Explaining Behavior with What-If Thinking: A Replication and Extension

Batuhan Erdogan and Zeren Kösoğlu

Abstract

We replicated the study by Brockbank et al. (2024), which proposed that people explain behavior by mentally simulating alternative scenarios, also known as counterfactuals. Their model predicted human judgments about whether observed actions are due to internal traits or external situations. We re-applied the original model and re-analyzed the data from their experiment, confirming their reported results. The counterfactual simulation model predicted situational attributions strongly and trait-based attributions moderately. We then developed and tested a novel hybrid model that combines counterfactual difference-making with heuristic expectations. This hybrid model performed best in predicting trait judgments and matched the heuristic model in situation judgments. Our findings reinforce the importance of modeling structural alternatives, and show that carefully calibrated hybrids may offer competitive explanatory power.

1 Introduction

We often make sense of others' actions by asking ourselves questions such as 'What if this person had made a different choice?' or 'What if the situation had been different?' These are counterfactual questions imagining alternative versions of events to understand what caused them. Although the dominant position in psychology asserts that people rely mainly on trait judgments rather than counterfactual assessments, recent work in cognitive science suggests that people do use such counterfactual simulations to explain others' behavior.

The computational model developed by Brockbank et al. (2024) puts this idea into practice. Their model simulates possible outcomes by altering either the agent's personality (trait) or the environment (situation) and compares those alternative outcomes to the actual one. This helps predict whether people will attribute behavior more to apparent traits (such as being an optimist) or to the situation at hand (such as starting out with a disadvantaged position).

Our study replicates and extends their work. We examined whether their results could be independently verified and explored whether similar or improved predictions could be obtained using a hybrid strategy that combines simple difference-making with heuristic shortcuts.

2 Summary of the Original Study

Participants: The original study recruited 100 adults through Prolific. Participants were randomly assigned to one of five conditions in a between-subjects design.
Task: Each participant observed an agent in a grid-world game. The agent's task was to collect berries by navigating a 10x10 map. Two factors were manipulated: the agent's *trait* (optimist or pessimist) and the *starting location* (situation). After watching the behavior, participants answered questions about whether the result was due to the trait or the situation.
Model: The counterfactual simulation model used Monte Carlo methods to simulate alternative scenarios—either changing the agent's expectations (trait) or changing their start position (situation). Based on the difference between these simulated outcomes and the real outcome, the model predicted how people would respond.
Findings: The model's predictions matched human judgments well for situational causes (correlation $r = 0.82$), and somewhat less well for trait-based causes ($r = 0.37$). It performed better than simpler baselines.

3 Our Replication Study

Goals

We had two main goals. First, to replicate the findings of Brockbank et al. through an independent execution of their model and analyses. Second, to test whether a new hybrid model, combining coun-

terfactual sensitivity with heuristic evaluations, could match or exceed the predictive power of the original.

Methodology

We analyzed and recompiled the experimental setup with all the relevant models, based on descriptions in the original paper and the repository provided by the researchers. This includes the grid-world simulation, agent types, and the counterfactual reasoning procedures.

To evaluate model performance, we used both trial-level Pearson correlations between model predictions and human judgments, and full Bayesian model comparisons using leave-one-out cross-validation (LOO) via the `brms` package.

Our hybrid model introduces an additional layer to the heuristic used in the original research: it calculates counterfactual difference scores both for trait (C_{trait}) and situation (C_{start}), and then normalizes these into weights w_{trait} and w_{start}. These weights determine how much influence to assign to the heuristic components. In particular:

- C_{trait} is the expected reward difference when swapping the agent's trait.
- C_{start} is the expected reward difference when swapping the start location.
- $w_{trait} = \frac{C_{trait}}{C_{trait}+C_{start}}$, and likewise for w_{start}.
- These weights are used to scale the values from the heuristic model: expected rewards and symbolic rules.

4 Results

Replication of the Original Model

We successfully reproduced the performance reported by Brockbank et al. For situational attributions, our counterfactual simulation model reached $r = 0.82$ with human judgments. For trait attributions, the correlation was $r = 0.37$. Bayesian model comparisons confirmed the superiority of the counterfactual model over baseline models.

Hybrid Model Performance

After iteratively refining the hybrid model, we achieved substantial improvements. It now outperforms all other models in predicting trait judgments and performs comparably to the heuristic model for situation judgments.

Bayesian model comparison confirms this:

Trait judgments (LOO):

- Hybrid model (best): Δelpd = 0.0
- Heuristic model: Δelpd = -35.9 (SE = 8.8)
- Simulation model: Δelpd = -44.6 (SE = 12.4)
- Baseline: Δelpd = -86.7 (SE = 11.9)

Situation judgments (LOO):

- Simulation model (best): Δelpd = 0.0
- Hybrid model: Δelpd = -57.8 (SE = 11.4)
- Heuristic model: Δelpd = -57.6 (SE = 11.8)
- Baseline: Δelpd = -92.3 (SE = 12.8)

Due to computational constraints, we were unable to generate individual trial-level Bayesian fits for the hybrid model as we did for the others. Thus, its evaluation is based on aggregate results.

5 Discussion

Our results confirm and extend the original study's conclusions. First, we successfully replicated the superior performance of the counterfactual model for situational judgments. Second, we improved upon the trait modeling by combining structural counterfactual sensitivity with flexible heuristics. The resulting hybrid model not only captured more variance in trait judgments, but also achieved the best predictive accuracy.

This suggests that human causal reasoning may flexibly integrate both simulation-based and heuristic mechanisms, and that weighting heuristic components by difference-making measures (e.g., w_{trait}, w_{start}) provides a principled and cognitively plausible mechanism.

While the hybrid model did not outperform in situation judgments, it matched the heuristic model and came close to the simulation model. We believe that its efficiency and flexibility make it a promising tool for future cognitive modeling work.

6 Conclusion

Our independent replication confirms the findings of Brockbank et al. (2024), while our extension shows that hybrid reasoning models can be competitive. By

combining causal difference-making with heuristic expectations, our hybrid model captures both structural depth and computational economy. This opens the door for models that approximate human social reasoning without full simulation, but still remain grounded in the causal structure of counterfactual alternatives.

References

Brockbank, E., Panos, A., & Ullman, T. D. (2024). Without his cookies, he's just a monster: A counterfactual simulation model of social explanation. *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*