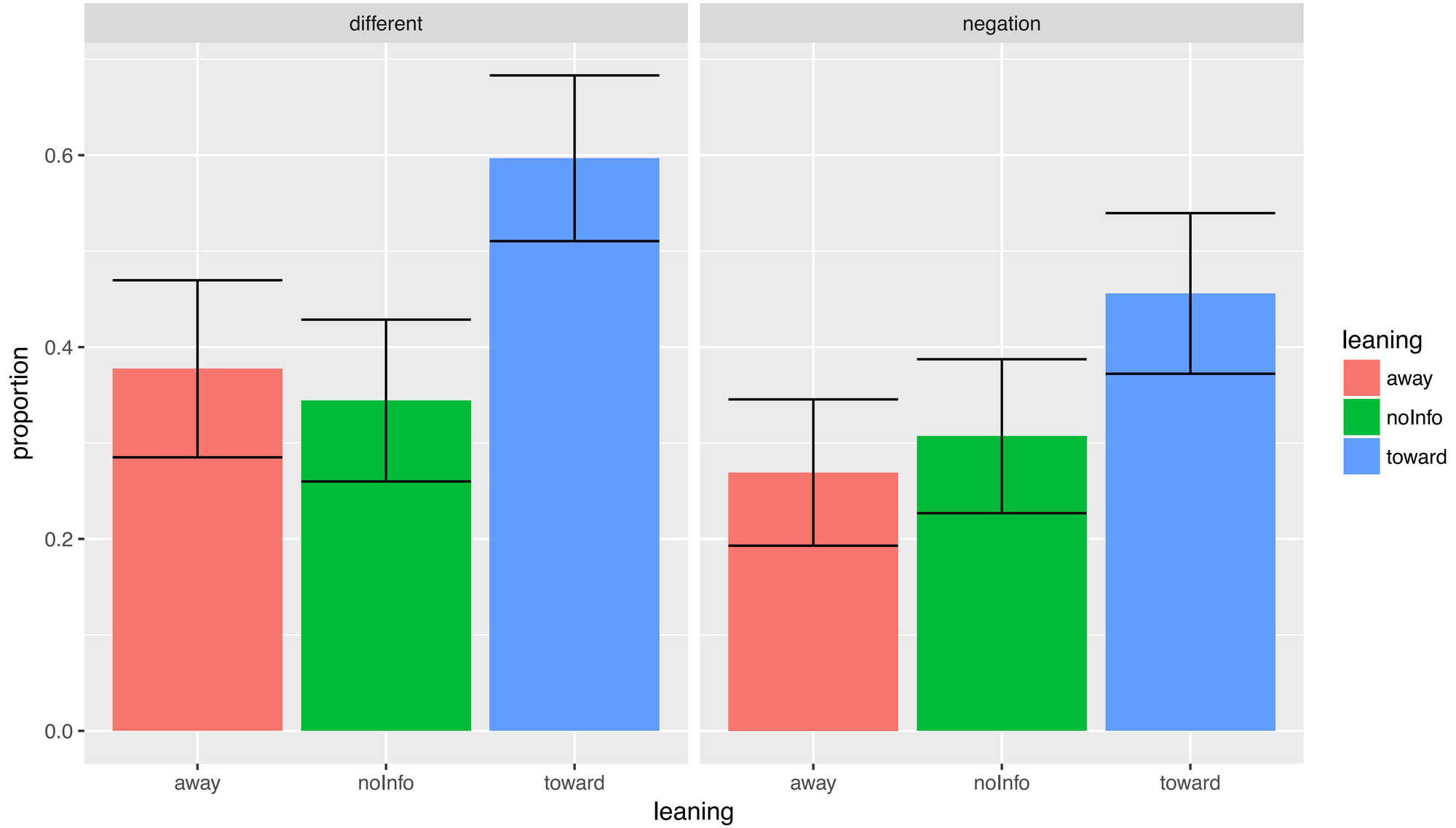
Pearl: no predictions

CZC, Santorio, Briggs: No effect of non-categorical shifts in probability

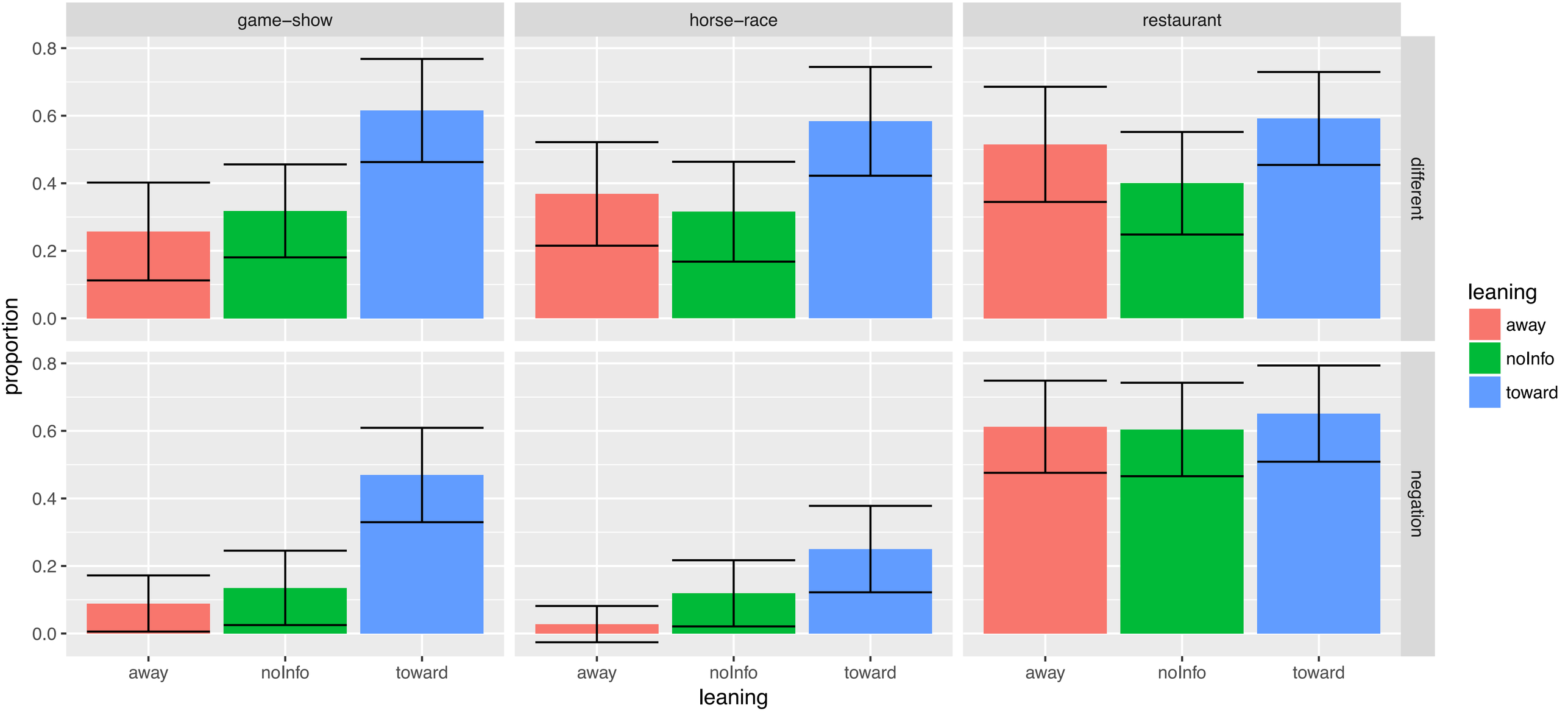
- Wiggle room: treat low probability events as irrelevant

Me: graded effects of shifts in probability

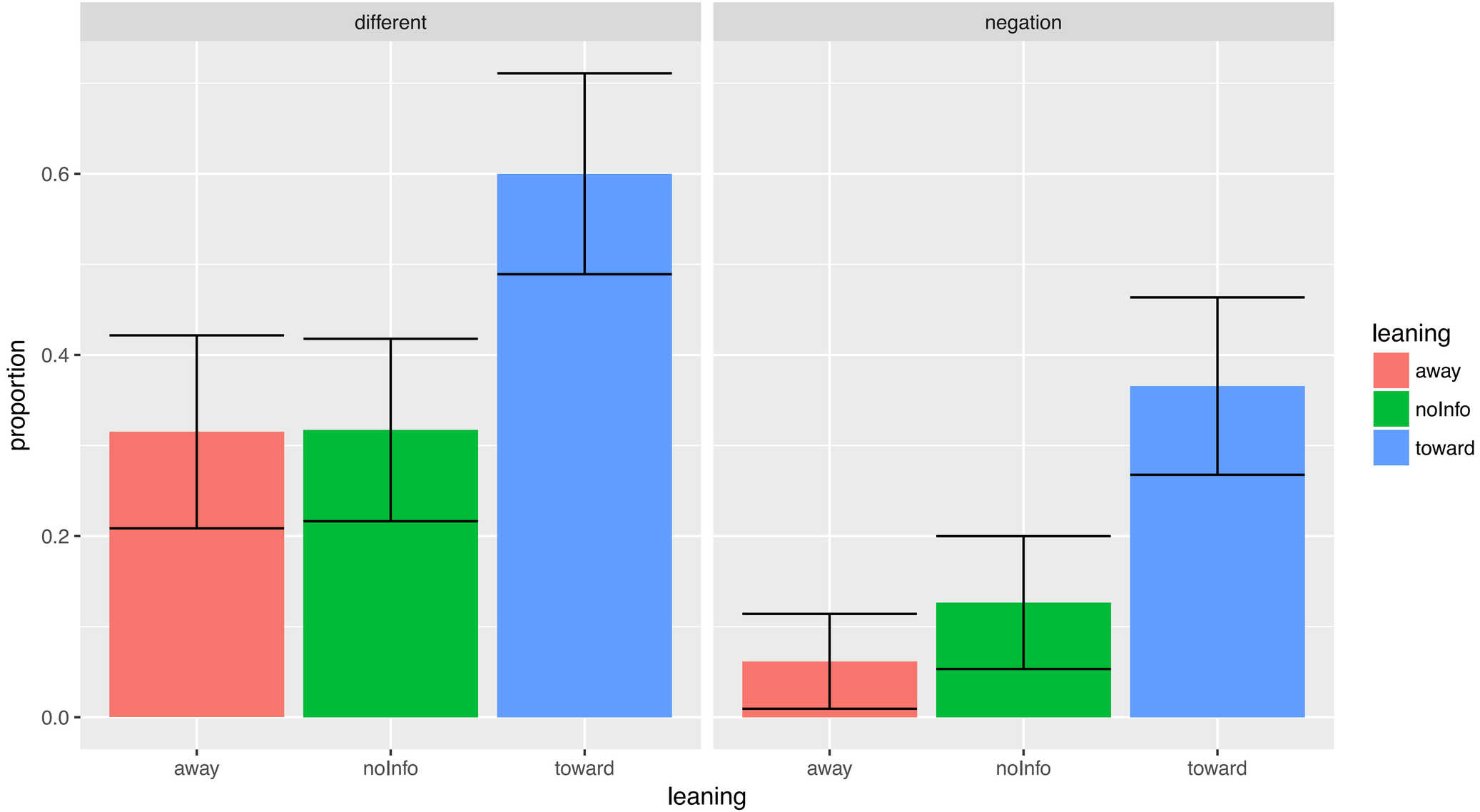
Aggregate results



Forced-choice T/F by scenario

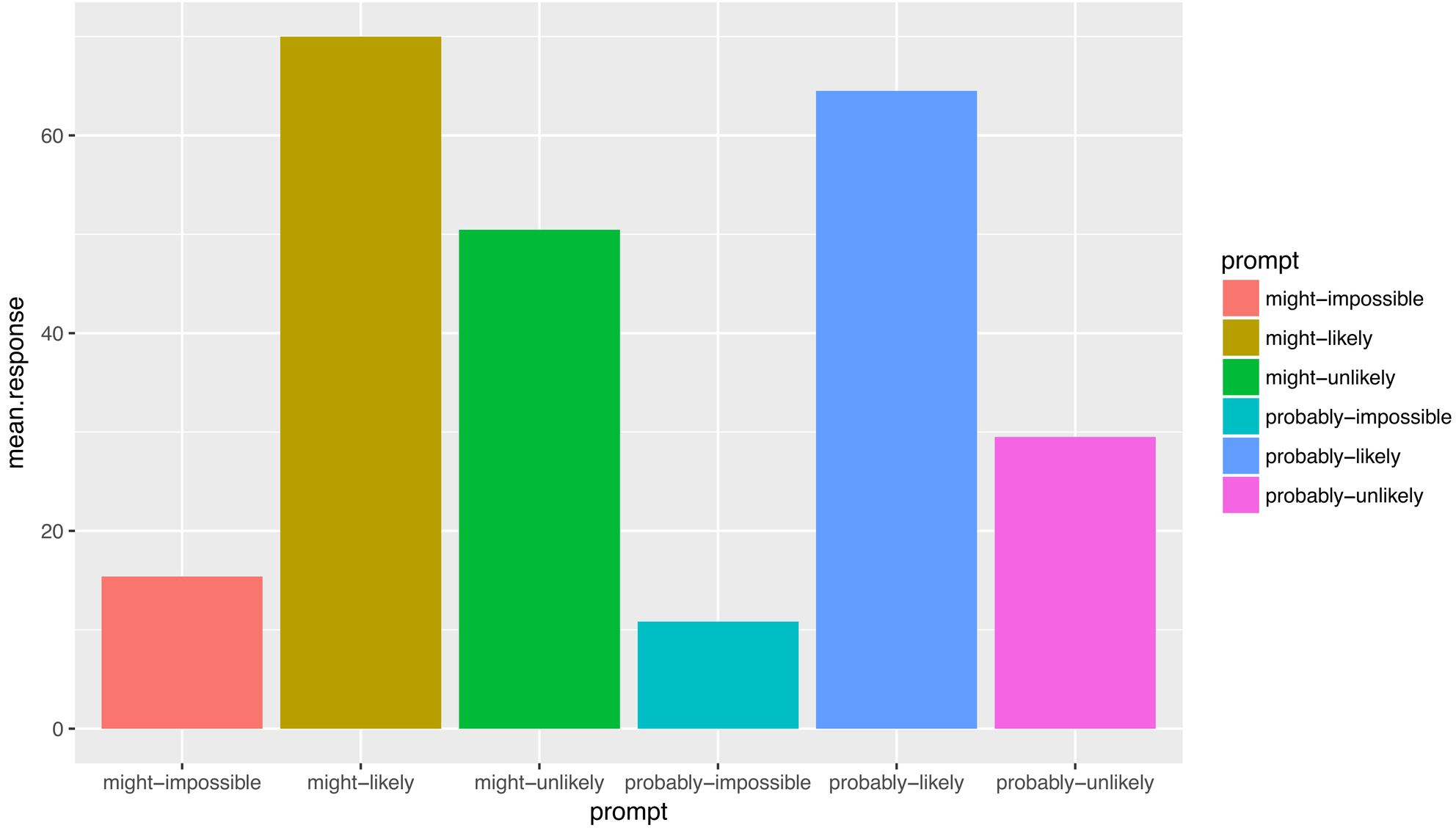


Aggregate results w/o the troublesome scenario



Manipulation check

- Probably: Ss pick up on probability manipulation
- Might: Unlikely events are not lumped with impossible



Experiment 2: Discussion

- Explanatory reasoning affects the way antecedents are mapped to interventions

- Preference for more likely scenarios
  - is soft (probabilistic)
  - can’t be attributed to domain restriction/etc

- This is predicted by explanatory intervention choice
  - but not by Pearl, CZC, etc

PILOT 3:
TESTING QUANTITATIVE PREDICTIONS

(JOINT WORK WITH TOBI GERSTENBERG
& DISHA DASGUPTA)

Concept

- Multiple ways to instantiate a quantified antecedent
- Clear predictions about qualitative & quantitative patterns

Instructions

In this experiment, your task is to answer questions about different systems of pipelines that supply a factory with water. We will show you diagrams like the ones in Figure 1 to illustrate how the different systems work.

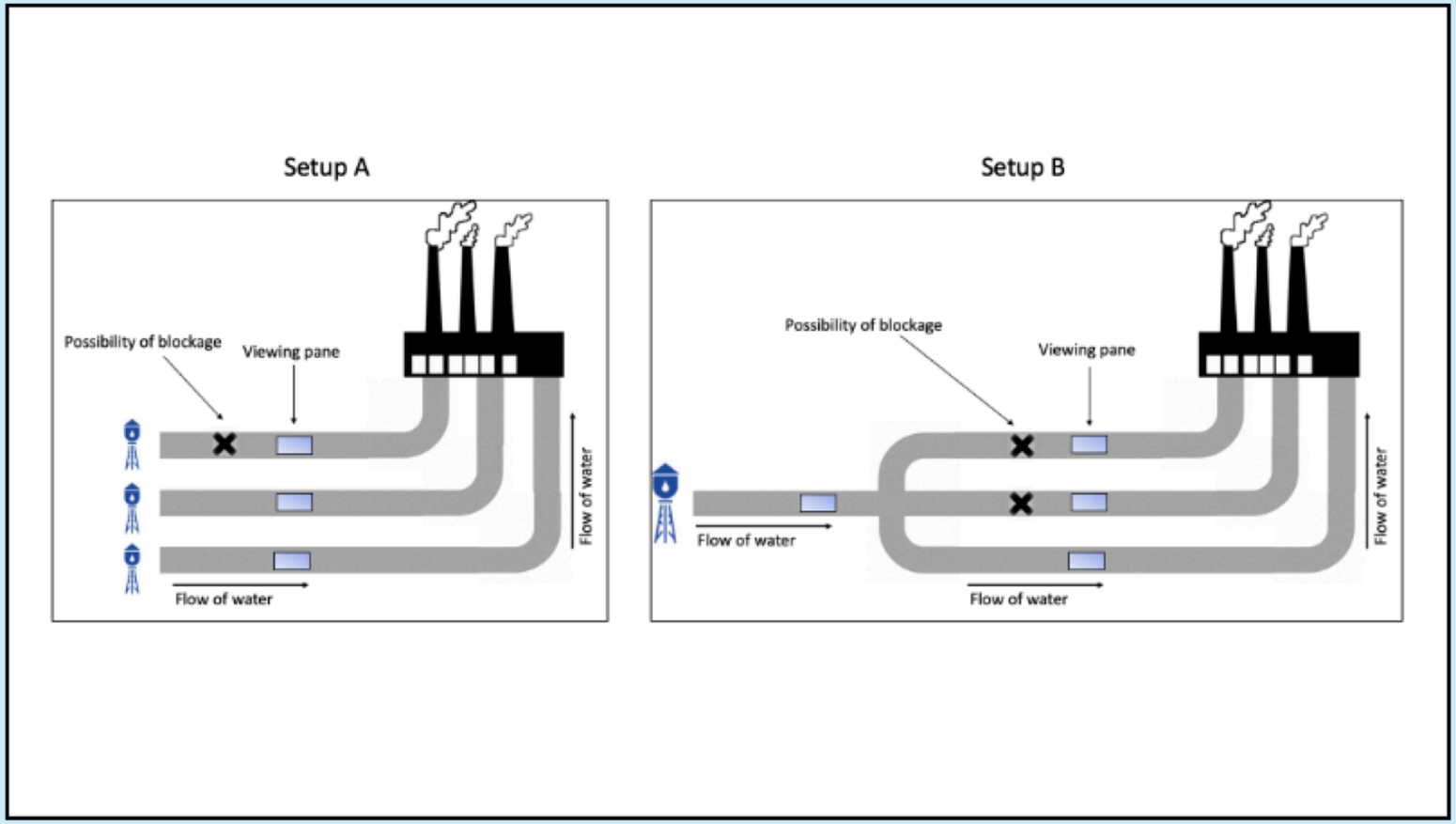


Figure 1. Diagrams illustrating two different water pipe setups.

Water supply:

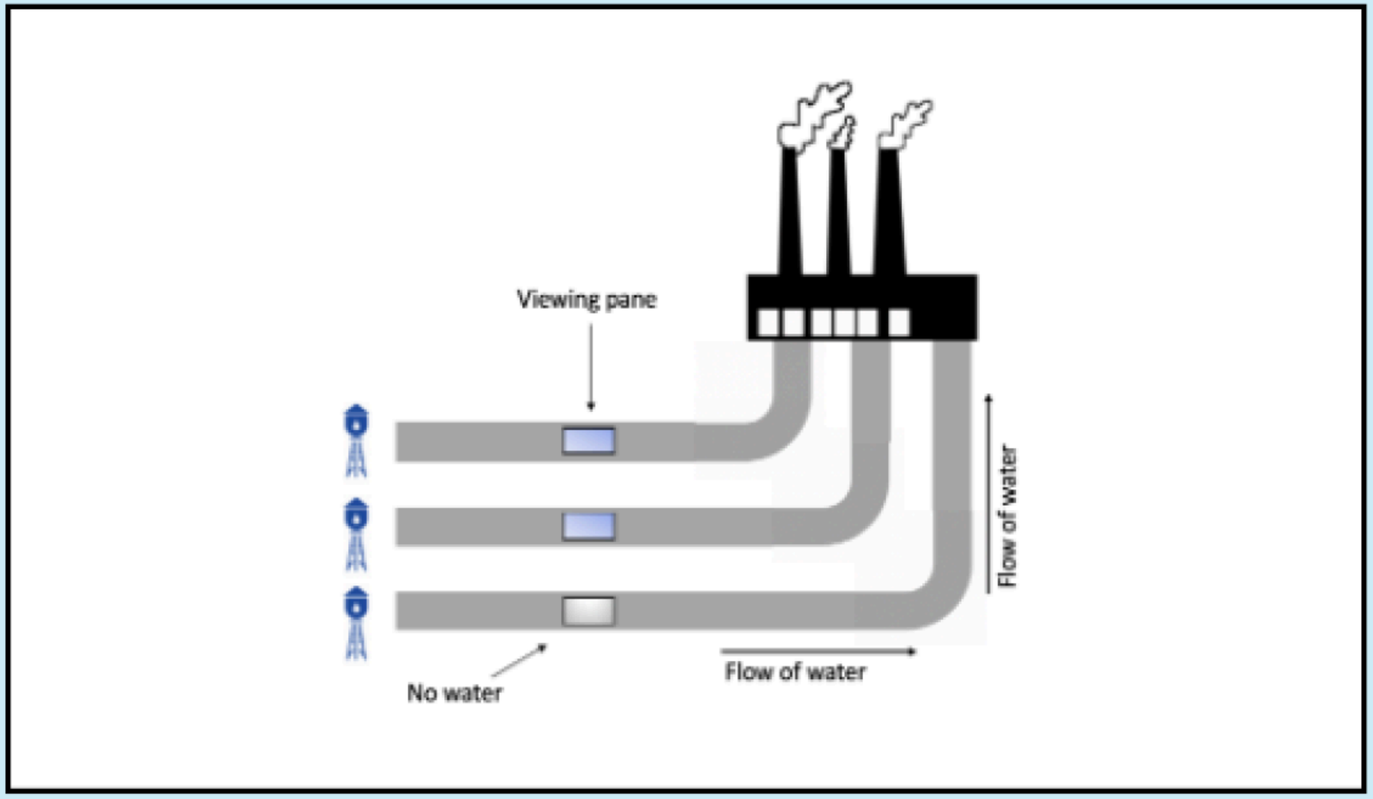
Each factory setup has several pipes that supply the factory with water. The water flows from the water pumps on the left to the factory on the right. Setup A has three separate water pumps that supply the factory with water though separate pipes. Setup B has a single water pump that supplies the factory's water.

In order for the factory to have water, water has to flow through at least one of the pipes that end in the factory. For

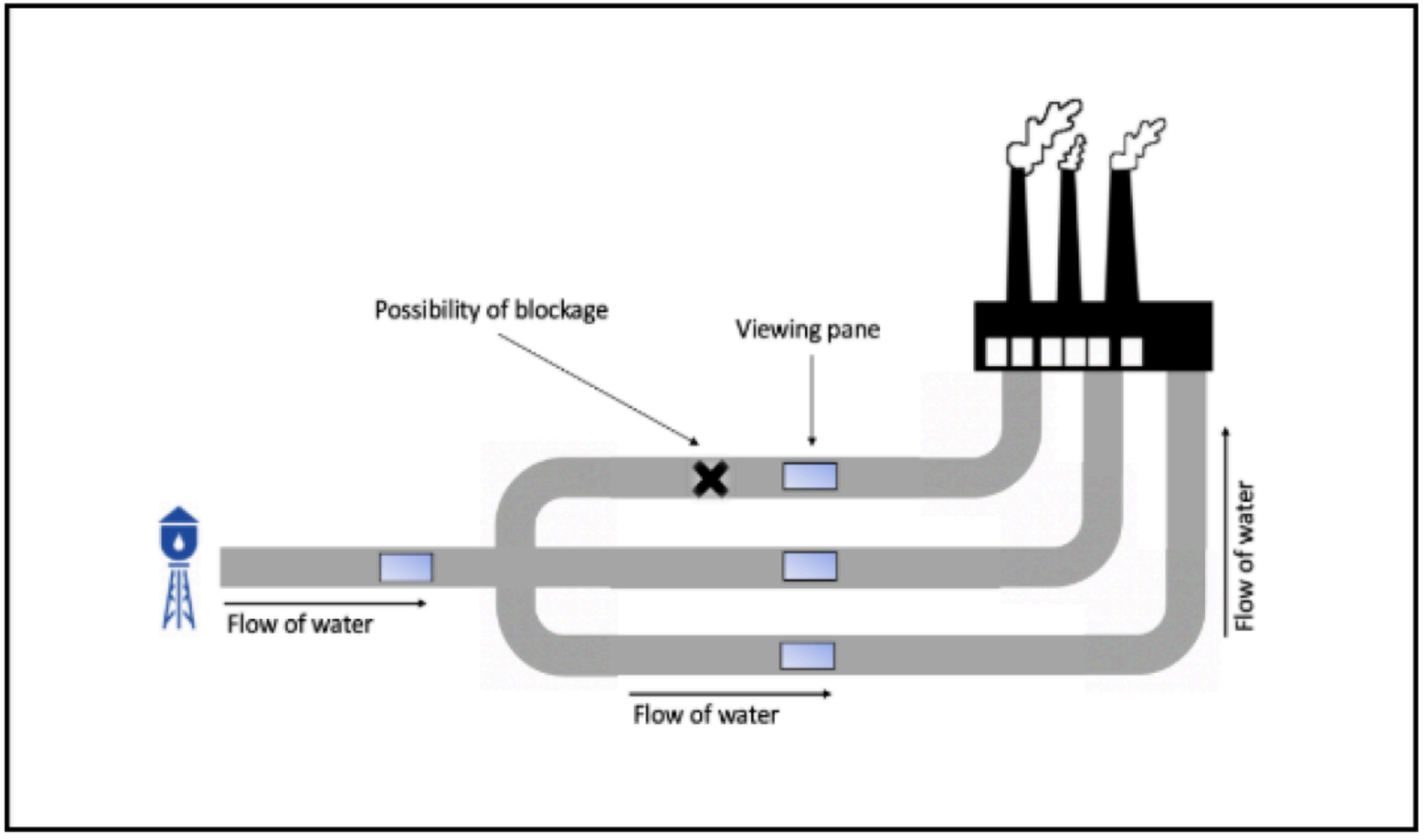
Check questions

Please answer the following questions:

1. In the pipeline system, a gray viewing window on a pipe shows that there is no water flowing through that pipe. In the image below, would there be water at the factory?



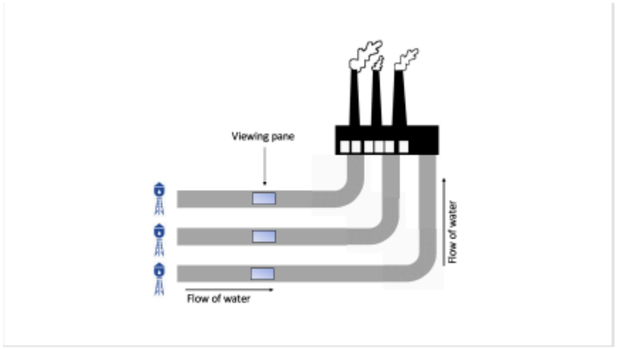
- Yes
- No



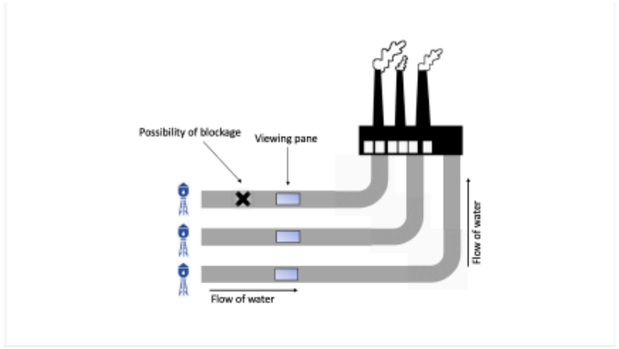
Given this particular setup of water pipes, to what extent do you agree with the following statement?

If some of the pipes did not have water, the factory would still have water.

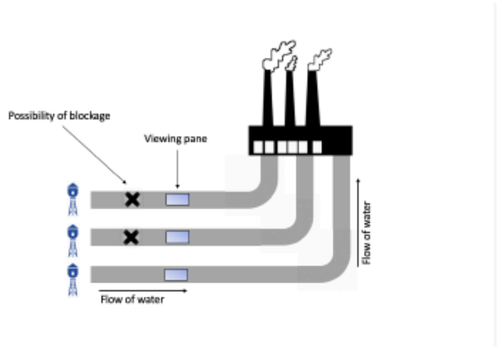
Not at all Unsure Very much



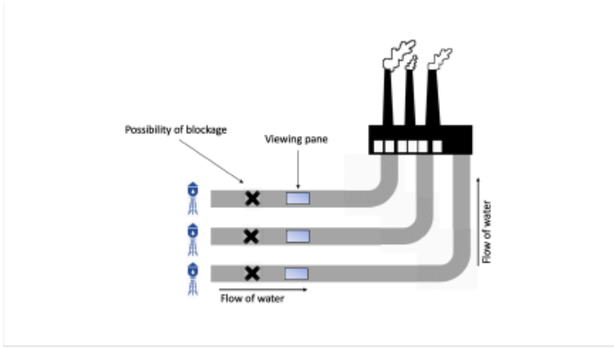
(1)



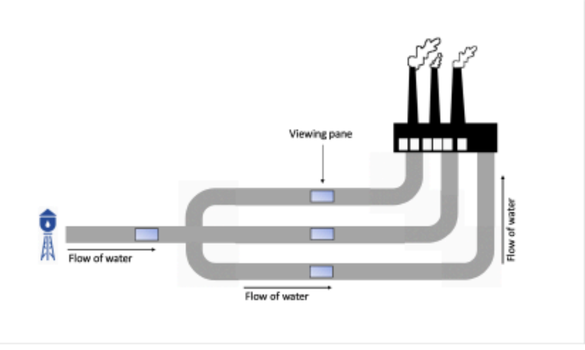
(2)



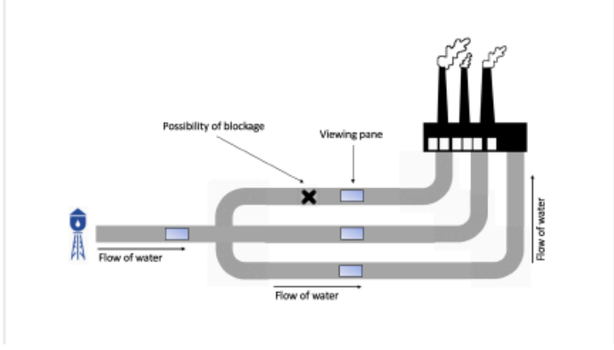
(3)



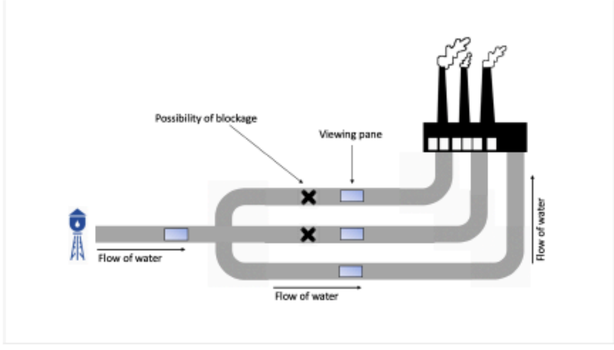
(4)



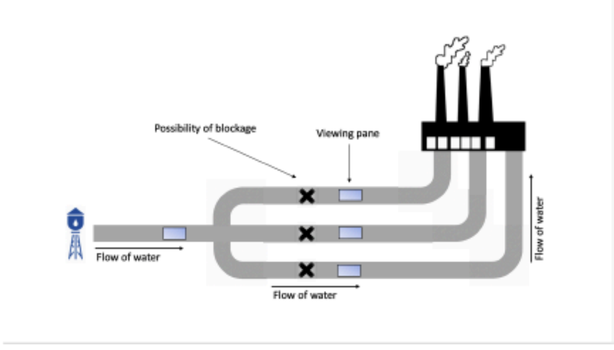
(5)



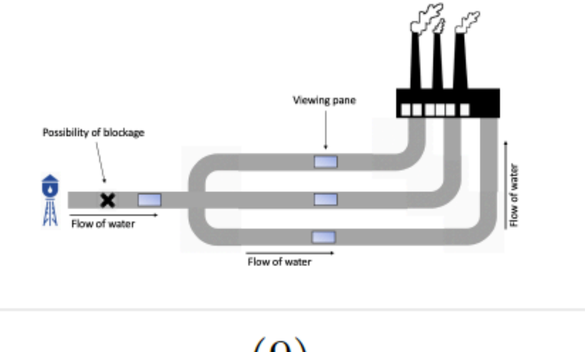
(6)



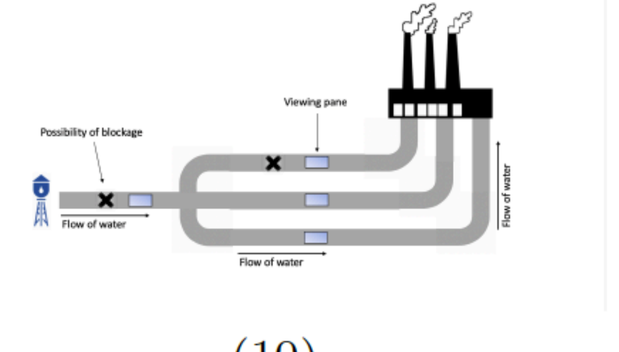
(7)



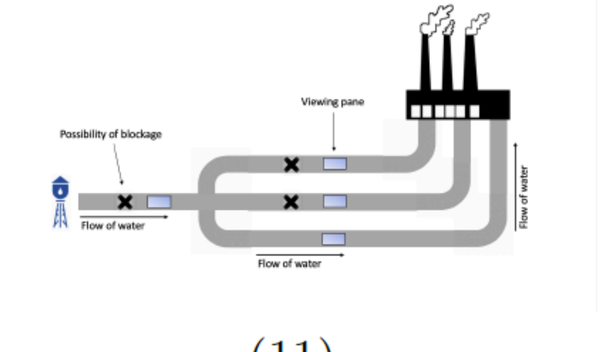
(8)



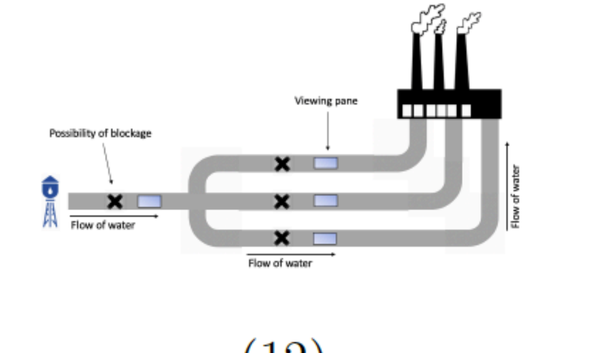
(9)



(10)

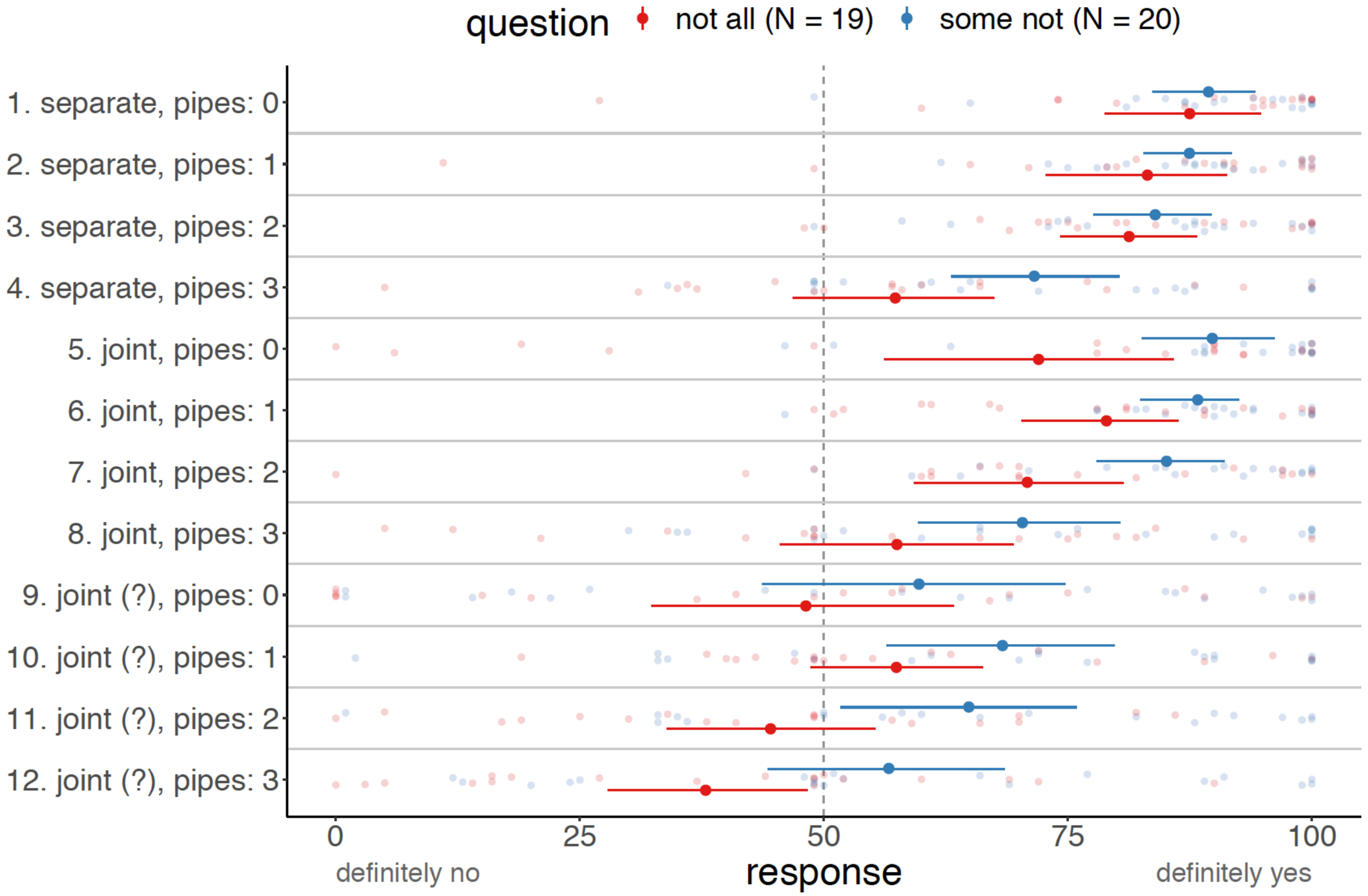


(11)

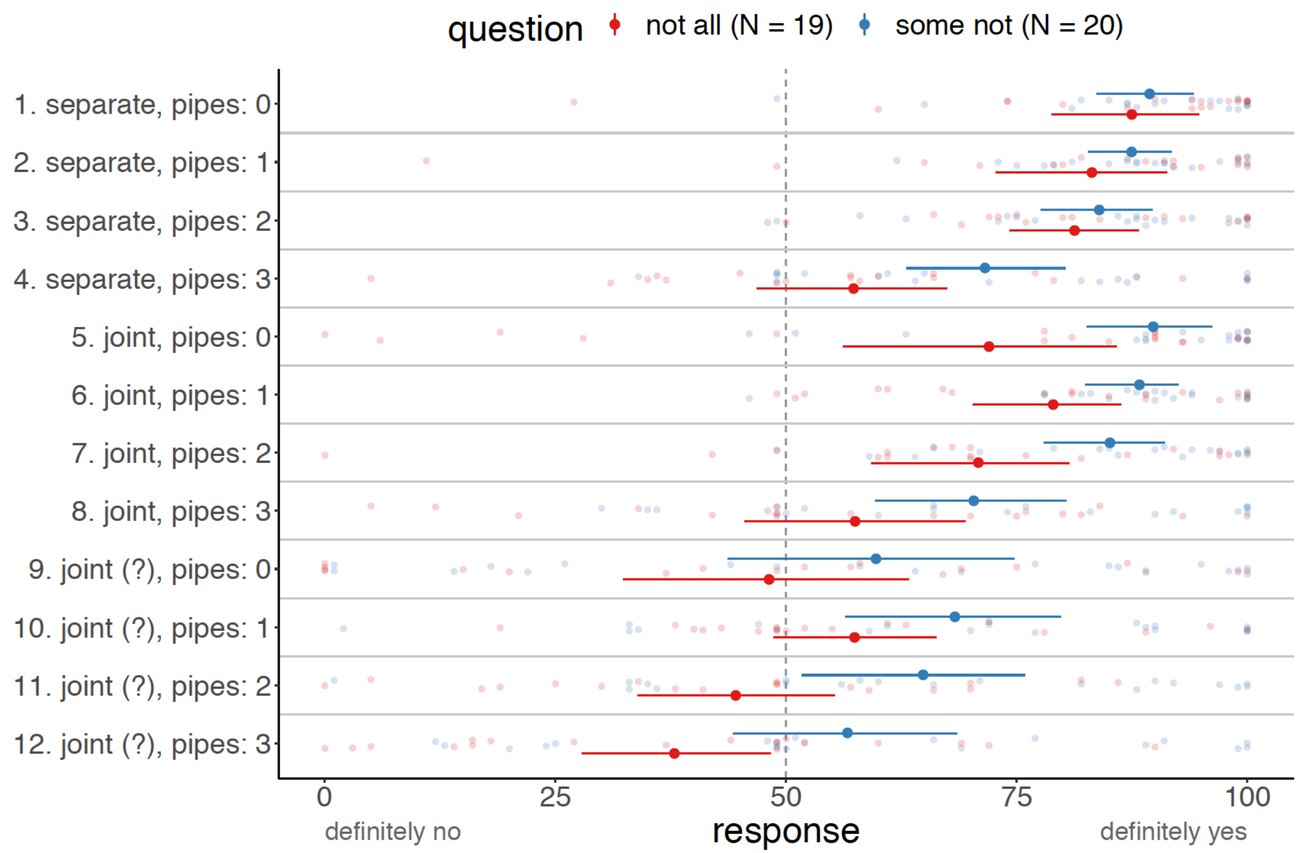


(12)

Results


question
not all (N = 19)
some not (N = 20)
1. separate, pipes: 0
2. separate, pipes: 1
3. separate, pipes: 2
4. separate, pipes: 3
5. joint, pipes: 0
6. joint, pipes: 1
7. joint, pipes: 2
8. joint, pipes: 3
9. joint (?), pipes: 0
10. joint (?), pipes: 1
11. joint (?), pipes: 2
12. joint (?), pipes: 3
0
25
50
75
100
definitely no
response
definitely yes

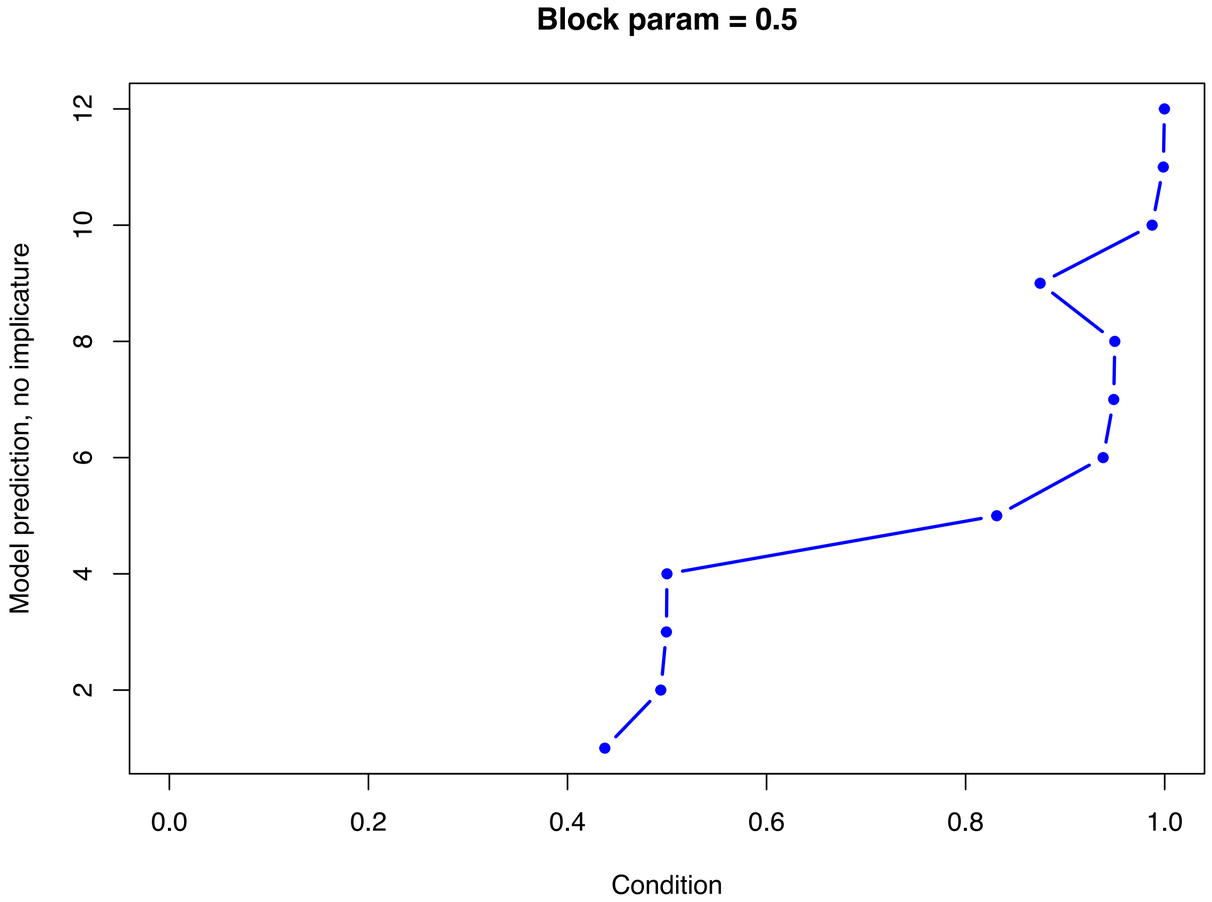
Experimental data



Model

parameters:

- P(failure | ?) = .5
- P(failure | no ?) = .05



summary

- Semantics for counterfactuals built on causal models
- The problem of complex antecedents
  - Ciardelli, Zhang, & Champollion's contribution
  - Some lingering issues
- How to choose your intervention
- Some (preliminary) experimental evidence

summary

- Causal-models semantics for counterfactuals
- The problem of complex antecedents
- How to choose your intervention
- Experimental evidence

Thanks!

Contact: danlassiter@stanford.edu