The role of counterfactual reasoning in responsibility judgments
Sarah Wu (sarahawu@stanford.edu) & Tobias Gerstenberg
Department of Psychology, Stanford University

**Introduction**

- Defendent vs. Reasonable person
  - “Reasonable person” legal standard
  - “Ghosting” in AI sports analytics

Responsibility ↔ Causality ↔ Counterfactuals

- Extending counterfactual theories of causation
  - Responsibility as difference-making: judging by comparing what happened with what would have happened in relevant counterfactual situations
  - Counterfactual potency = if-likelihood × then-likelihood

**Experiment 1**

**Setup:**
- Three agents contribute equally to a positive outcome
- Agents may be busy, but others can take their place
- Manipulate number of possible replacements

**Results:** Responsibility decreases with the number of replacements, even when the outcome and individual contributions are the same.

What if participants are simply mapping the number of replacements without computing any counterfactuals?

**Hypotheses:**
- **Contribution model:** Responsibility is about individual contribution
  - Prediction: Uniform judgments across all trials
- **Difference-making model:** Responsibility is about counterfactual difference-making
  - Prediction: number of replacements ↑, responsibility ↓

**Discussion**

- In judging responsibility, people consider counterfactual scenarios and assign responsibility to the extent that counterfactual outcomes would have been different
  - Here: “what if the highlighted agent had been busy?”
- Responsibility ratings well predicted by then-likelihood

**Future directions:** exploring more complex and explicitly simulated counterfactuals

**Experiment 2**

**Setup:**
- Manipulate number and “busyness” of replacements

**Results:** Responsibility decreases with the probability of replacement, not simply the number.

**Hypotheses:**
- **Number of replacements model:** Mapping number of replacements without counterfactuals
  - Prediction: number of replacements ↑, responsibility ↓
  - but no difference with varying busyness
- **Difference-making model:** Responsibility is about counterfactual difference-making
  - Prediction: probability of finding available replacement (based on number and busyness) ↑, responsibility ↓

**References**