Leave a trace: Recursive reasoning about deceptive behavior

Verona Teo, Sarah A. Wu, Erik Brockbank, Tobias Gerstenberg

Introduction

- When people act deceptively, they must reason about how others will interpret their actions and adjust their behavior accordingly.
- How do people plan deceptive actions, and how do observers make inferences based on the evidence left behind?

Experiment 1: Suspects

Draw a path for the agent to take the snack back to their room.

- Participants acted as suspects planning paths to and from the fridge.
- They were either told to get or steal a snack (without getting caught).
- Suspects' paths left behind physical evidence.

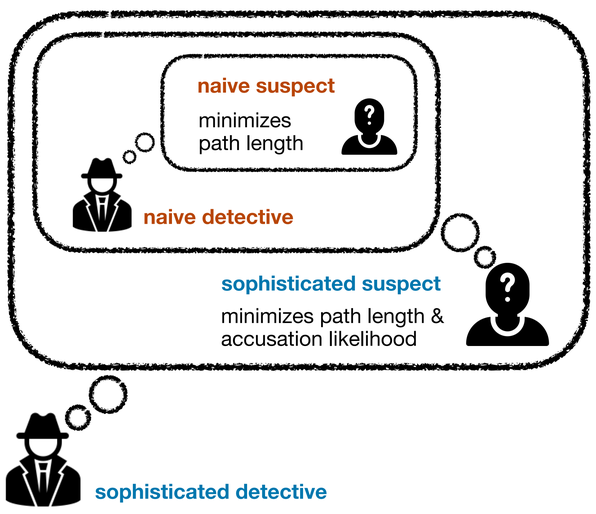
Experiment 2: Detectives

- As detectives, participants were shown the final scene after someone had left evidence behind and were asked to figure out who did it.
- Detectives were either told that someone had taken or stolen a snack.

Models

Recursive Simulation Model (RSM)

- RSM combines inverse planning with recursive theory of mind to select actions and reason over evidence.
- There are three components: path planning, evidence generation, and an inference mechanism.
- It simulates agents as level-k reasoners, either naive $(k = 1)$ or sophisticated $(k = 2)$.

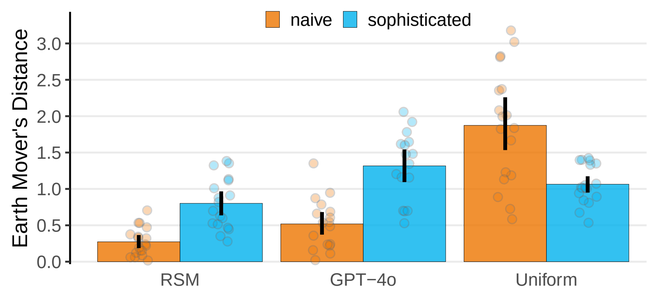


Alternative Models

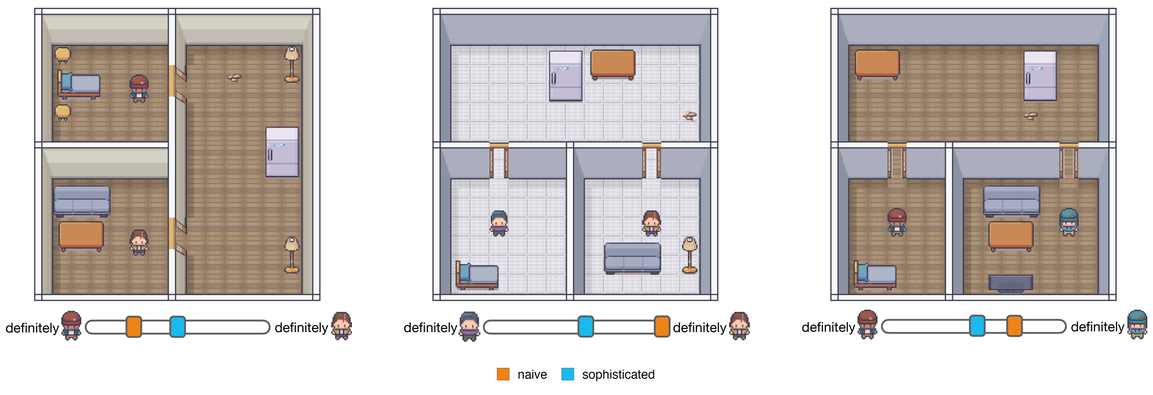
- *Uniform simulation*: Samples paths uniformly, so it does not distinguish between naive and sophisticated agents
- *Empirical simulation*: Uses paths generated by participants in place of simulated paths
- *Heuristic*: Linear model with features based on directly observable features, e.g., distance
- *GPT-4o*: How well does a VLM do?

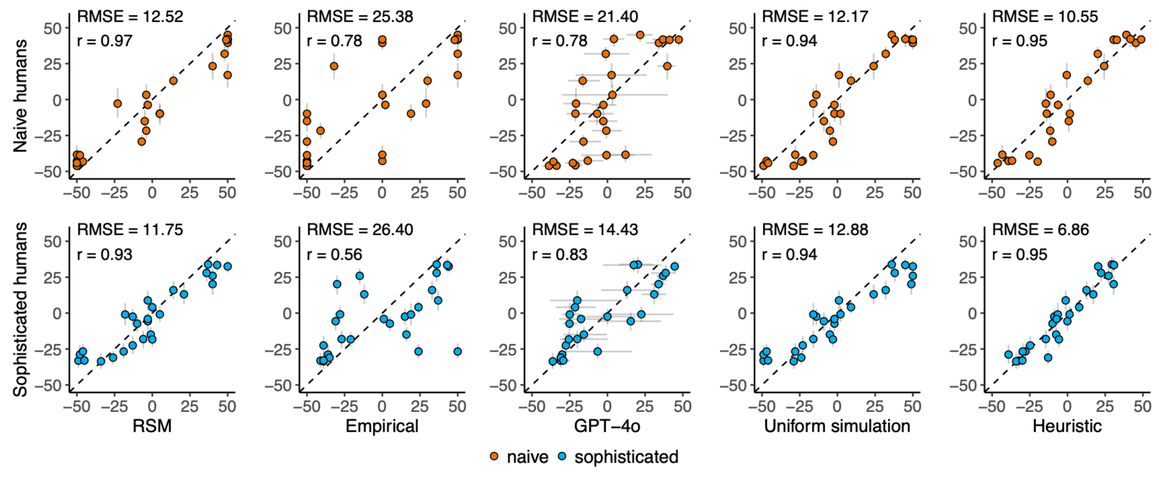
Results

Suspects



Detectives





Conclusion

- People are adept at acting as deceptive suspects, demonstrating complex theory of mind in action planning.
- Detective uncertainty about deceptive agents suggests limits in recursive reasoning when interpreting others' actions.
- Future work includes extending the paradigm to incorporate additional modalities like audio.

Paper Link