That was close! A counterfactual simulation model of causal judgments about social agents
Sarah A. Wu (sarahawu@stanford.edu)1, Shruti Sridhar2, & Tobias Gerstenberg1
1Department of Psychology, Stanford University  2Department of Computer Science, Stanford University

Introduction
How do people evaluate each others’ actions?

- Extending the counterfactual simulation model of causal judgments from physical1 to social scenarios
- People can invert generative mental models of others’ actions using their intuitive theories of psychology2-4
- But people also consider social evaluations5-6
- 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}

Experiment 1
Setup:
- 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)

Counterfactual: “The agent would have won if they had taken the red path this time.”

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

Experiment 2
Setup:
- 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.”

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: