Computational Cognitive Science

Actual causation

Guest lecturer: Tadeg Quillien

Chancellor’s Fellow

Department of Psychology, University of Edinburgh

In previous lectures: causal inference

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

Wood

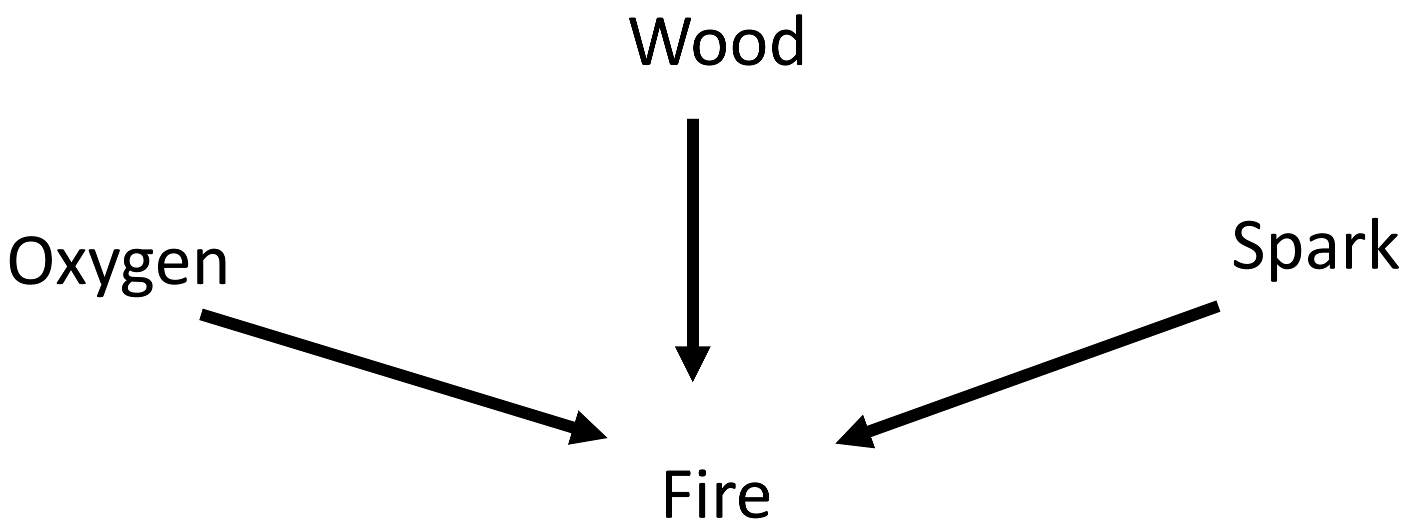
Oxygen

Spark

Fire

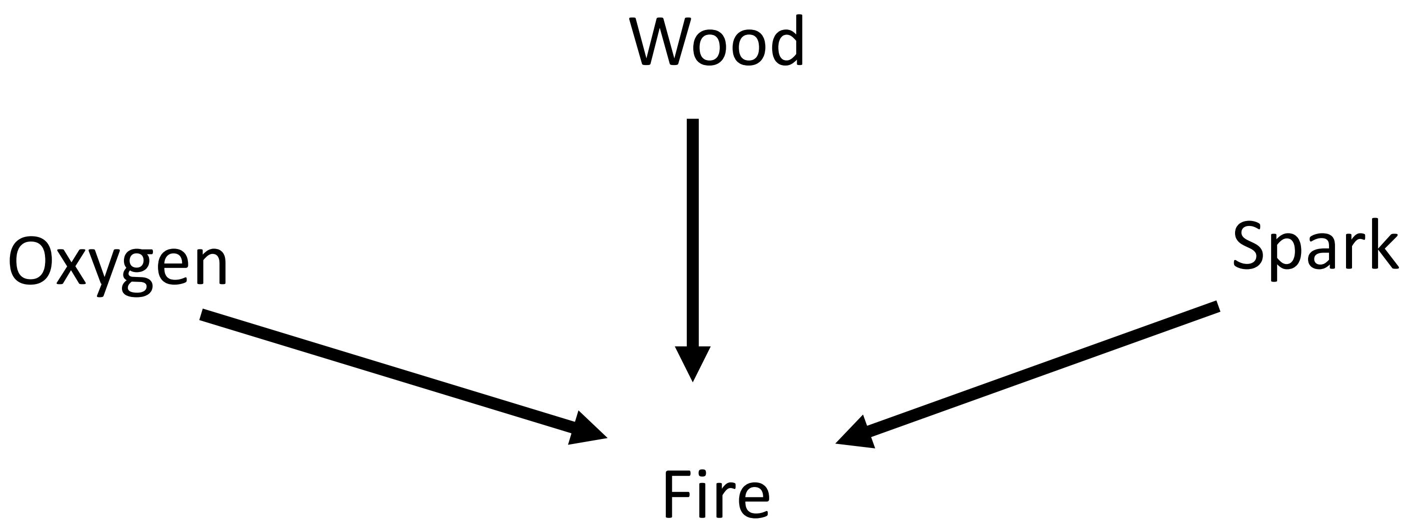
In previous lectures: causal inference

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?

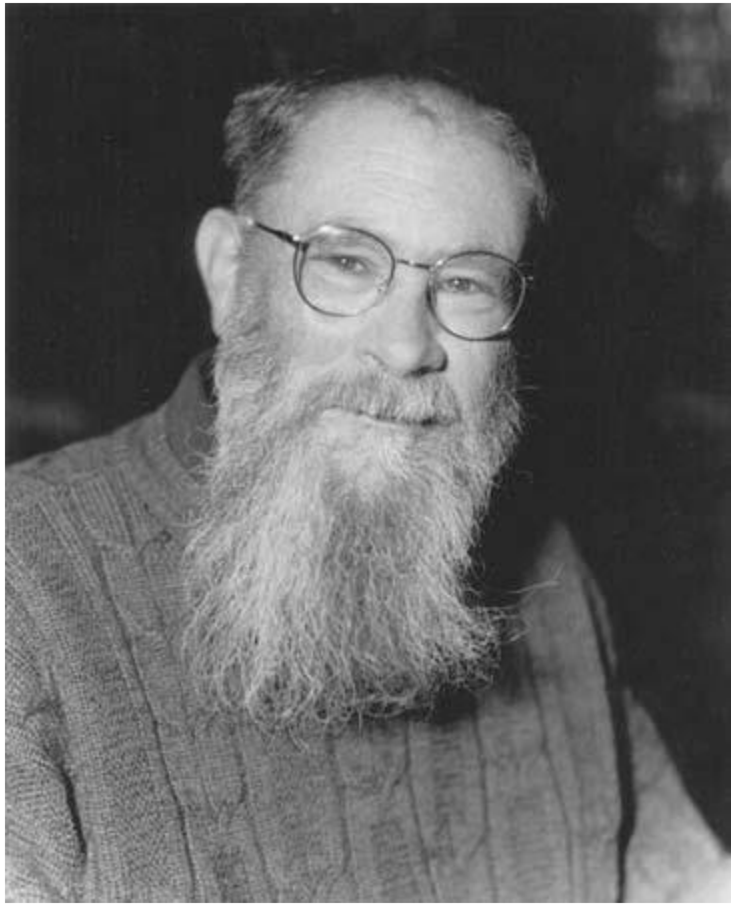


Counterfactual theory of causation (e.g. David Lewis)

- 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

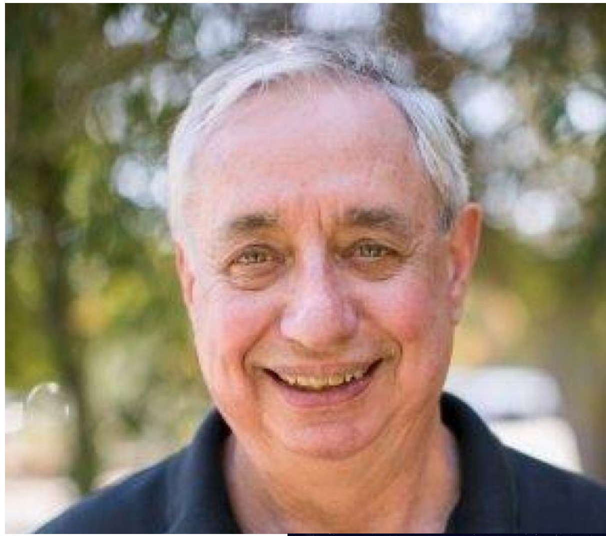
- -> None of the soldiers caused the prisoner's death!

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



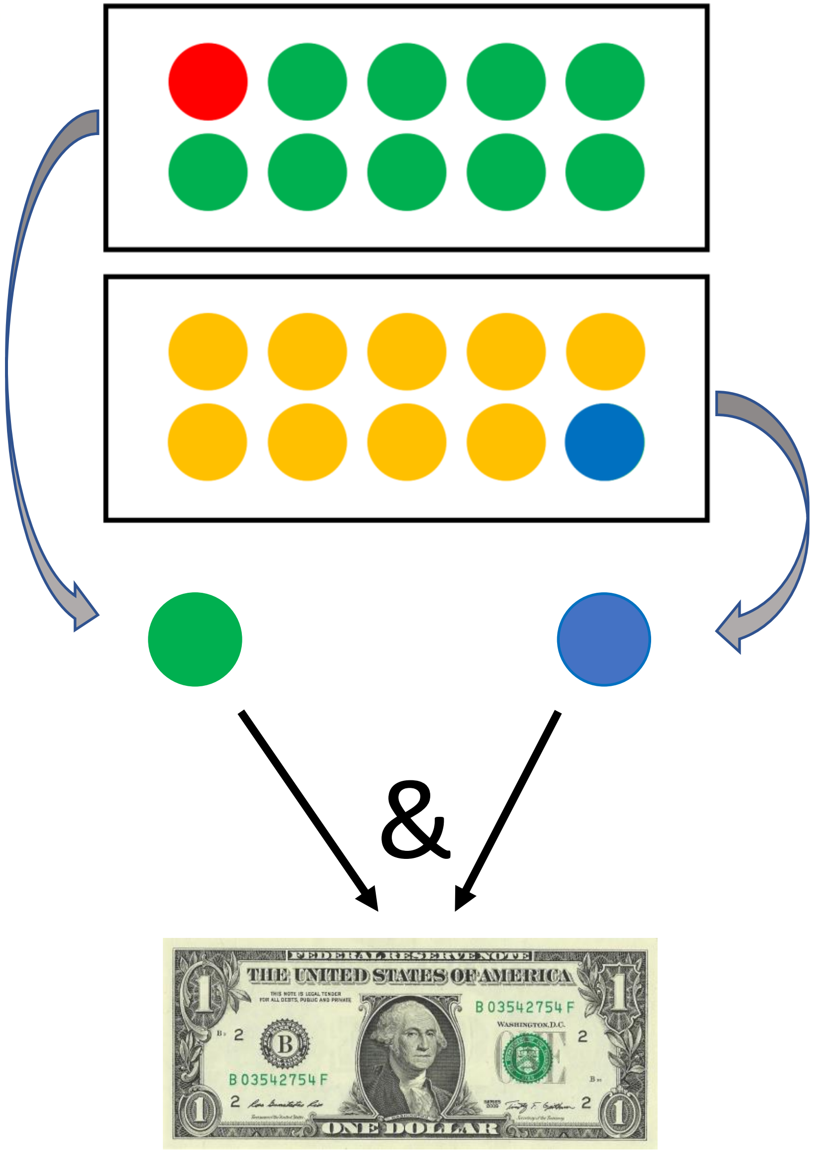
Saving the counterfactual theory: “invariant” counterfactual dependence (Jim Woodward)

- 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 you win a dollar because you drew a green ball, or because you 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:

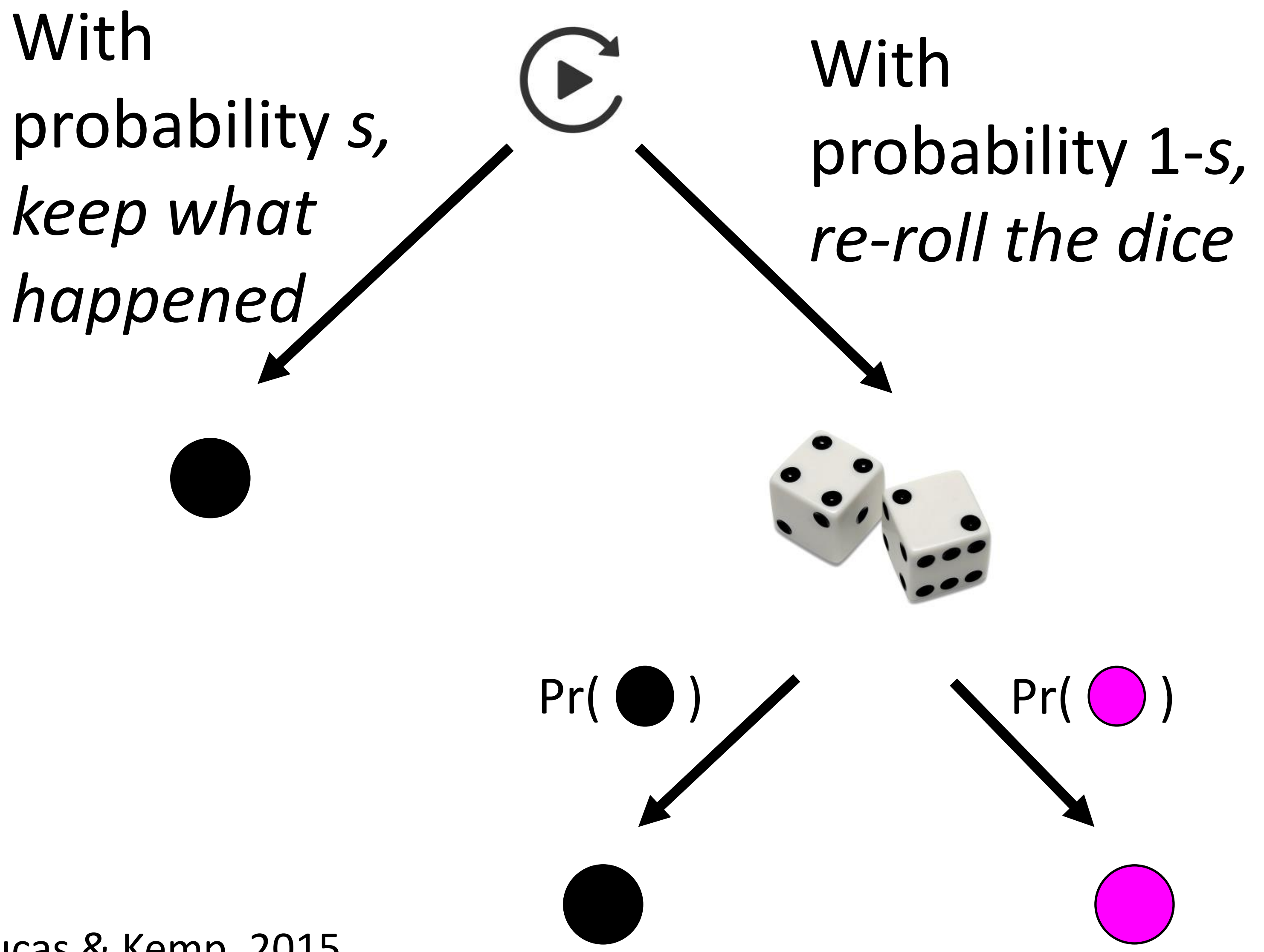
  ‘sample’ counterfactuals from the set of possible outcomes

  Quantify the average causal effect of C on E across counterfactuals

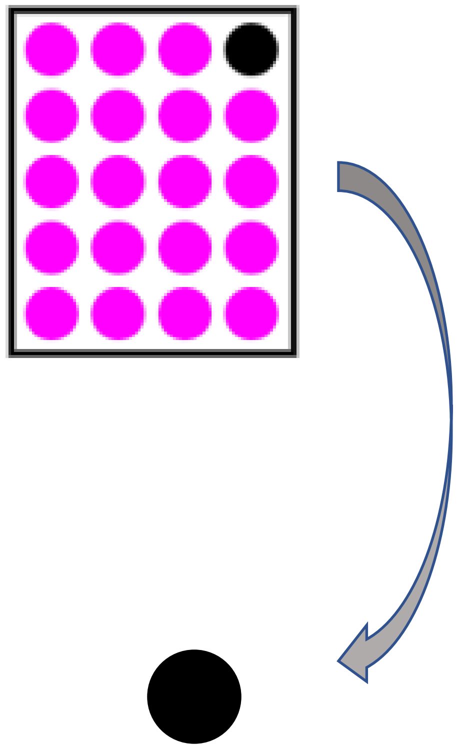
Sampling counterfactuals

- We assume people sample from a probability distribution S over possible worlds.

- This distribution is inspired by past research on counterfactual reasoning.



What happened in the actual world



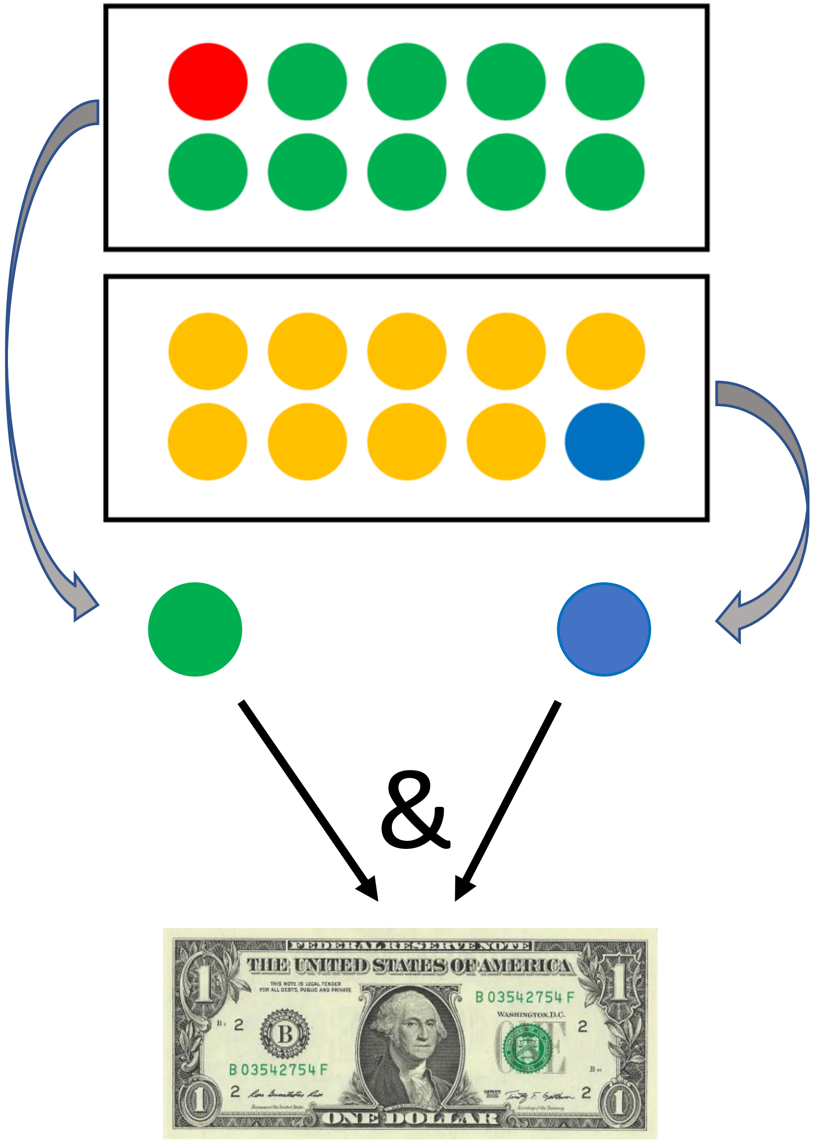
(Lucas & Kemp, 2015, *Psychological Review*)

Computing an average causal score from this distribution

- Average causal score: $S(E|do(C))- S(E|do(\neg C))$
  → This is the causal equivalent of a regression coefficient

Sample counterfactuals by mental simulation

Ball from top box	Ball from bottom box	Outcome



Here we have:

$S(E|do(G))- S(E|do(\neg G)) = 1/4$

$S(E|do(B))- S(E|do(\neg B)) = 3/4$

Computing an average causal score from this distribution

- Average causal score: $S(E|do(C)) - S(E|do(\neg C))$

  → This is the causal equivalent of a regression coefficient

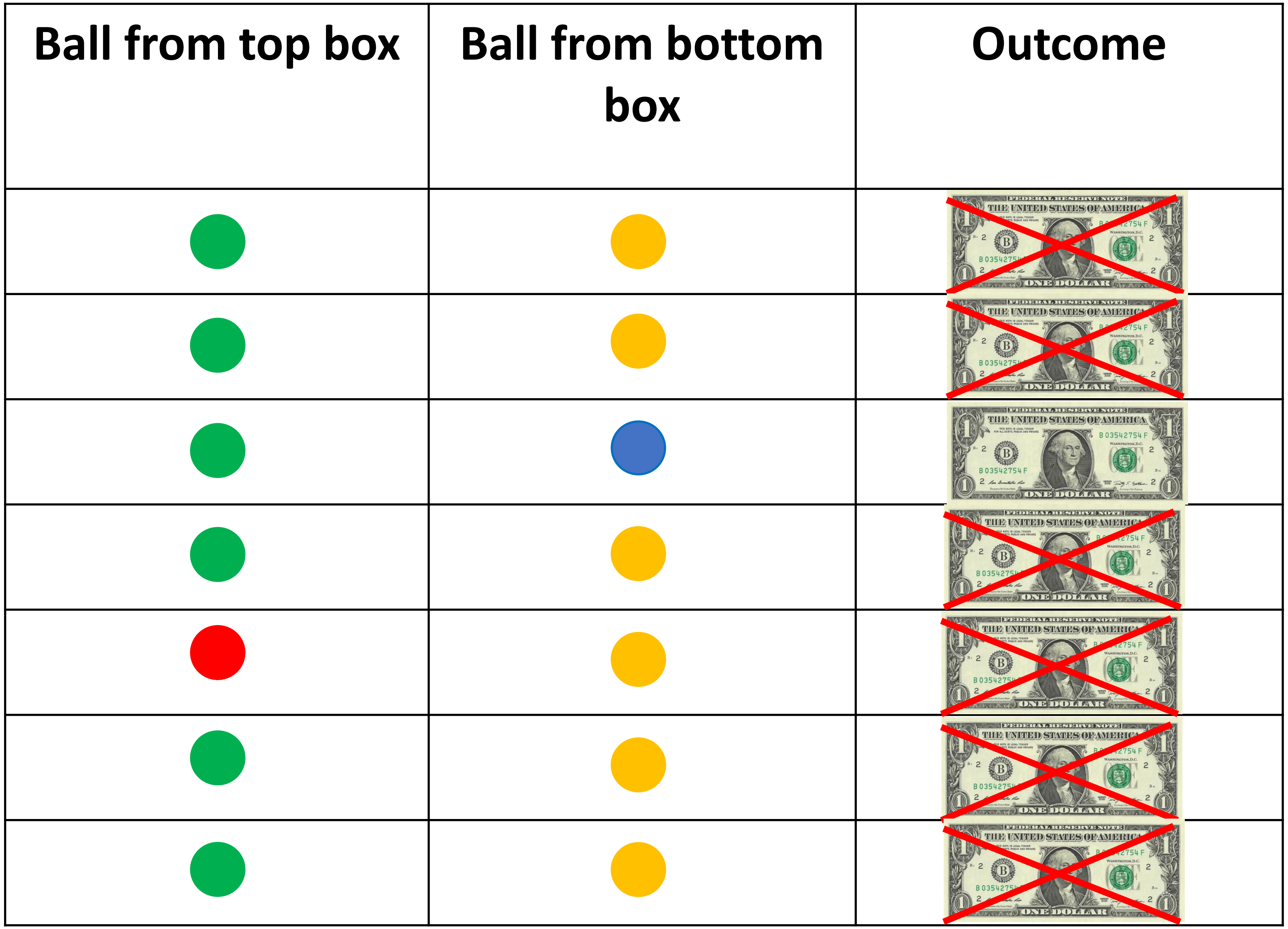
- Standardization factor σ_C / σ_E

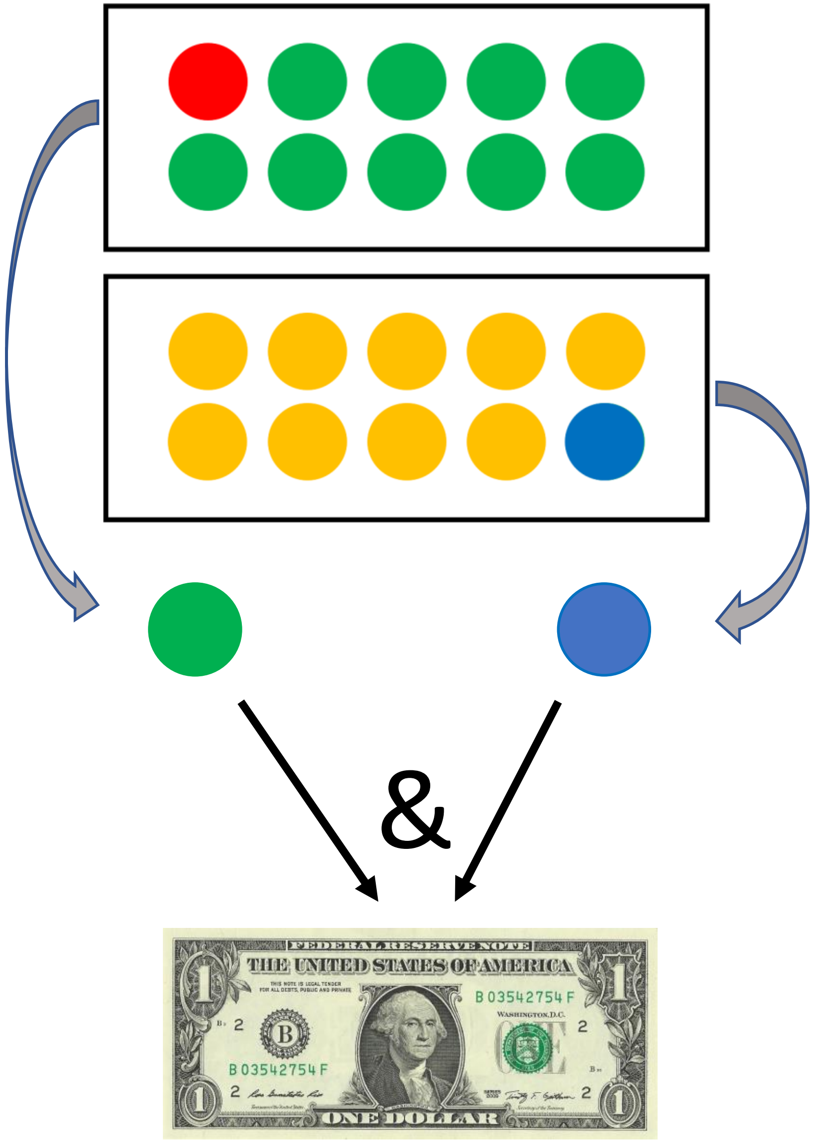
- Causal effect size: Average causal score * Standardization factor

$= [S(E|do(C)) - S(E|do(\neg C))] * (\sigma_C / \sigma_E)$

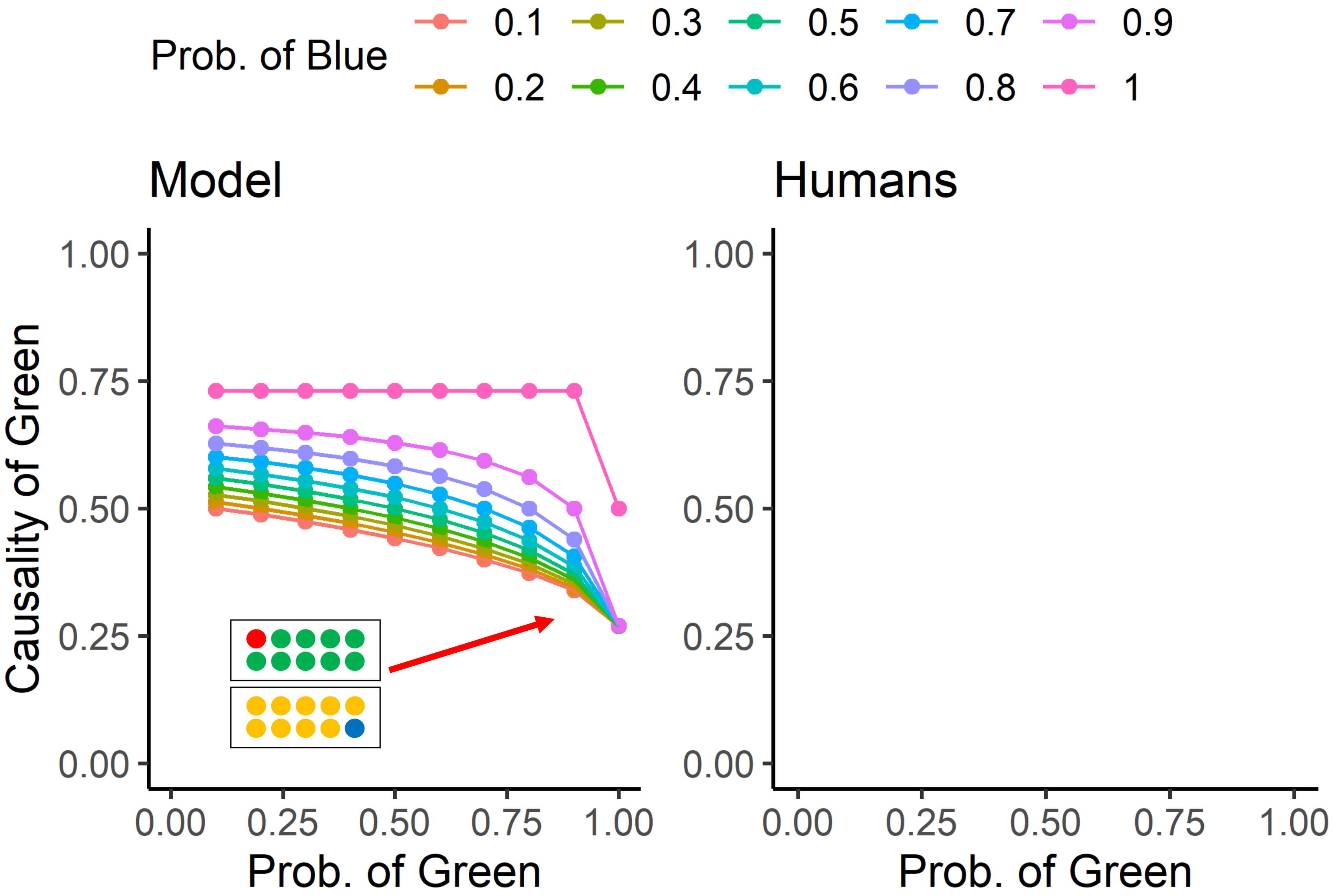
  → This is the causal equivalent of a correlation coefficient!

Sample counterfactuals by mental simulation





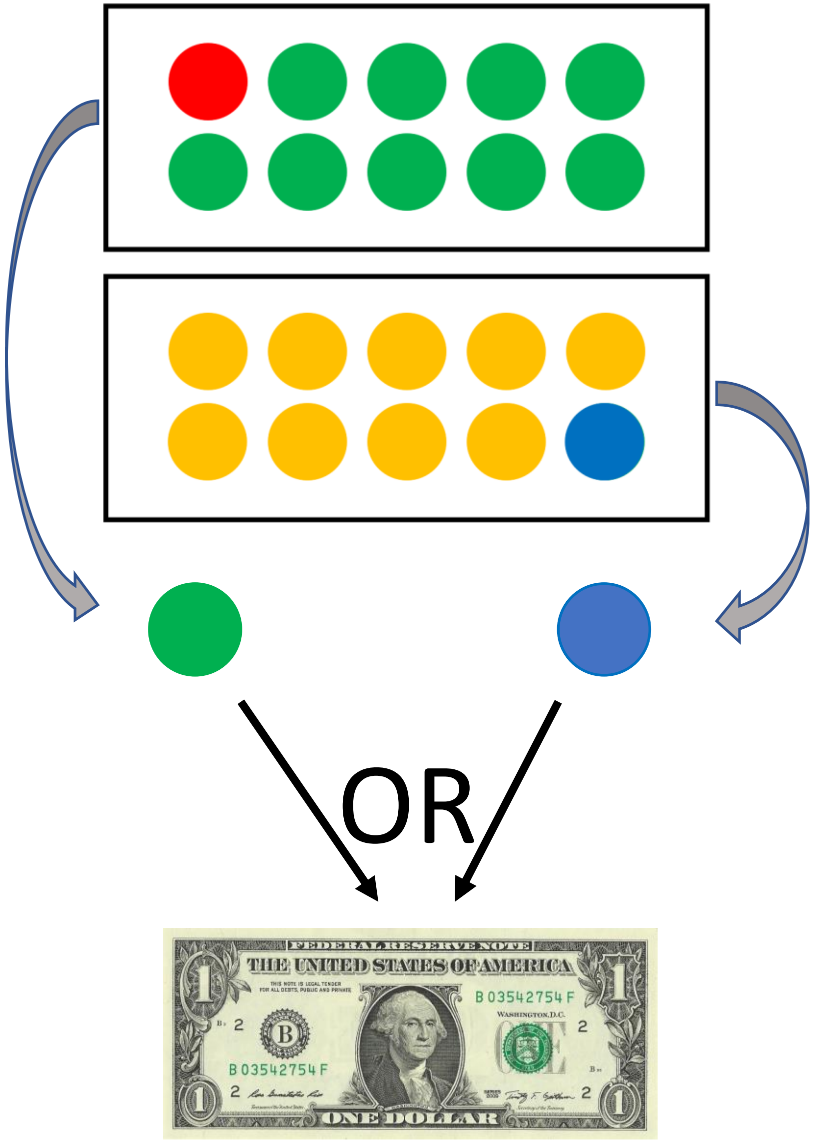
Counterfactual effect size model



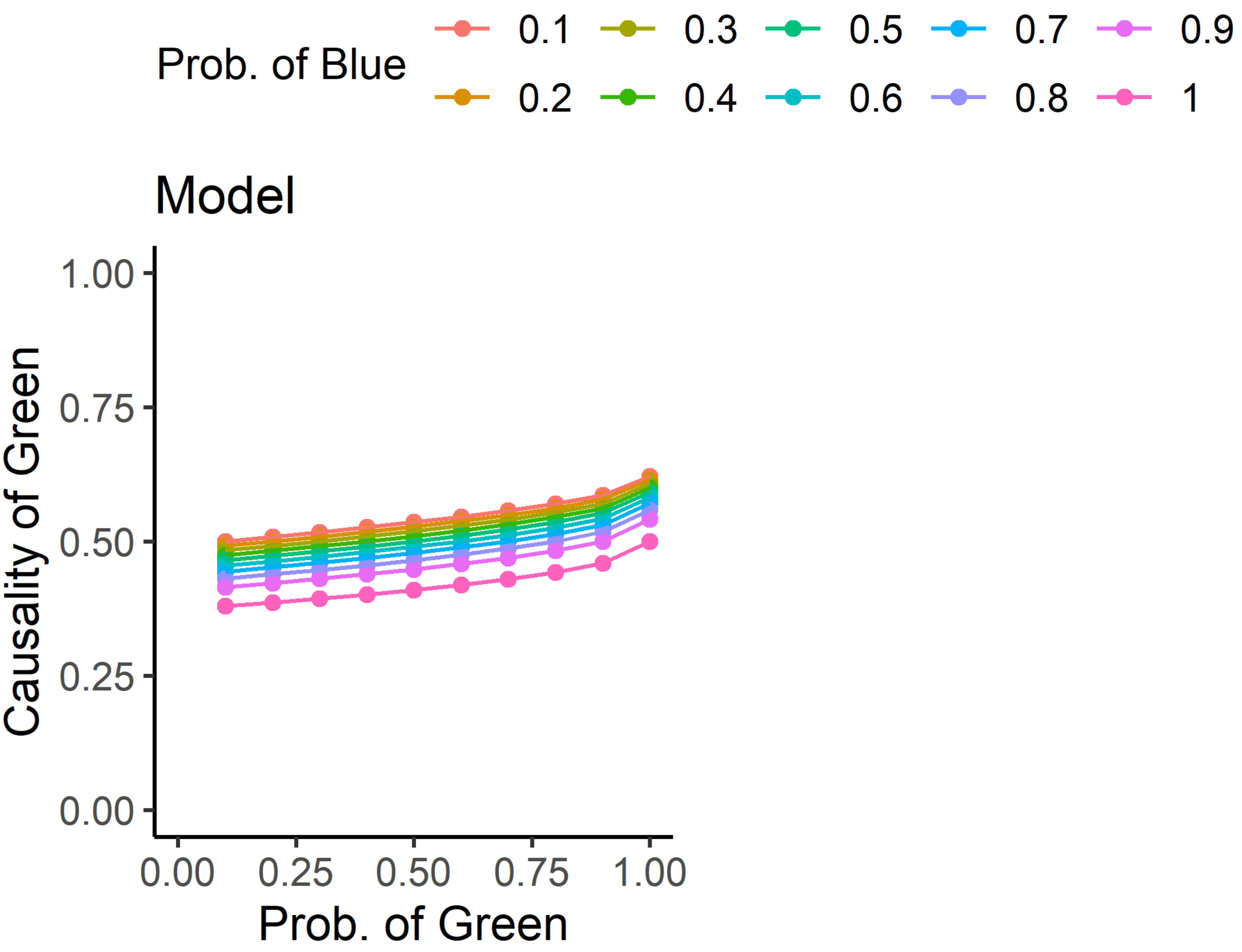
r = .89

Data from Exp 1 in Morris et al., 2019, PLoS One

Ball from top box	Ball from bottom box	Outcome



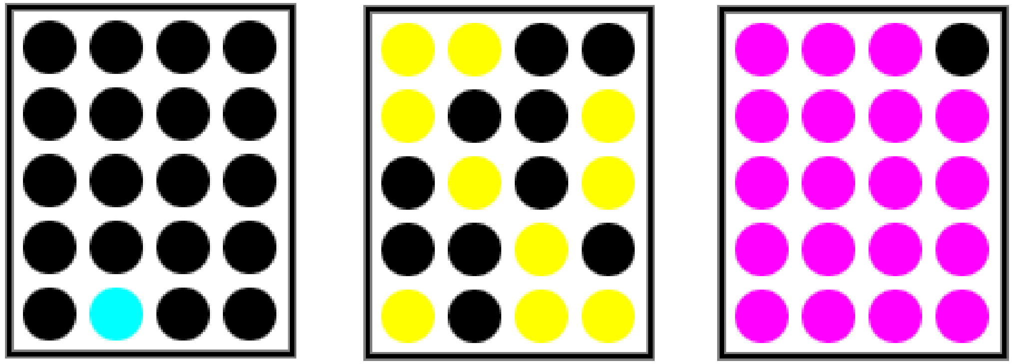
OR
Structure



Data from
Morris et al.,
2019

New experiment (Quillien & Lucas, 2023)

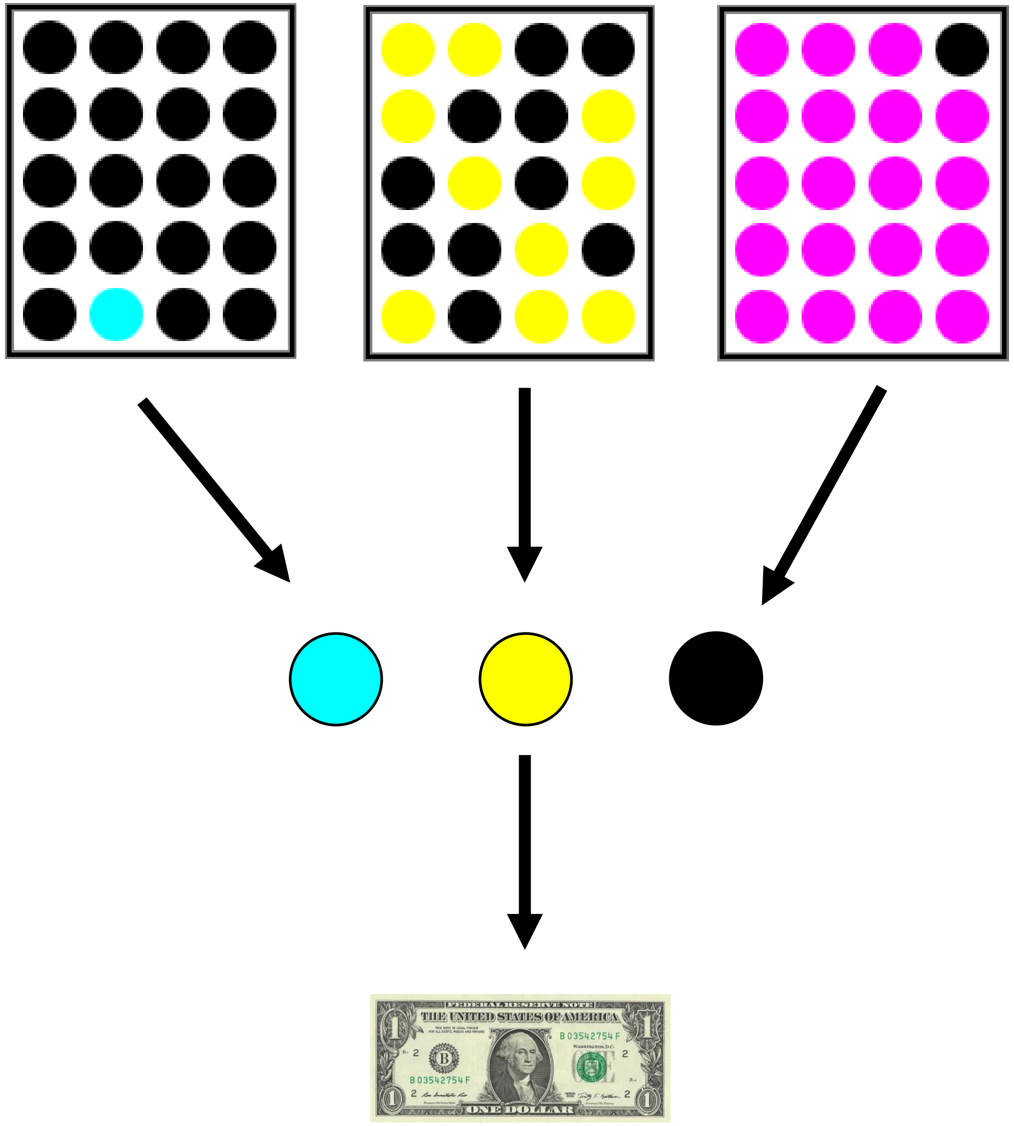
- Causal judgments should be sensitive to:
  - The prior probability of events
  - The details of what actually happened
- We predict an *interaction* between the two



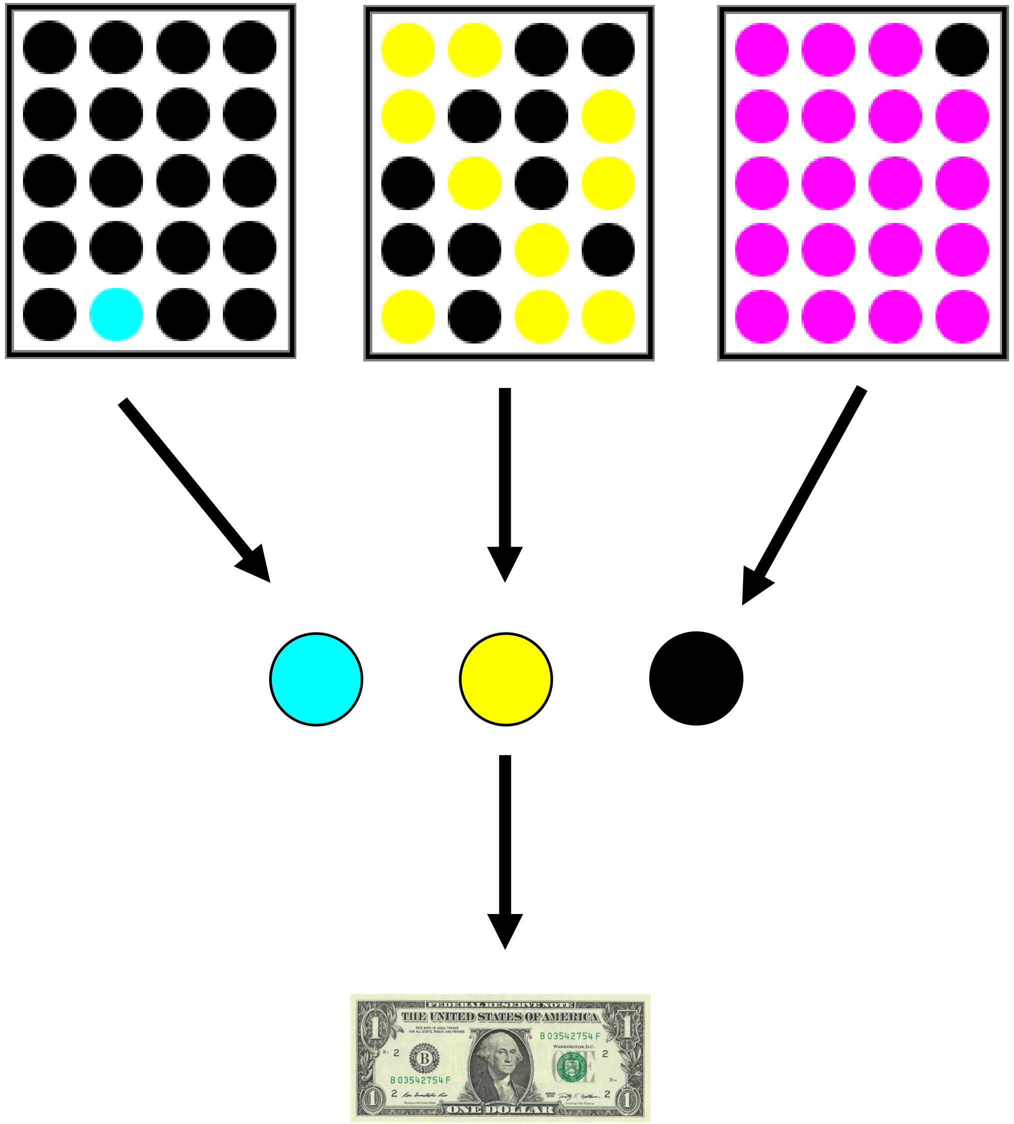
2 colored
balls or more

Did you win because you drew a blue ball?

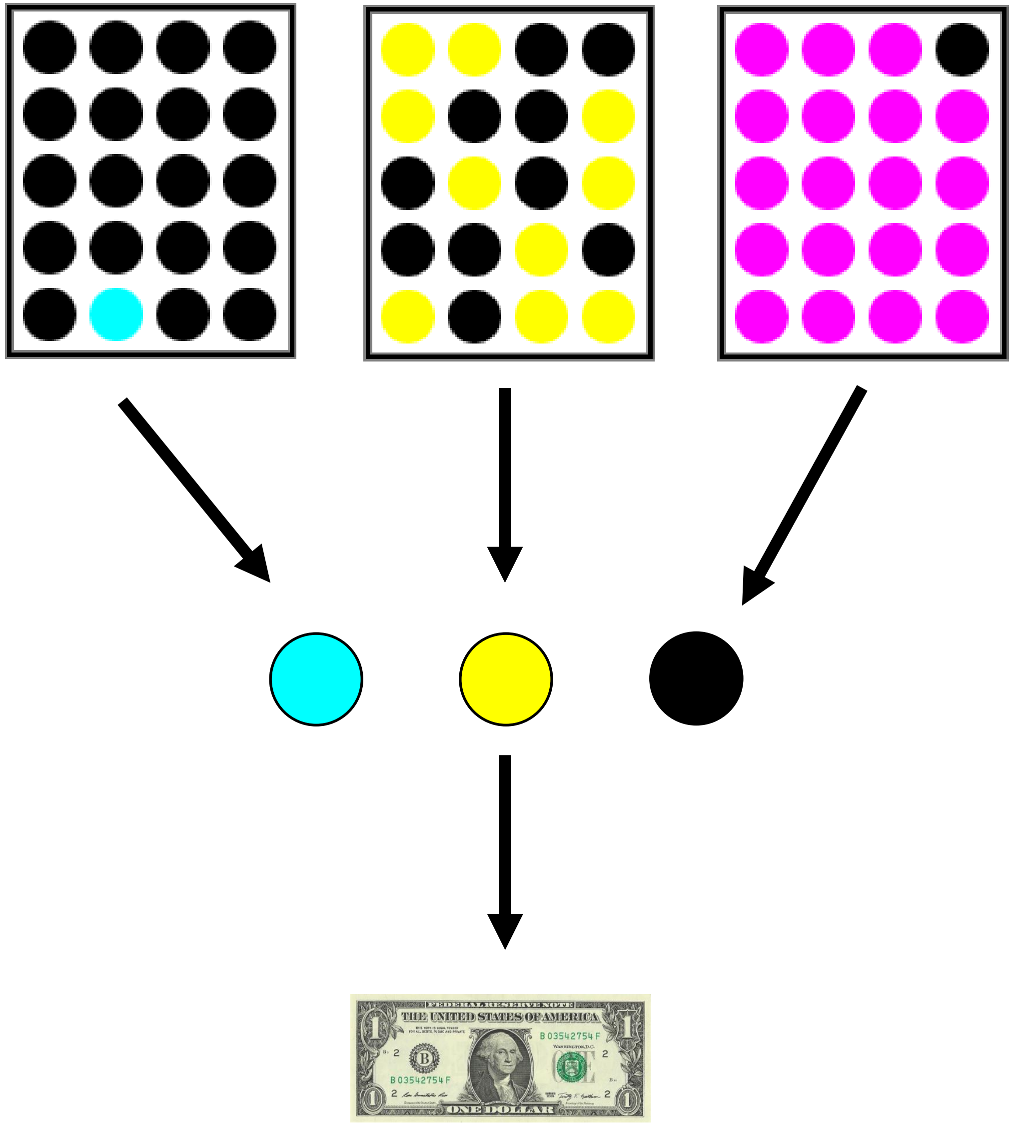
Because you drew a yellow ball?



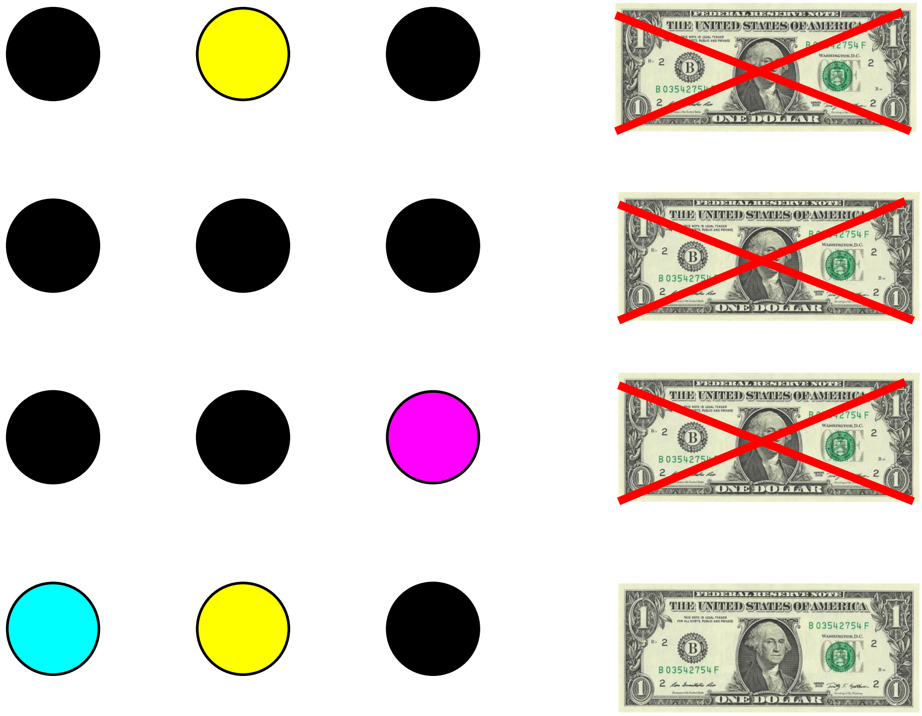
Actual World



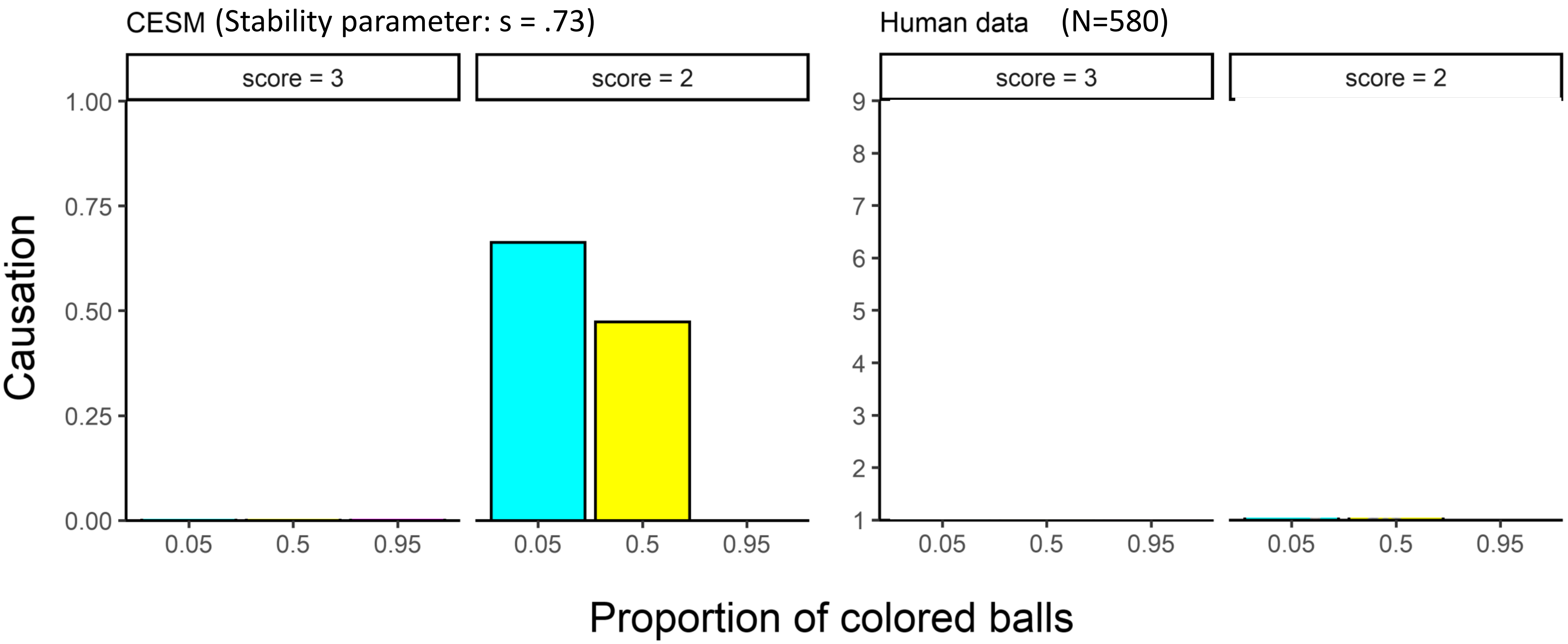
Actual World



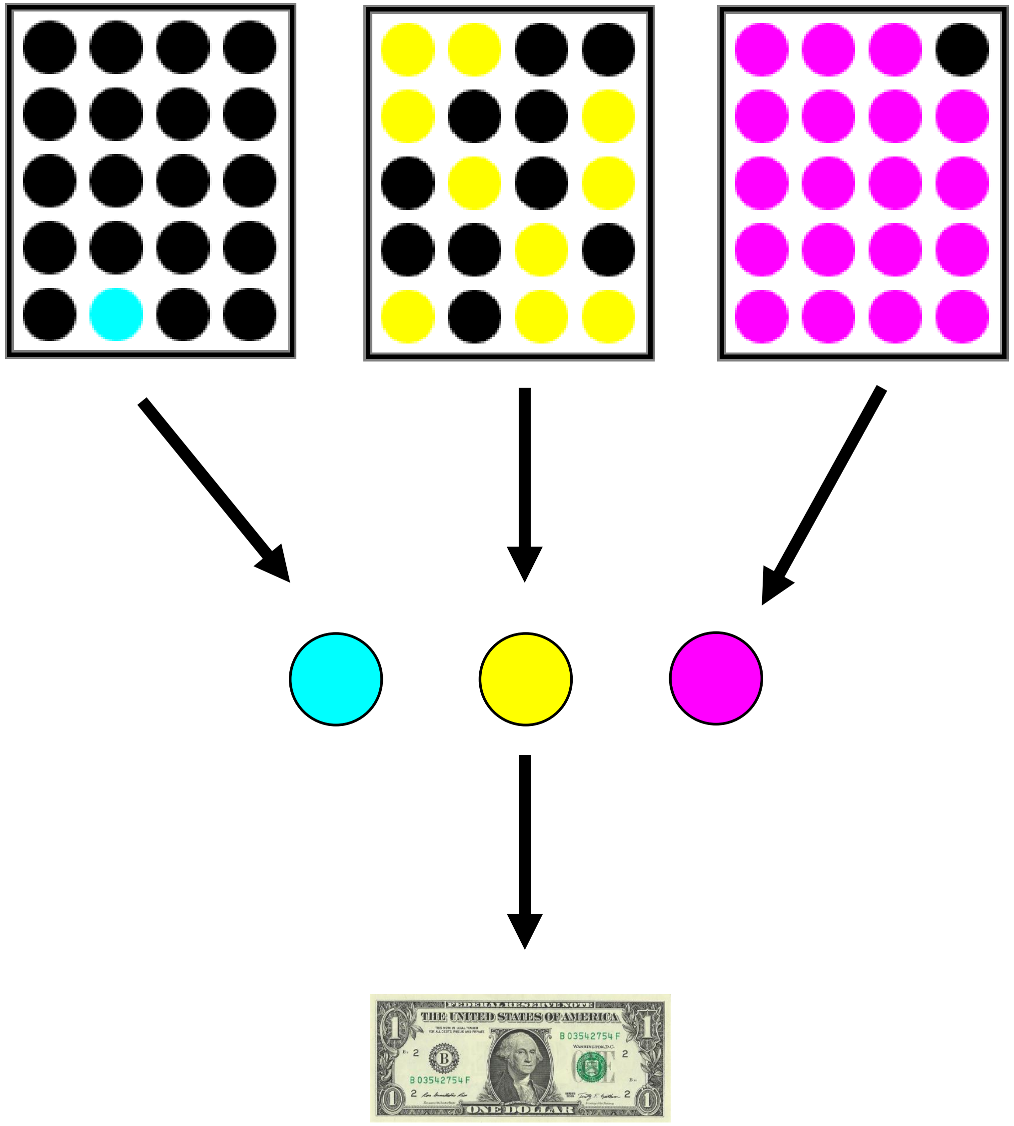
Counterfactuals



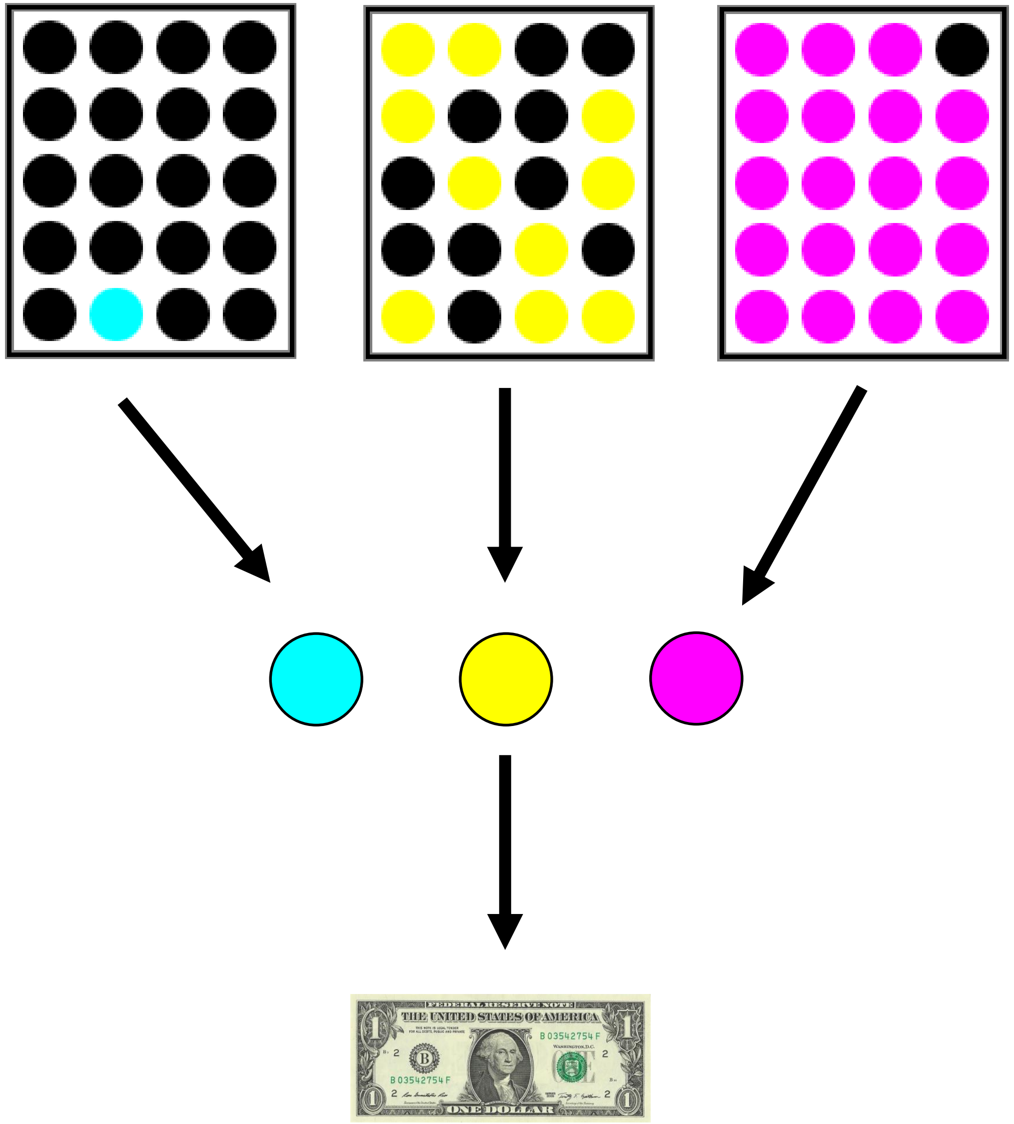
etc

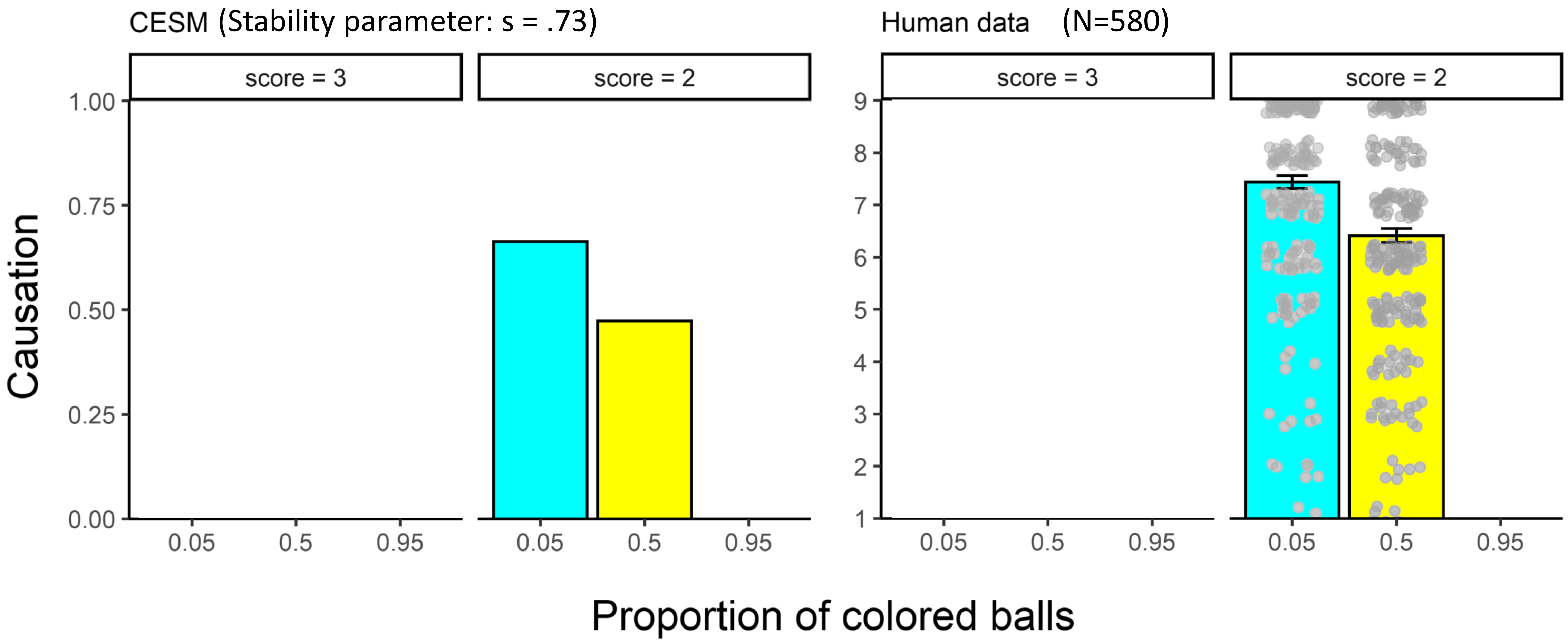

CESM (Stability parameter: s = .73)
Human data (N=580)
score = 3
score = 2
score = 3
score = 2
Causation
1.00
0.75
0.50
0.25
0.00
9
8
7
6
5
4
3
2
1
0.05
0.5
0.95
Proportion of colored balls

Did you win because you drew the blue ball? The yellow ball? The purple ball?



Actual World





CESM (Stability parameter: s = .73)
Human data (N=580)
score = 3
score = 2
score = 3
score = 2
1.00
0.75
0.50
0.25
0.00
9
8
7
6
5
4
3
2
1
Causation
0.05
0.5
0.95
Proportion of colored balls

Ongoing research questions

- What other factors affect the distribution over 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

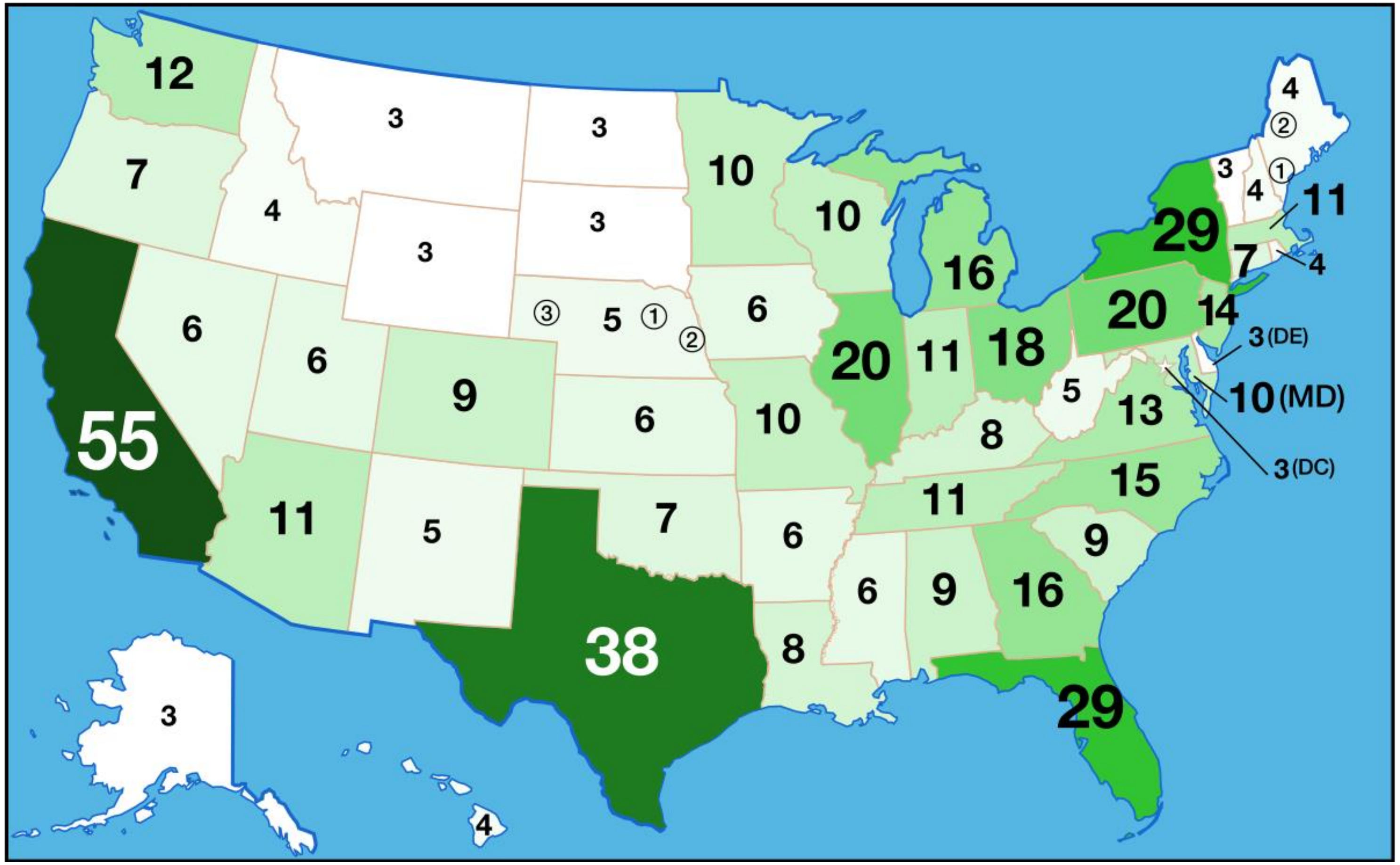
References

- 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.
- Lucas, C., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review.*
- Quillien, T. (2020). When do we think that X caused Y?. *Cognition*, *205*, 104410.
- Quillien, T., & Lucas, C. (2023). Counterfactuals and the logic of causal selection. *Psychological Review.*

Appendix

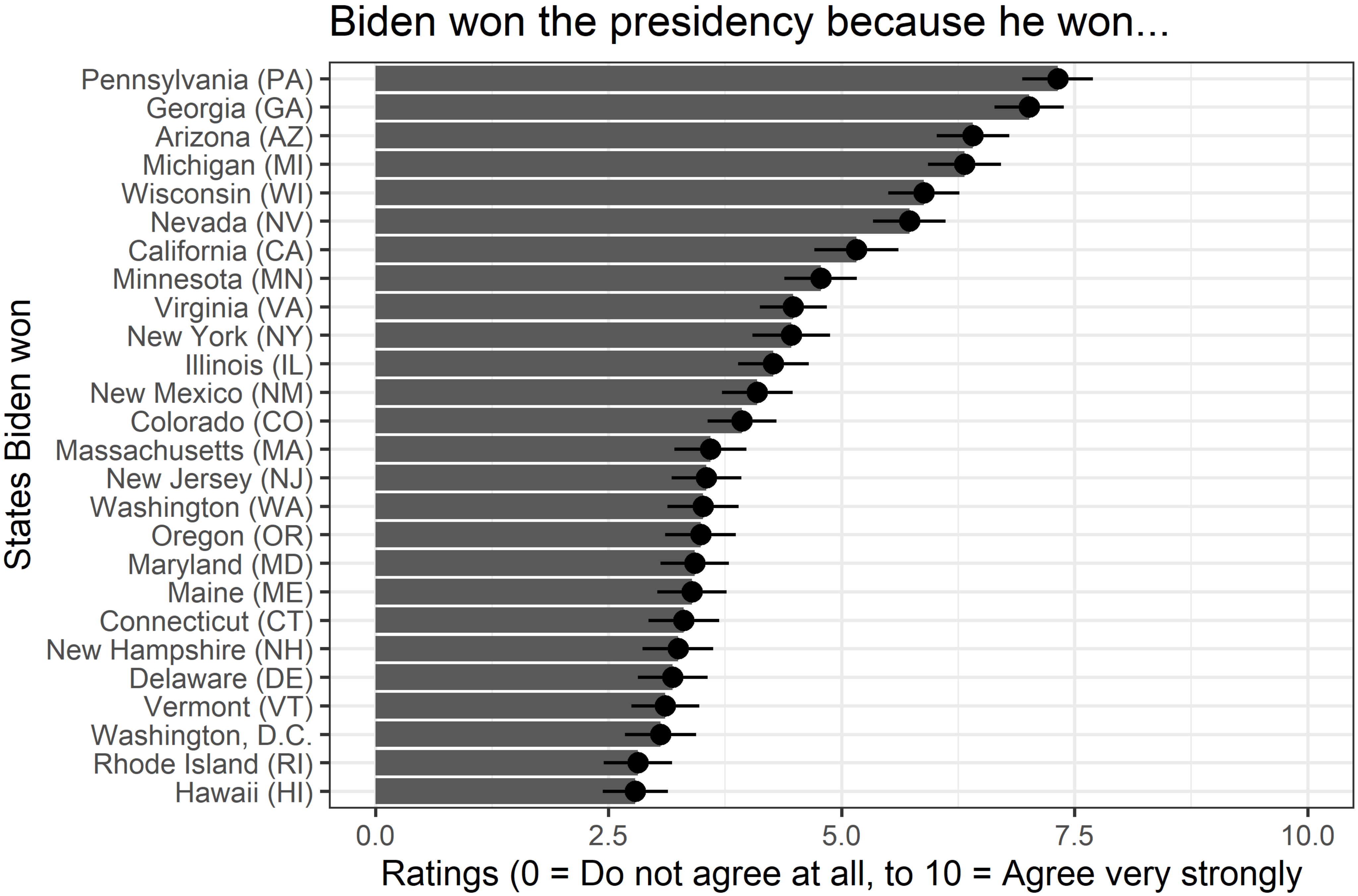
Testing the model with a real-world example

Which state caused Biden to win the election?



Average human judgments

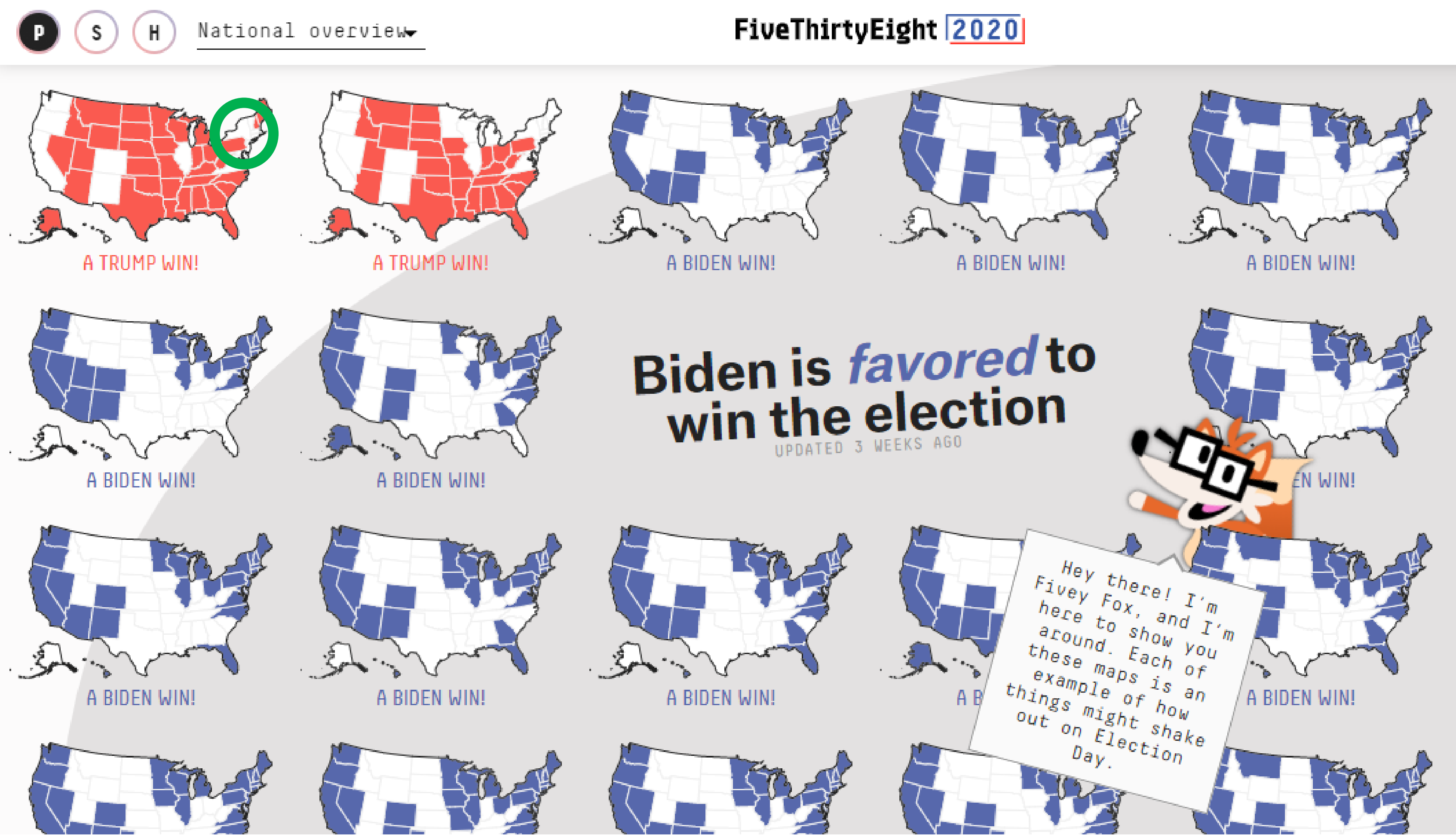
N=207

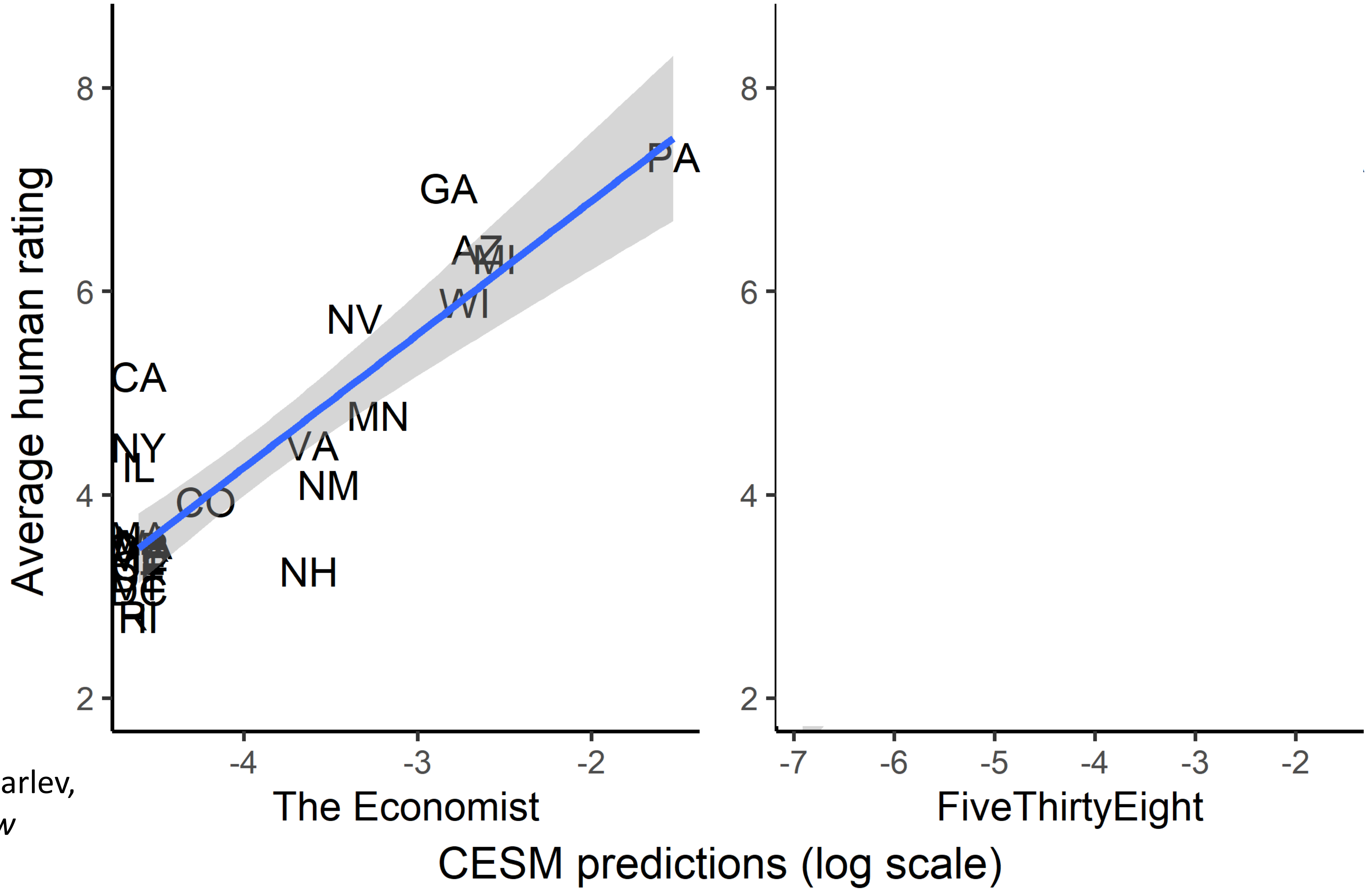


Quillien & Barlev, *under review*

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