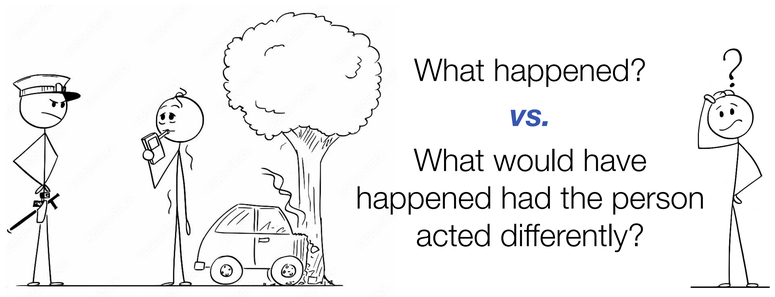
That was close! A counterfactual simulation model of causal judgments about social agents

Sarah A. Wu (sarahawu@stanford.edu)[1], Shruti Sridhar[2], & Tobias Gerstenberg[1]

[1]Department of Psychology, Stanford University [2]Department of Computer Science, Stanford University

Introduction

How do people evaluate each others' actions?



What happened?

vs.

What would have happened had the person acted differently?

- Extending the ***counterfactual simulation model*** of causal judgments from physical [1] to social scenarios
- People can invert generative mental models of others' actions using their intuitive theories of psychology [2-4]
- But people also consider social evaluations [5-6]

Hypotheticals	***vs.***	*Counterfactuals*
What would happen in the future, if they acted differently in the present?		What would have happened in the present, if they had acted differently in the past?

- Are counterfactuals necessary? (or are hypotheticals sufficient?) [7]

Computational Model

Generative model:

- Rational planning (graph search and Q-learning) in gridworlds formalized as Dec-MDPs
- Exp. 1: contrast = a single agent's binary decision
- Exp. 2: contrast = a second agent's (possible helping or hindering) interactions with the first agent

Modeling causal judgments:

- *Hypothetical simulation:* predict outcome under contrast
- *Counterfactual simulation*: predict outcome under contrast, conditioning on observed environment events
- *Heuristic:* linear regression using visual features of scene

Modeling intention inferences:

- Bayesian inference over possible goals = {help, hinder}

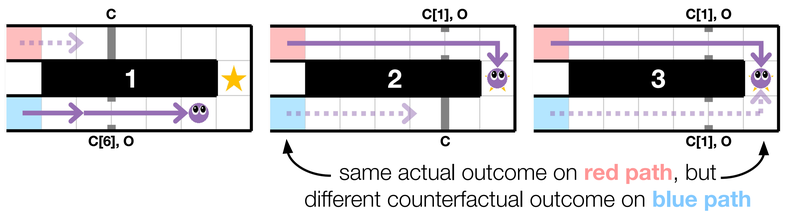
$$p(g_i \mid s, a_i) \propto p(a_i \mid s, g_i) \propto \exp\left(\beta \times \underbrace{Q_i(s, a_i)}\right)$$

agent *i*'s goal · state · agent *i*'s action · expected future reward

Experiment 1

Setup: OSF

- Agent chooses **red** or **blue** path and has 10 timesteps to reach goal ⭐, but can only pass through open doors



Hypothetical: "The agent would win if they took the red path this time." (asked at beginning of trial) (n = 50)

Counterfactual: "The agent would have won if they had taken the red path this time." (n = 50)

don't agree at all — agree very much

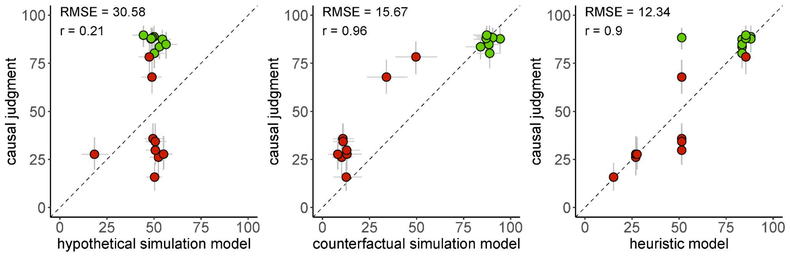
Results:

- Simulation model captures empirical hypothetical (r = 0.83) and counterfactual (r = 0.94) judgments well

Experiment 2

Causal: "The agent lost because they took the blue path this time." (n = 50)

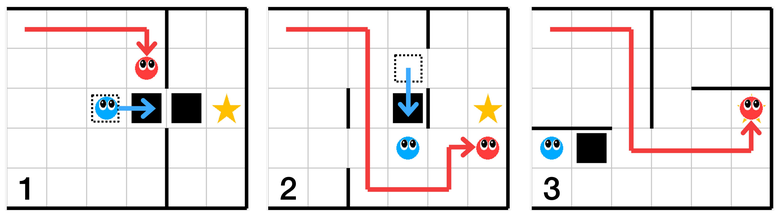
- Causal judgments best explained by counterfactual simulations, not hypothetical simulations or heuristics



Ongoing Experiment

Setup: OSF

- Red agent has 10 timesteps to reach goal ⭐
- Blue agent can push/pull boxes in order to help/hinder



Causal: "The red agent lost because of the blue agent."

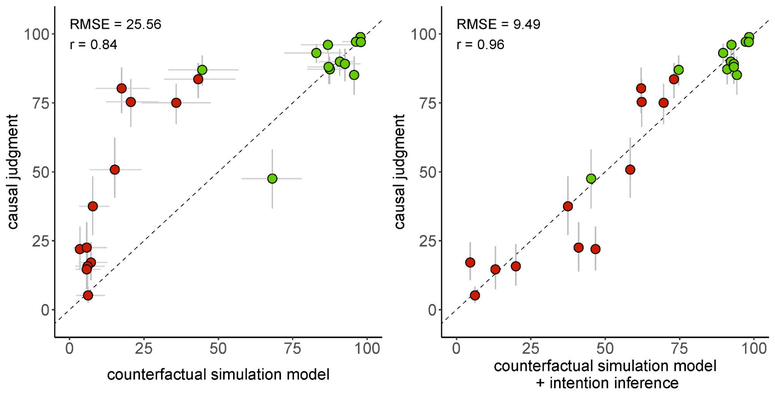
Counterfactual: "The red agent would have won if the blue agent hadn't been there."

don't agree at all — agree very much

Intention: "What was the blue agent intending to do?"

definitely hinder — unsure — definitely help

Results:



- Model also captures counterfactual judgments (r = 0.93) and empirical intention inferences (r = 0.97) well

Discussion

- Causal judgments about outcomes resulting from agents' actions are best explained by considering relevant ***counterfactual simulations*** as well as ***social inferences*** (here, intentions) about those agents
- Future directions: more complex settings, the problem of counterfactual selection, the process of mental simulation

References: 1. Gerstenberg et al. (2021). *Psychol Rev.* 2. Baker et al. (2017). *Nat Hum Behav.* 3. Gerstenberg & Tenenbaum (2017). *Oxf Handbk Caus Reas.* 4. Jara-Ettinger et al. (2016). *Trends Cog Sci.* 5. Langenhoff et al. (2021). *Cog Psychol.* 6. Sosa et al. (2021). *Cognition.* 7. Gerstenberg (2022). *Philos Trans R Soc B: Bio Sci*. Illustrations: Zdenek Sasek.