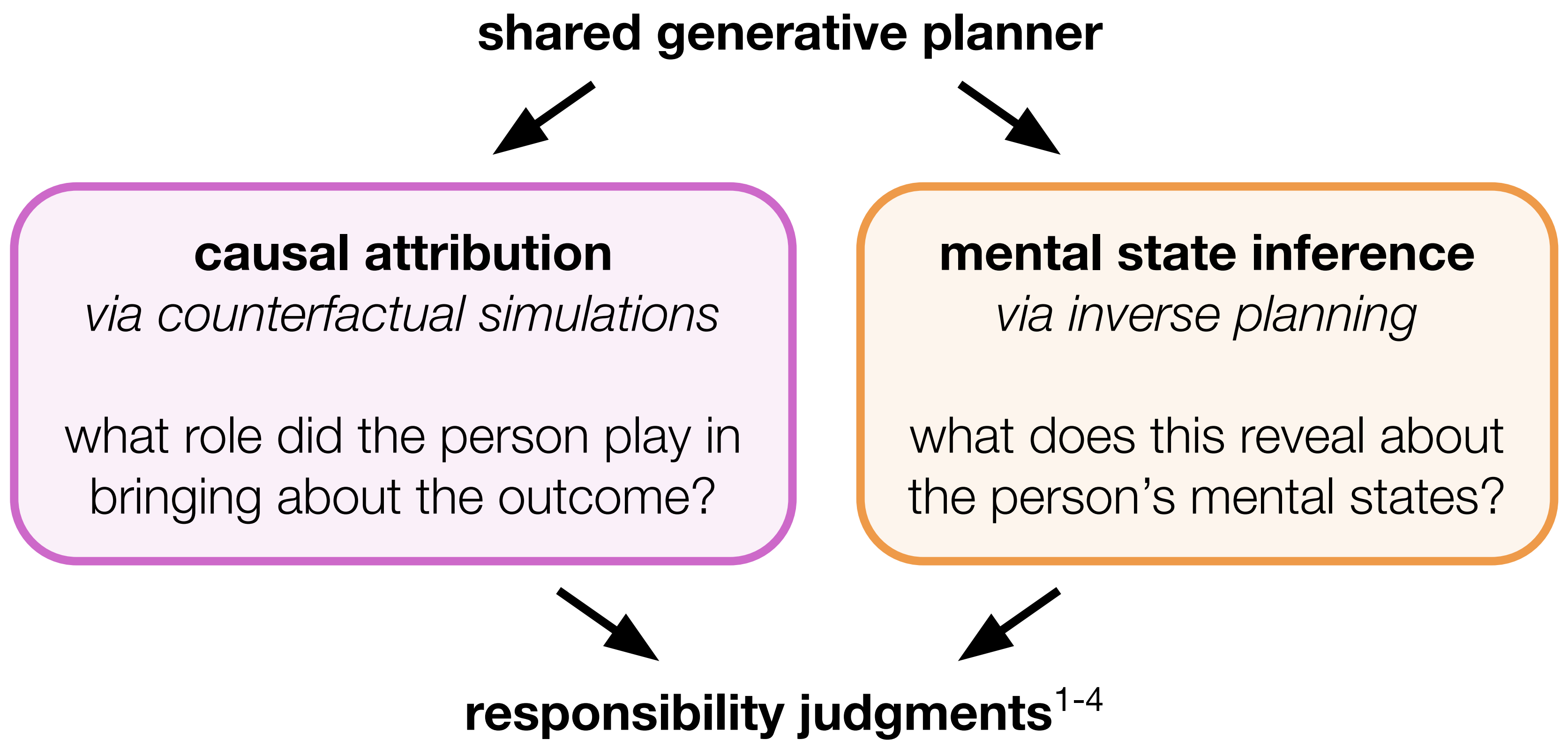
A computational model of responsibility judgments from counterfactual simulations and intention inferences

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 hold others responsible in social interactions?



shared generative planner

causal attribution
via counterfactual simulations

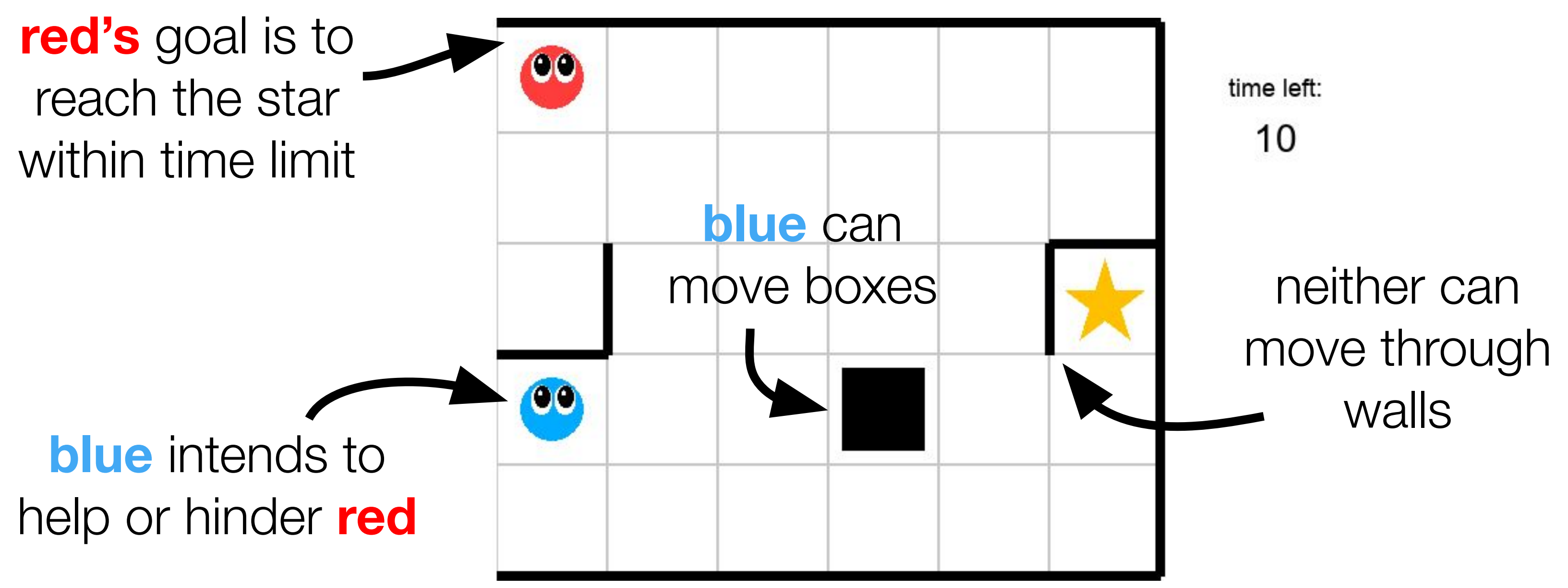
what role did the person play in bringing about the outcome?

mental state inference
via inverse planning

what does this reveal about the person's mental states?

responsibility judgments[1-4]

Model

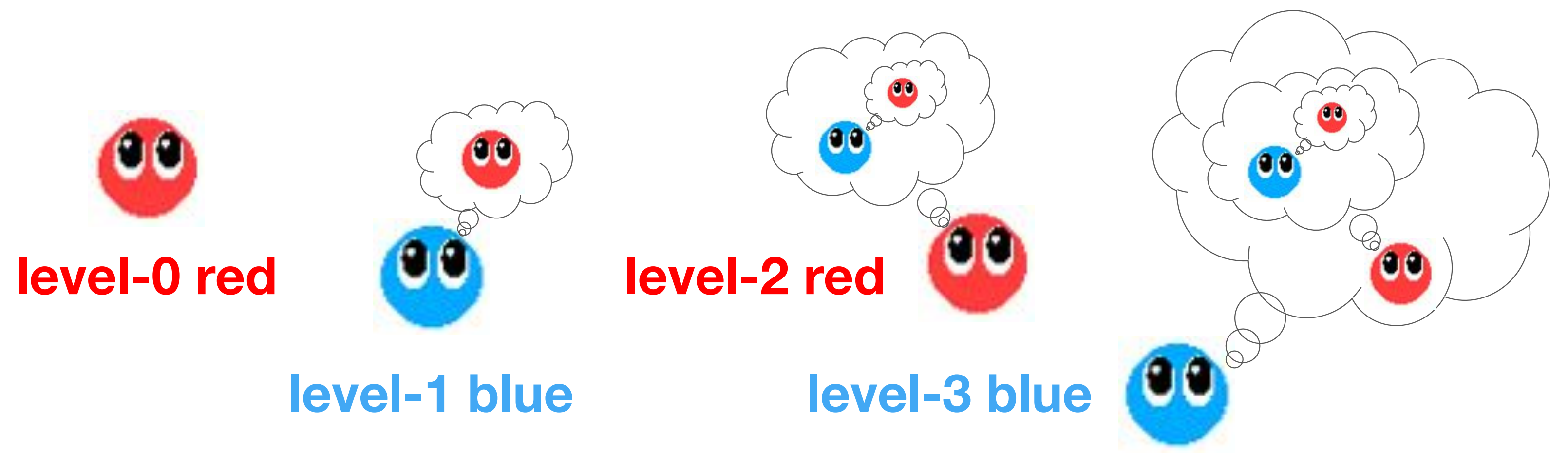


red's goal is to reach the star within time limit

blue can move boxes

neither can move through walls

blue intends to help or hinder **red**



Environments formalized as Social MDPs[5]:

$$M_i^l = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \chi_i, g_i, R_i^l, \gamma \rangle$$

χ_i = agent i's social goal

g_i = agent i's physical goal

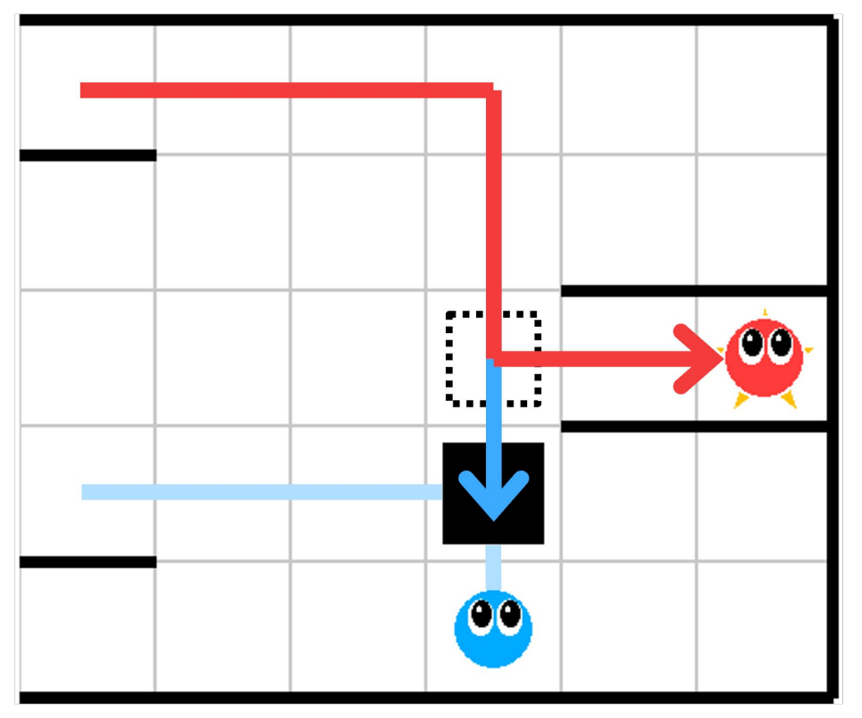
R_i^l = l-th level reward function for agent i

Counterfactual: What would have happened had blue not been there?

Mental state inference: What was blue intending to do?

Experiment 1

level-0 red and **level-1 blue**

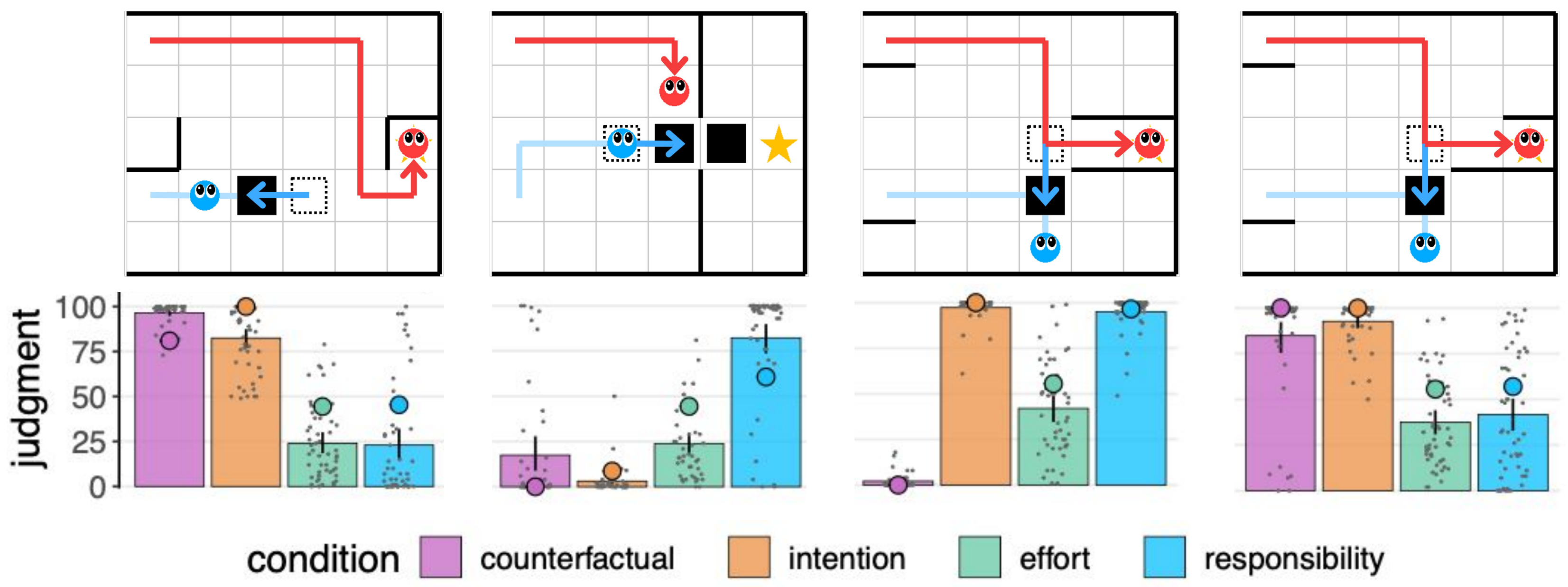


24 trials varying the actual outcome, the counterfactual outcome, and **blue**'s intentions

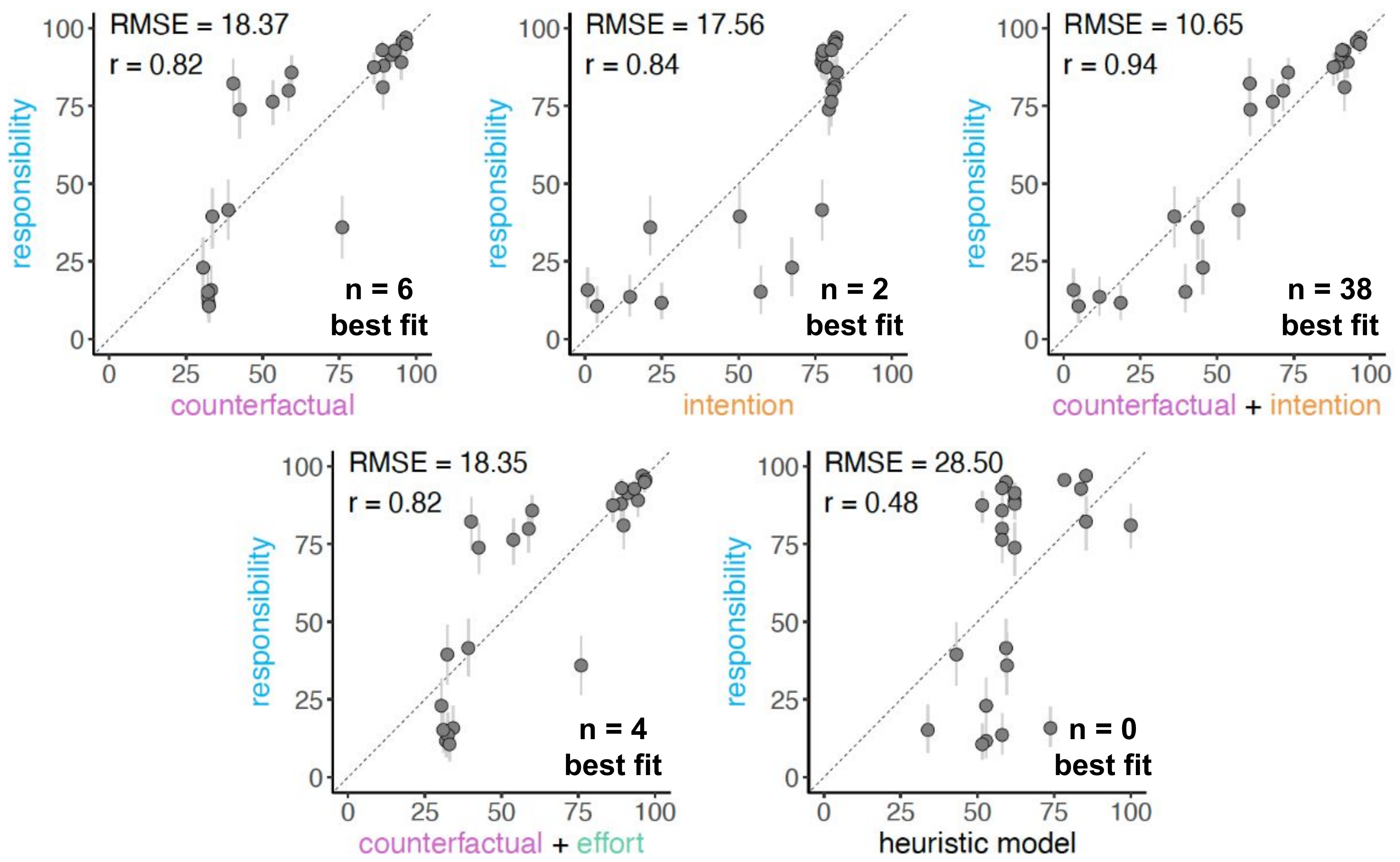
Participants in different conditions (n = 50 each) were asked:

1. **Counterfactual**: How much do you agree that **red** would (still) have succeeded if **blue** hadn't been there?
2. **Intention**: What was **blue** intending to do?)
3. **Effort**: How much effort did **blue** exert?
4. **Responsibility**: How responsible was **blue** for **red**'s success / failure?

Participants' judgments for select trials:

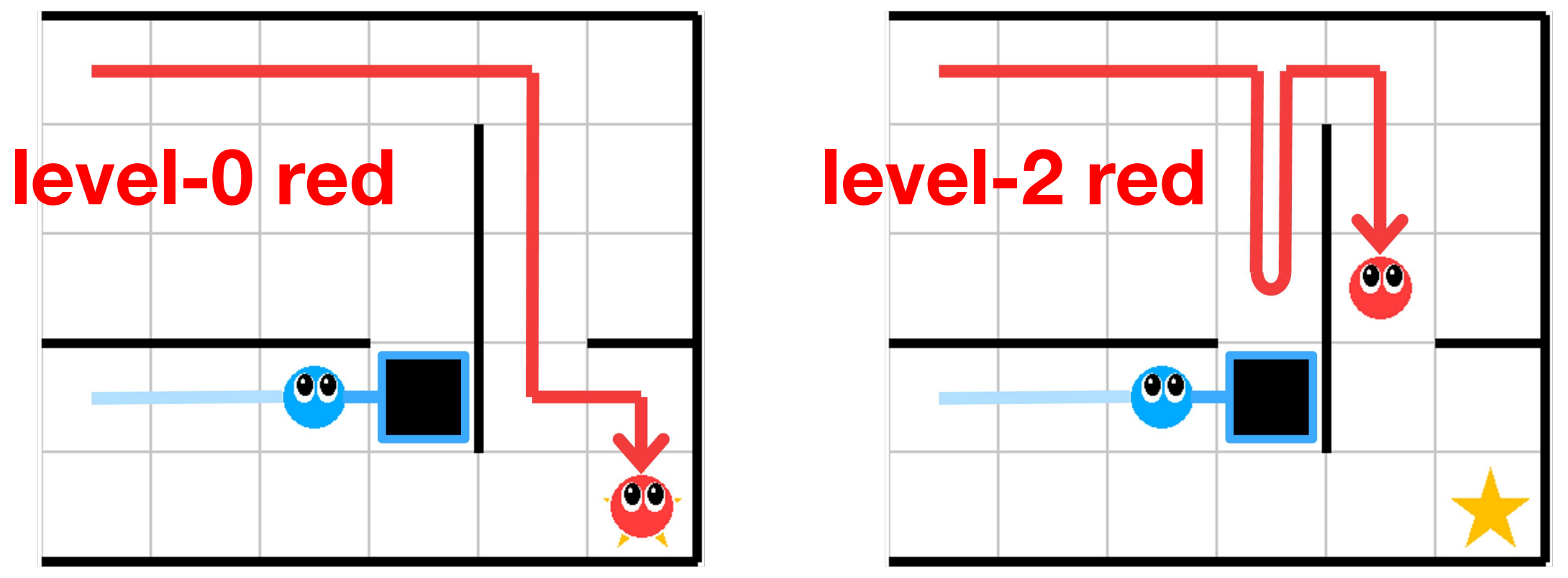


Responsibility model predictions:



Experiment 2

includes **level-2 red** and **level-3 blue**



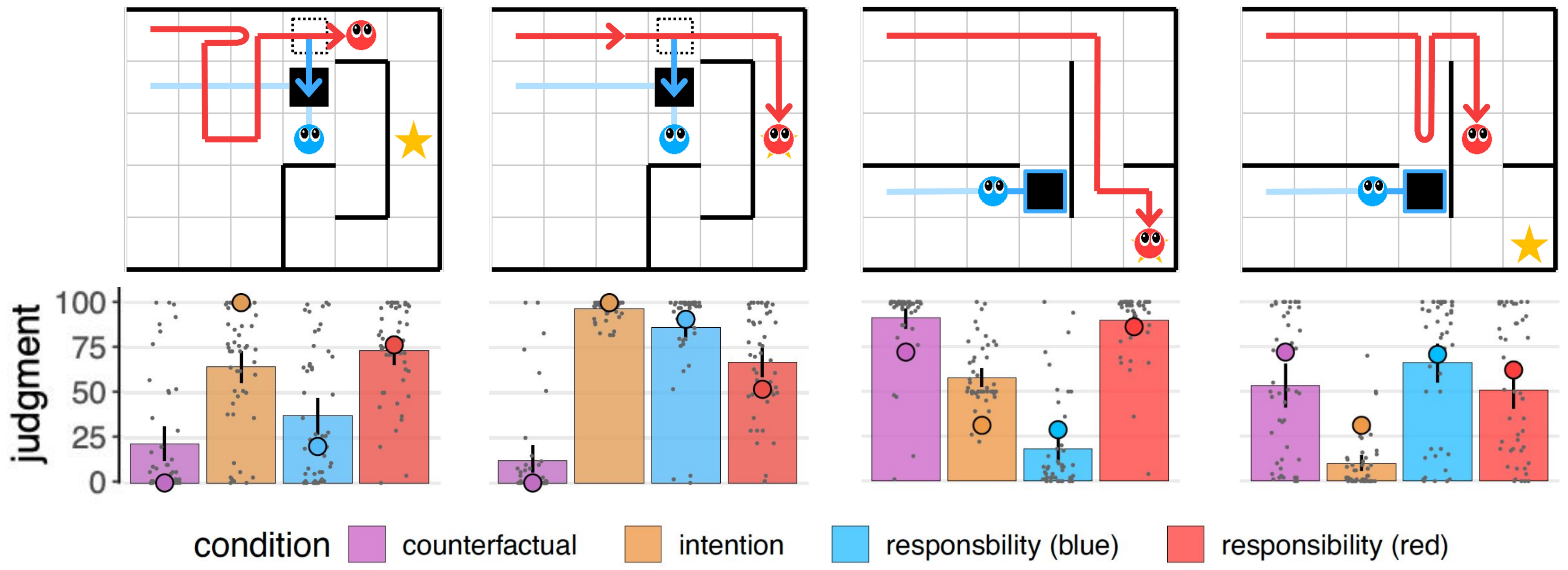
level-3 blue tricks **red** by appearing to be helpful, but not actually helping

12 pairs of trials differing only in whether **red** is level-0 or level-2

Participants in different conditions (n = 50 each) were asked:

1. **Counterfactual**: same as Experiment 1
2. **Intention**: same as Experiment 2
3. **Responsibility**: How responsible was **blue** for **red**'s success / failure? How responsible was **red** for the success / failure?

Participants' judgments for select trials:



Responsibility model predictions:

- **Counterfactuals** + **intentions** model again explained responsibility judgments best (r = 0.94, lowest RMSE, n = 26/50 best fit)
- Responsibility towards **blue** vs. **red** were highly anti-correlated!

Discussion

Responsibility judgments are best explained by a combination of counterfactual simulations ("what would have happened otherwise?") and mental state inferences ("what was the agent intending?").

Future work:

- Further investigating communicative actions (signaling, deception)
- Exploring responsibility throughout repeated interactions ("fool me once, shame on you, fool me twice, shame on me!")

References: 1. Gerstenberg et al. (2018). *Cognition.* 2. Langenhoff et al. (2021). *Cog Psychol.* 3. Sosa et al. (2021). *Cognition.* 4. Carlson et al. (2022). *Nat Rev Psychol.* 5. Tejwani et al. (2021). *CoRL.*