

# Computational Cognitive Science

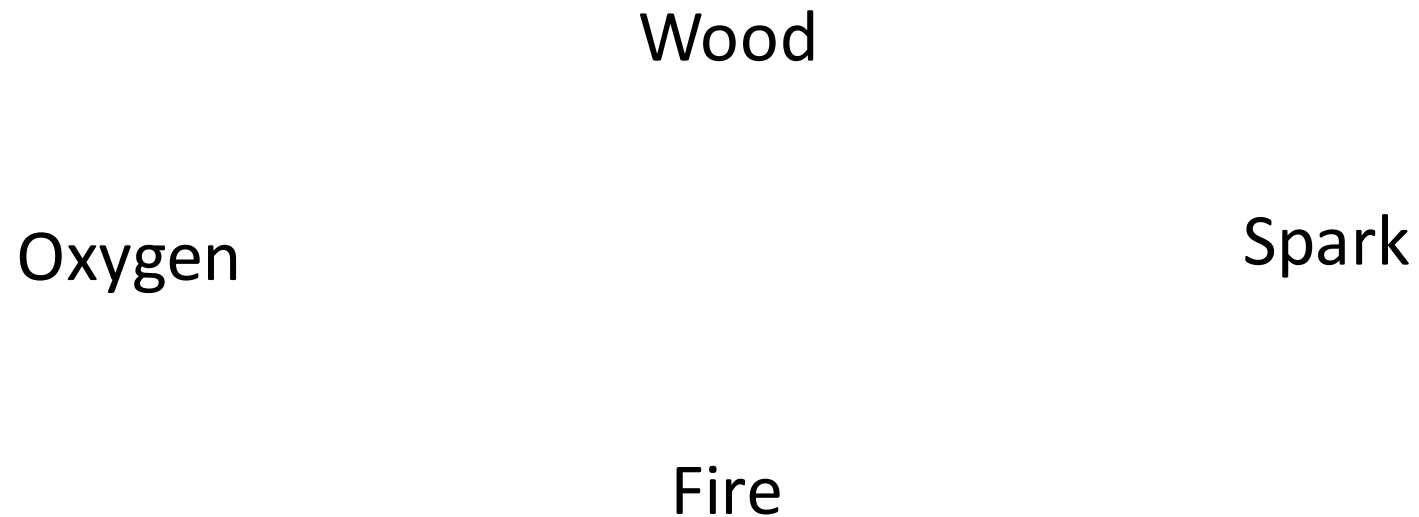
Lecture 12: Causality 3

Guest lecturer: Tadeq Quillien

School of Informatics, University of Edinburgh

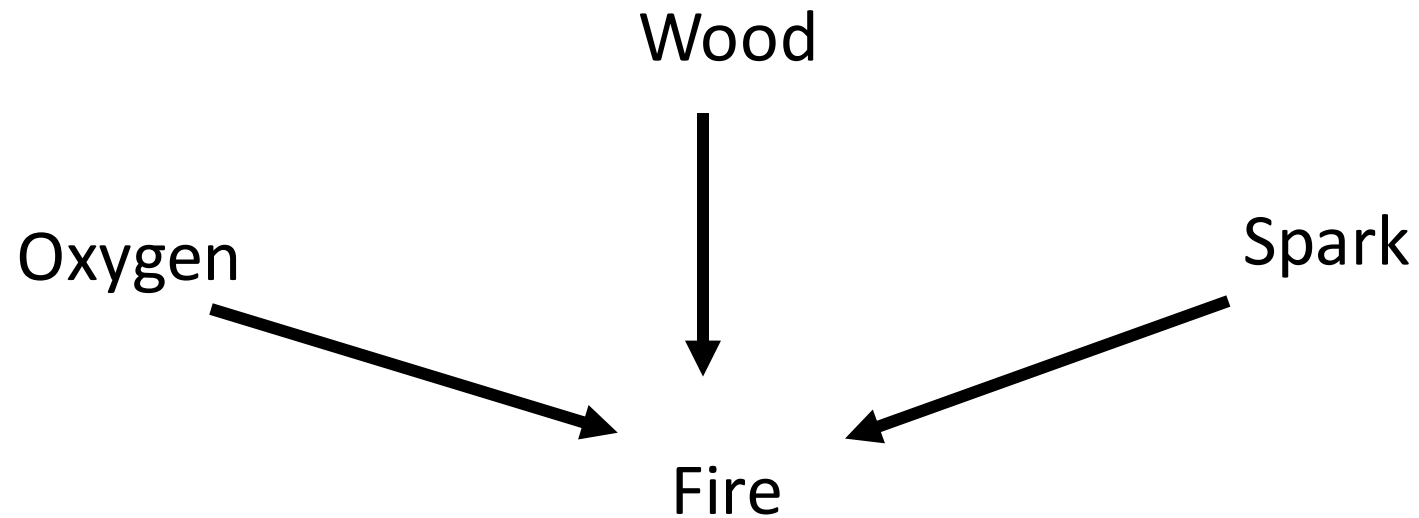
# Last week: causal inference

How can we discover the general causal relations among all these things?



# Last week: causal inference

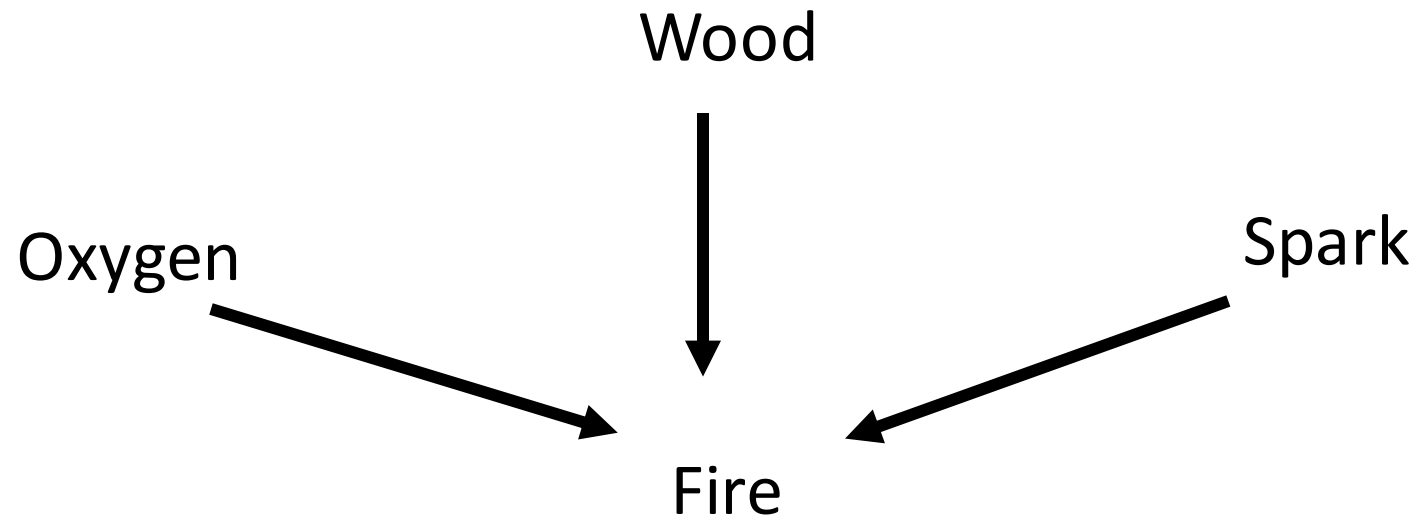
The goal is to discover the correct causal model:



# This week: 'actual causation'

Assume that we already know the causal model below

Suppose a friend asks you why a fire happened. What do you tell them?

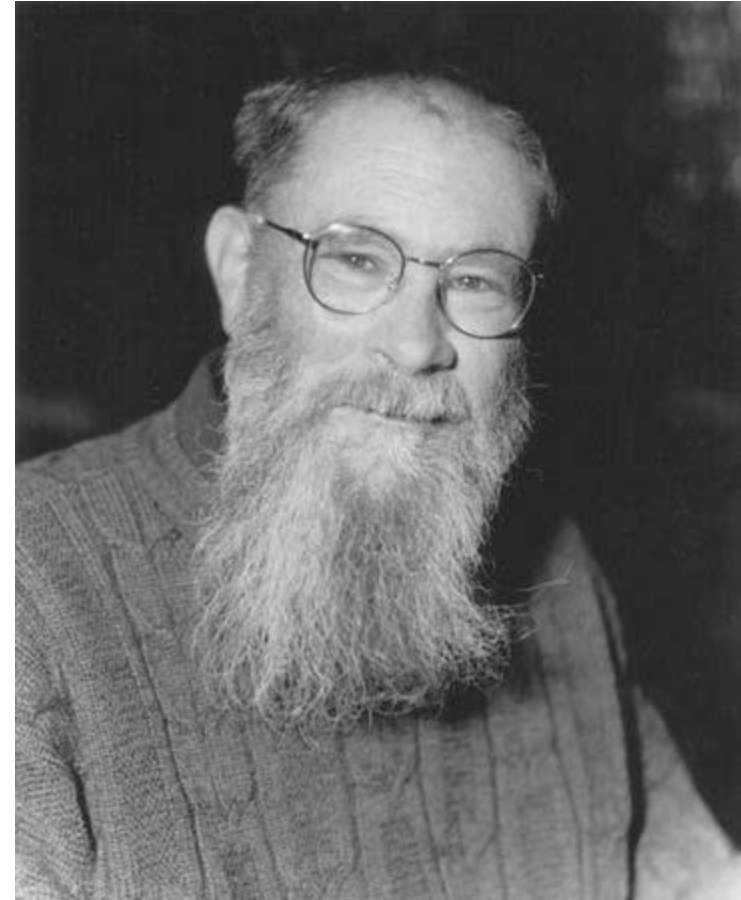


# Counterfactual theory of causation (e.g. David Lewis)

- C is a cause of E if:

If C had not happened, E would not have happened either

- Without the spark, the fire would not have started -> The spark caused the fire



# Problems with the counterfactual approach

- If a meteor had struck Edinburgh this morning, I would not be giving this lecture  
-> I am giving this lecture because no meteor struck Edinburgh this morning
  
- If there had been no oxygen in the air, the fire would not have started  
-> The fire started because there was oxygen in the air



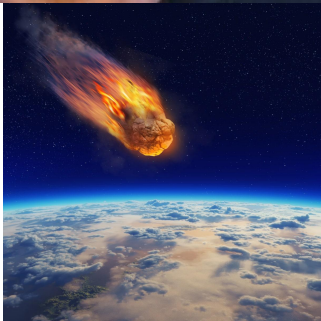
# Problems with the counterfactual approach

- The prisoner would be dead, even if soldier A had not shot
- The prisoner would be dead, even if soldier B had not shot
- -> None of the soldiers caused the prisoner's death!



# Saving the counterfactual theory: “invariant” counterfactual dependence (Jim Woodward)

- To be a cause of E, the link between C and E must be *invariant*
- I.e. C would have led to E even if the background conditions had been different
- The absence of meteor is not an invariant cause of my giving this lecture





# Saving the counterfactual theory: “invariant” counterfactual dependence (Jim Woodward)

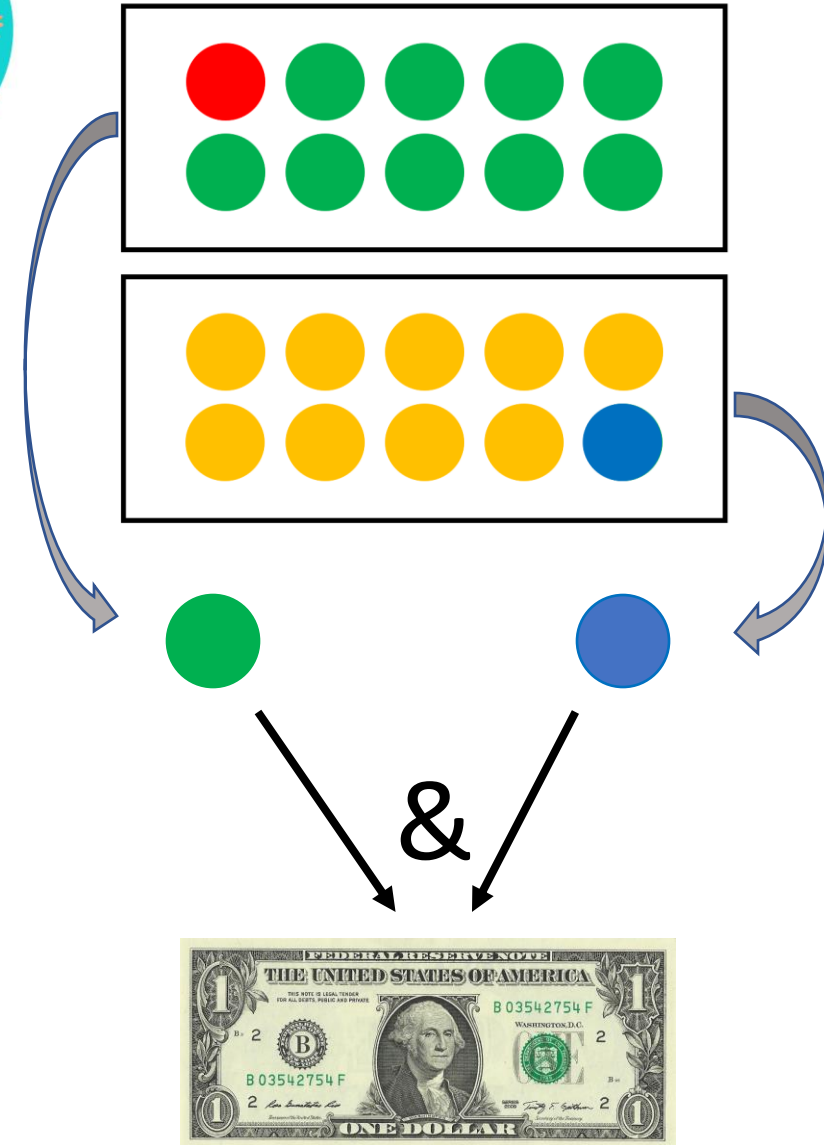
- Oxygen is not an invariant cause of the fire
- Soldier A shooting is an invariant cause of the prisoner’s death
- Is there experimental evidence for the role of invariance?



You win a dollar if and only if you get a green ball from the top box **AND** a blue ball from the bottom box.

Did Joe win a dollar because he drew a **green** ball, or because he drew a **blue** ball?

(Morris et al., 2019, PLoS One)



- “Invariance” is still a vague philosophical notion
- What computations actually underlie our sense of causation?






















# Counterfactual effect size model (Quillien, 2020)

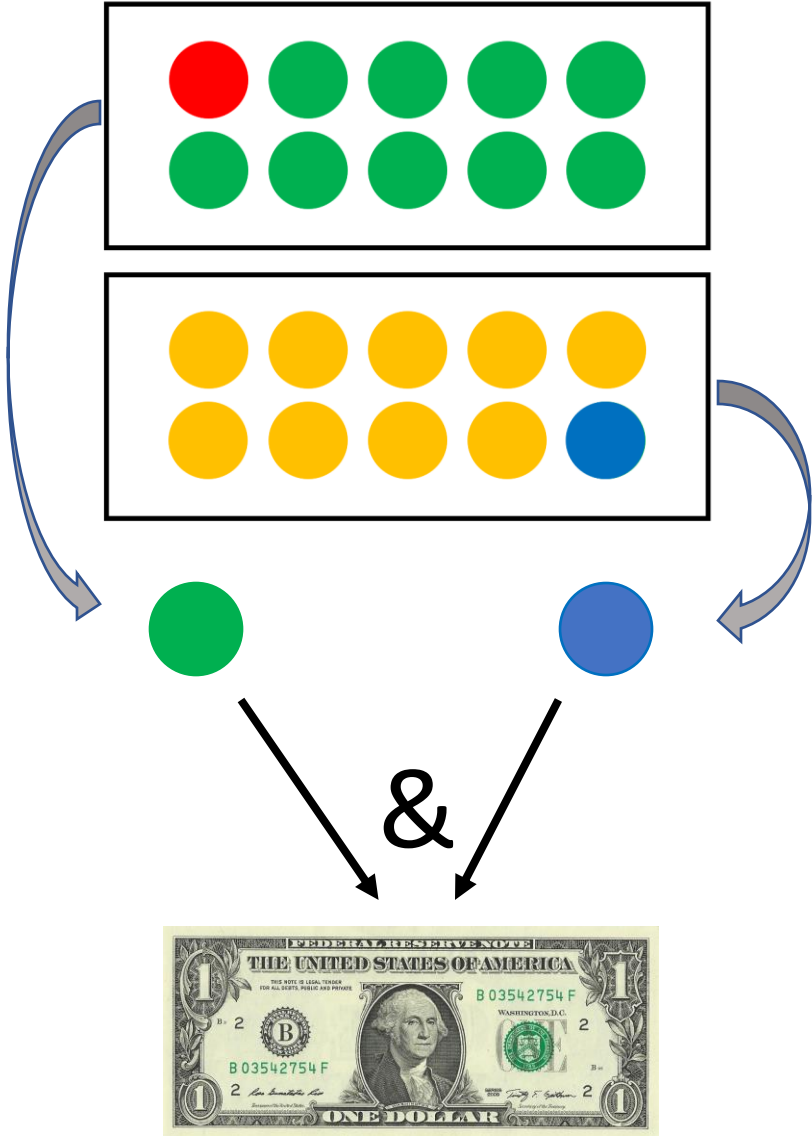
- To judge whether C caused E, people:

‘sample’ counterfactuals from the set of possible outcomes

compute the correlation between C and E across these counterfactuals

# Sample counterfactuals by mental simulation

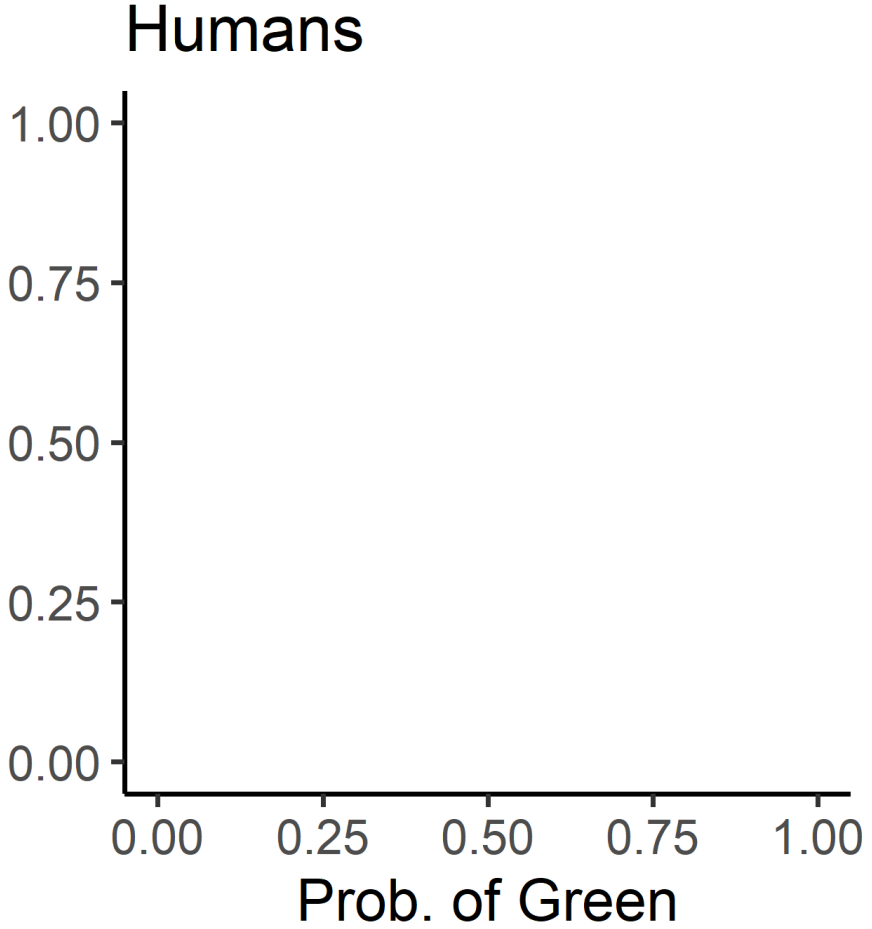
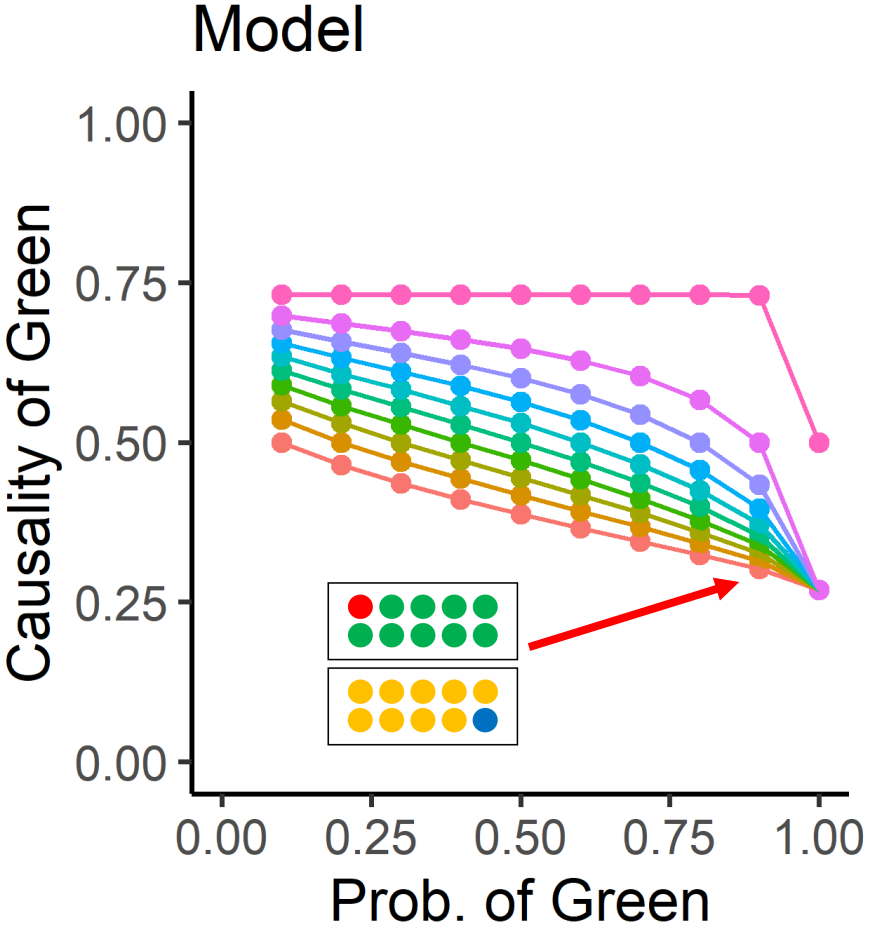
Ball from top box	Ball from bottom box	Outcome
		<del></del>
		<del></del>
		
		<del></del>
		<del></del>
		<del></del>
		<del></del>
























# Counterfactual effect size model

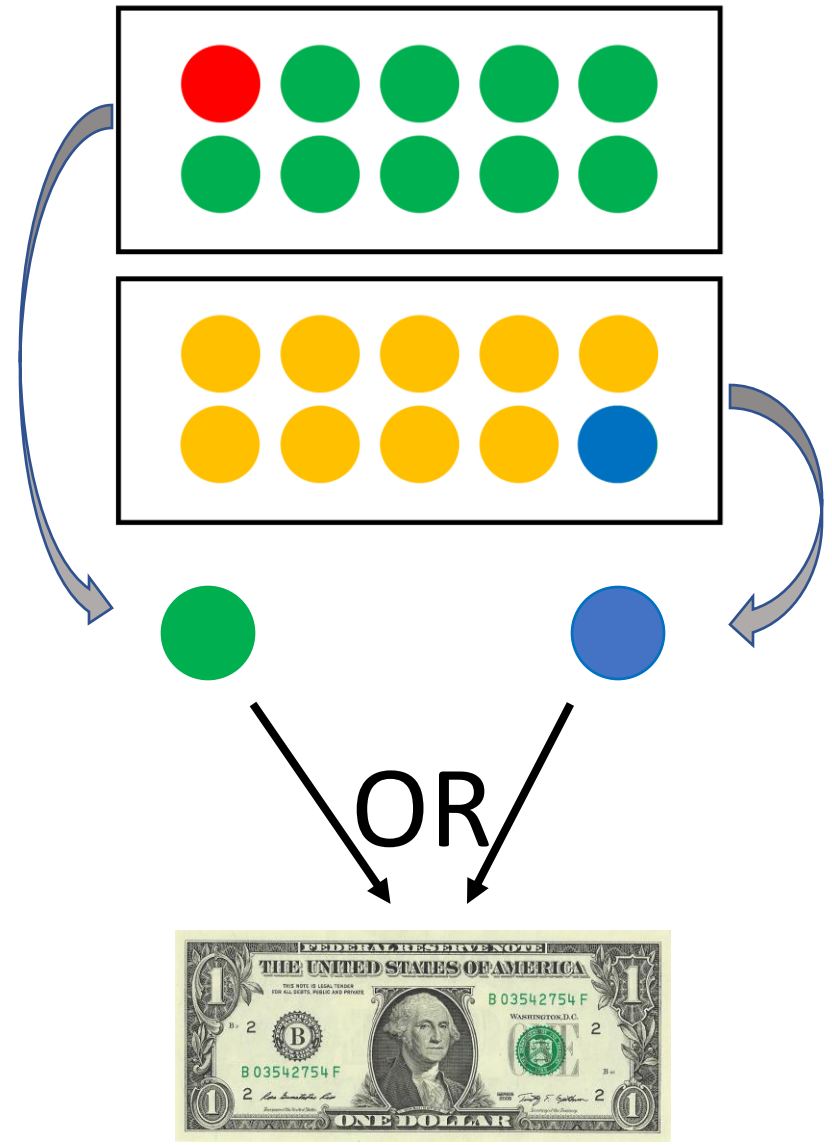
Prob. of Blue

● 0.1	● 0.3	● 0.5	● 0.7	● 0.9
● 0.2	● 0.4	● 0.6	● 0.8	● 1



$r = .89$   
 Data from Exp 1  
 in Morris et al.,  
 2019, PLoS One

Ball from top box	Ball from bottom box	Outcome
		
		
		
		
		<del></del>
		
		

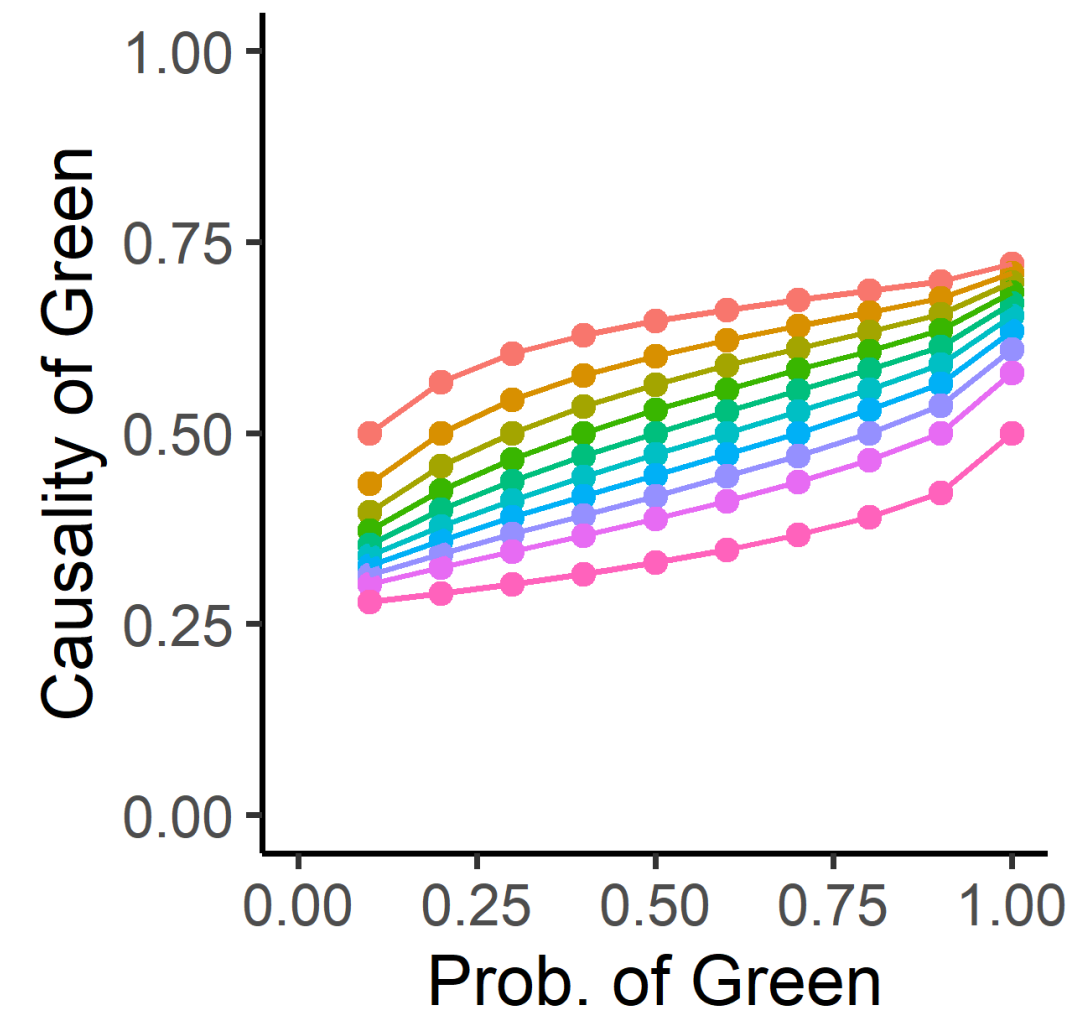




OR  
Structure



Model



Quillien, 2020

Data from  
Morris et al.,  
2019

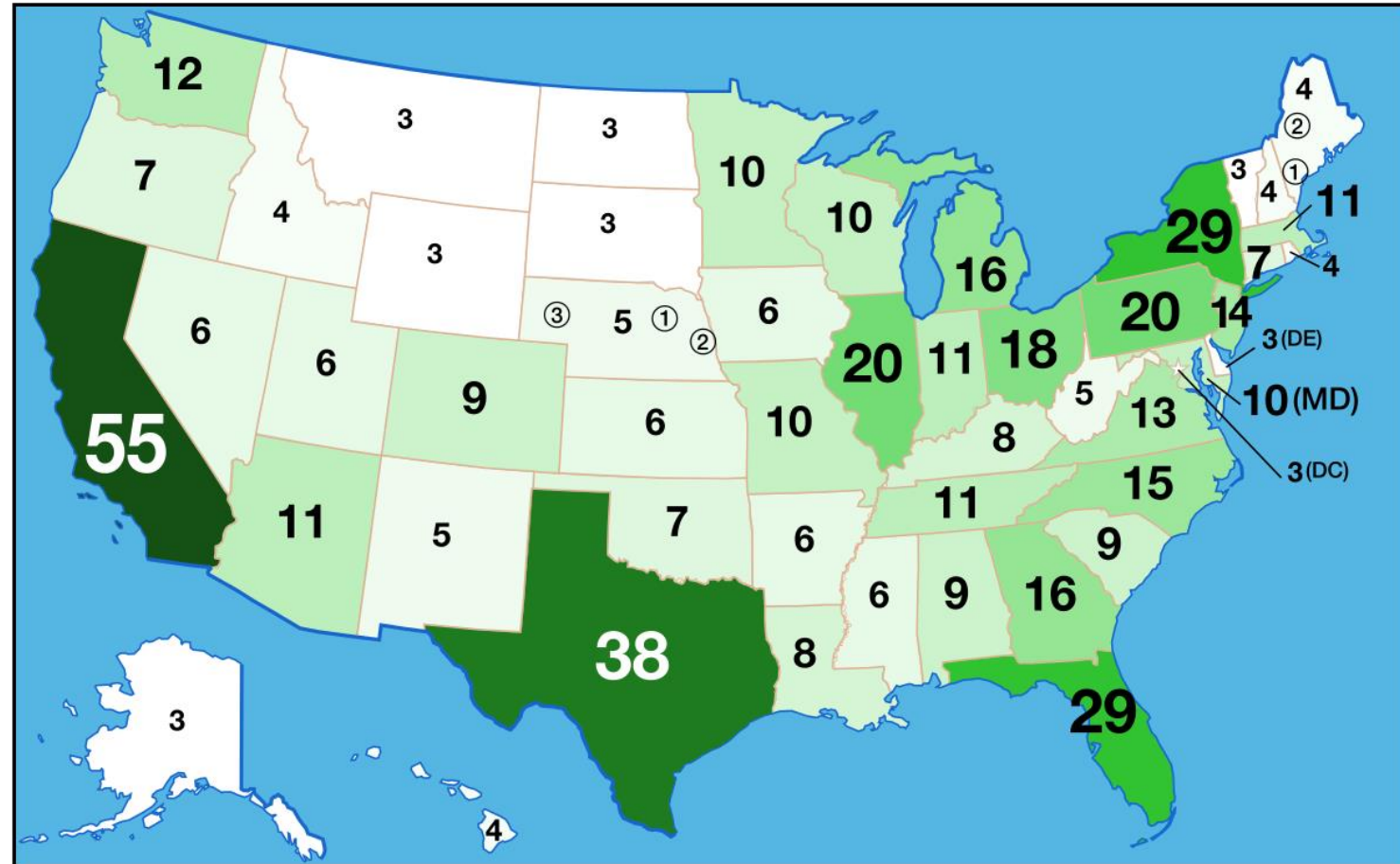


- The definition of “correlation” used by the model is slightly different than the ordinary statistical notion
- (Winning the prize is correlated with drawing a green ball, but does not cause it)
- See optional online readings for more details on the “interventionist” definition of correlation used by the model

# Testing the model with a real-world example



Which state caused Biden to win the election?



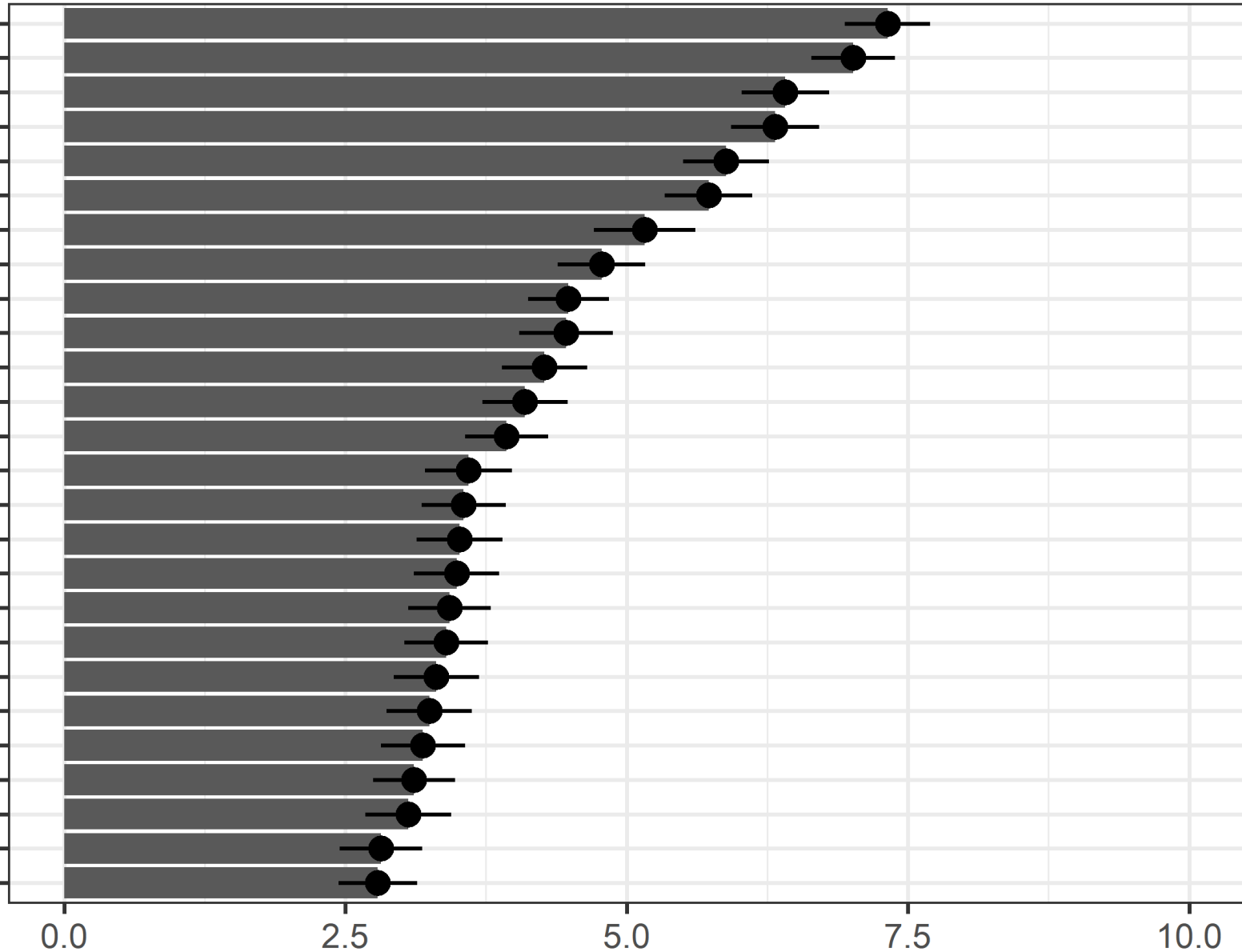
# Biden won the presidency because he won...

Average  
human  
judgments

N=207

States Biden won

- Pennsylvania (PA)
- Georgia (GA)
- Arizona (AZ)
- Michigan (MI)
- Wisconsin (WI)
- Nevada (NV)
- California (CA)
- Minnesota (MN)
- Virginia (VA)
- New York (NY)
- Illinois (IL)
- New Mexico (NM)
- Colorado (CO)
- Massachusetts (MA)
- New Jersey (NJ)
- Washington (WA)
- Oregon (OR)
- Maryland (MD)
- Maine (ME)
- Connecticut (CT)
- New Hampshire (NH)
- Delaware (DE)
- Vermont (VT)
- Washington, D.C.
- Rhode Island (RI)
- Hawaii (HI)

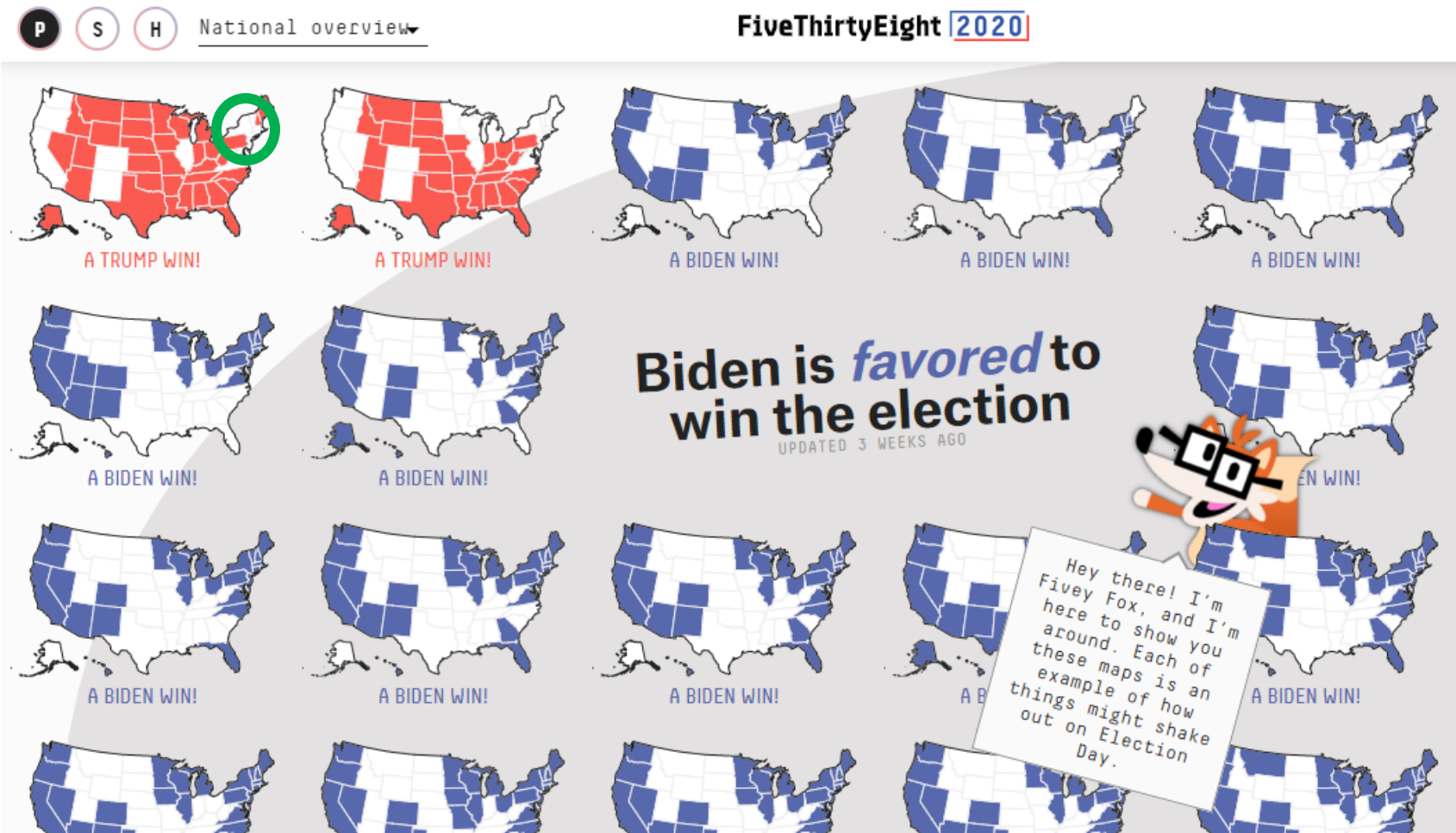


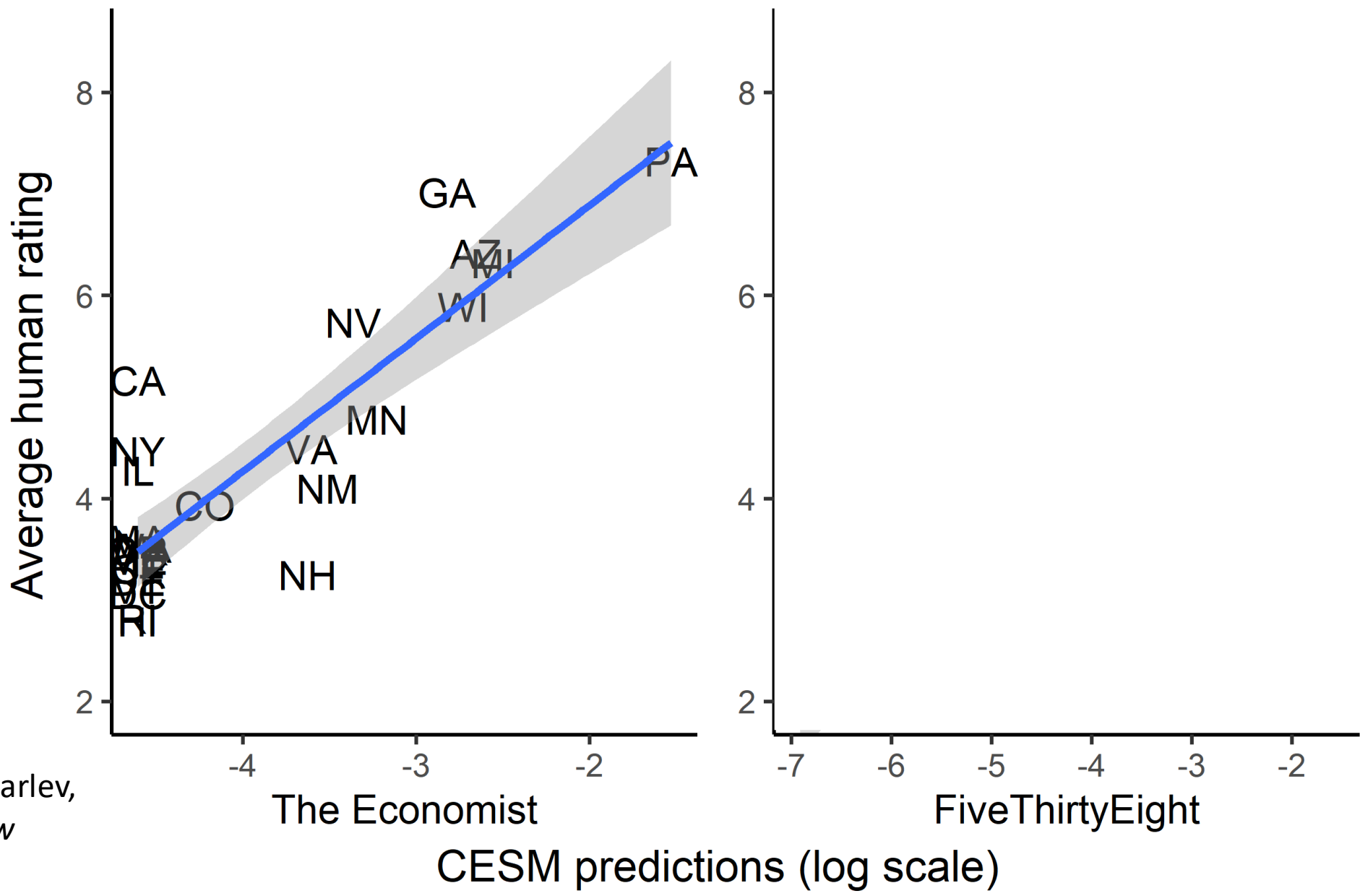
Ratings (0 = Do not agree at all, to 10 = Agree very strongly)

Quillien & Barlev,  
*under review*

# Model

- To compute the “causal strength” of the state of New York:
- Take the correlation, across all simulations, between “Biden wins in New York”, and “Biden wins the presidency”





Quillien & Barlev,  
under review

The Economist

CESHM predictions (log scale)

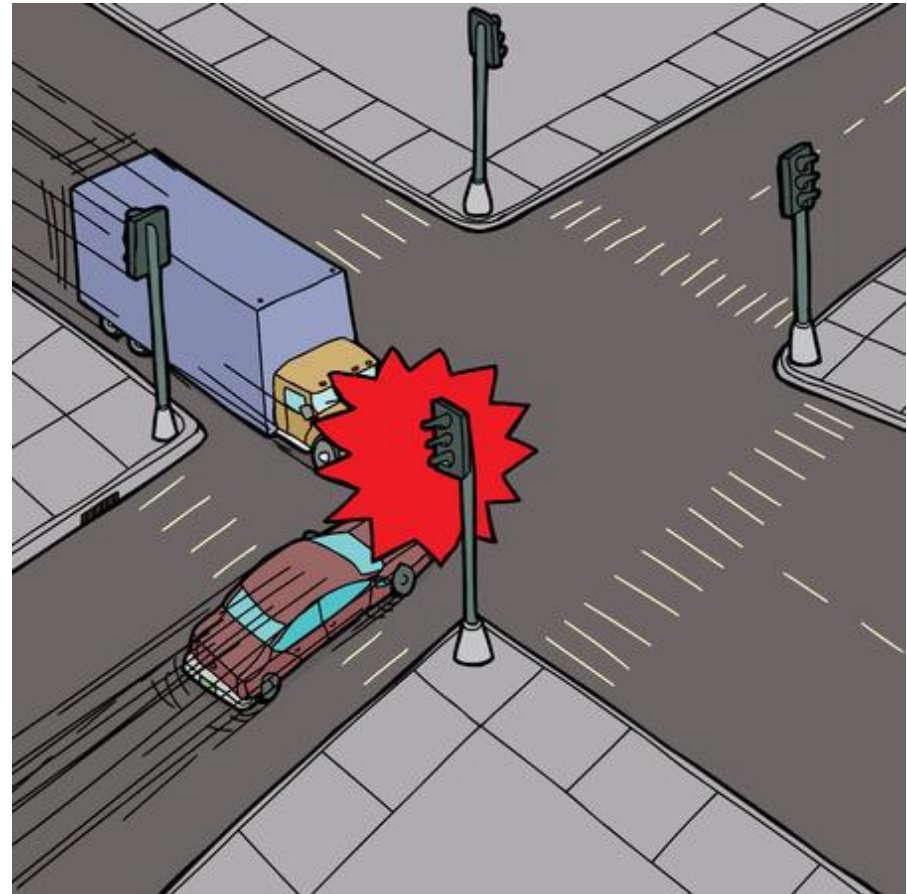
FiveThirtyEight

# Morality and actual causation (Hitchcock & Knobe, 2009)

Who caused the collision?

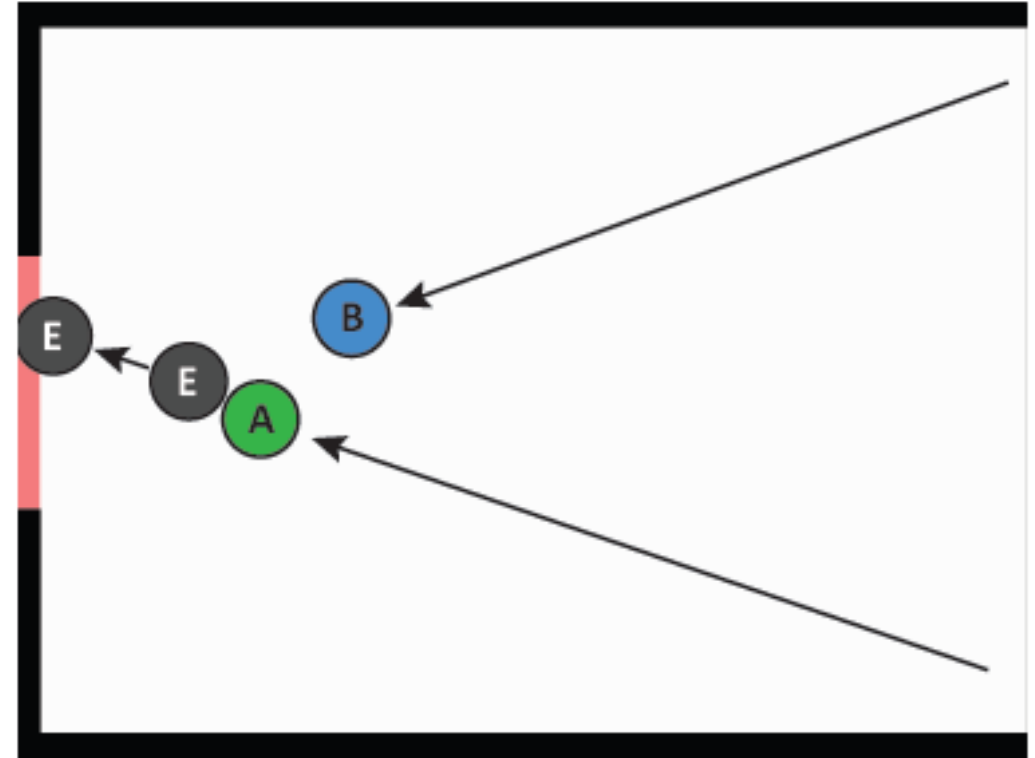
Counterfactuals are biased toward situations where people don't violate norms.

Across counterfactuals, the behavior of the car is more highly correlated with the collision



# Outstanding mysteries

- Did the green ball cause the black ball to go through the gate?
- Did the blue ball cause the black ball to go through the gate?
- Across counterfactuals, there is a correlation between the blue ball's presence and the black ball going through the gate -> incorrect causal attribution



*(this is called a case of "causal pre-emption")*

# Ongoing research questions

- How exactly do people sample counterfactuals?
- Does the way that judges attribute causal responsibility match our intuitive notion of cause?
- Does our intuitive notion of actual cause shape the way we use other concepts?
- etc



# References

- Lewis, D. (1973). Causation. *The journal of philosophy*, 70(17), 556-567.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.
- Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106(11), 587-612.
- Quillien, T. (2020). When do we think that X caused Y?. *Cognition*, 205, 104410.
- Quillien, T., & Barlev, M. (under review). Causal judgment in the wild: evidence from the 2020 US presidential election.