Leave a trace: Recursive reasoning about deceptive behavior

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

Department of Psychology, Stanford University, USA

veronateo@stanford.edu

Abstract

How do people reason about others when planning deceptive actions? How do detectives infer what suspects did based on the traces their actions left behind? In this work, we explore deception in a setting where agents steal other's snacks and try to determine the most likely thief. We propose a computational model that combines inverse planning with recursive theory of mind to select misleading actions and reason over evidence arising from such plans. In Experiment 1, we demonstrate that suspects strategically modify their behavior when acting deceptively, aligning with our model's predictions. Experiment 2 reveals that detectives show increased uncertainty when evaluating potentially deceptive suspects—a finding consistent with our model, though alternative explanations exist. Our results suggest that people are adept at deceptive action planning, but struggle to reason about such plans, pointing to possible limits in recursive theory of mind.

Keywords: social cognition; theory of mind; deception; mental simulation; causal inference

In the high-profile criminal trial of *The People vs. O. J. Simpson*, the defense argued that evidence against Simpson—which included the defendant's blood at the scene of the crime—had been planted by police investigating the case. This appeal by the defense reflects the potential for sophisticated reasoning about the causes of others' behavior; when inferring what happened, people can consider the possibility that evidence for a particular action was generated through deceptive motives. How do people plan misleading actions, and how do others incorporate possibly deceptive behavior in their inferences about what happened?

In order to plant evidence or cover one's tracks, one needs to represent what inferences an observer would make based on the evidence they see. Recent work has formalized these inferences in the context of lying and lie detection, arguing that both rely on a *recursive theory of mind (ToM)*—the ability to reason about others' reasoning about one's own reasoning—in which liars model the beliefs of lie detectors under different possible outcomes, and lie detectors model the expected behavior of liars under different possible world states (Alon et al., 2024; Oey & Vul, 2024; Oey et al., 2023; Schulz et al., 2023; Tan et al., 2024). In these settings, people's ability to craft believable lies, and to detect such lies when they are less credible, may arise in part from the relatively constrained space of possible actions and world states. In the real world, carrying out successful deceptive behavior requires planning *how* to do it.

To make sense of others' actions (even non-deceptive ones), *inverse planning* models propose that people reason about others as if they were approximately rational planners (Baker et al., 2009, 2017; Jara-Ettinger et al., 2016, 2020). This allows observers to infer unseen mental states such as goals and beliefs that could have given rise to others' behavior (Ullman et al., 2009; Wu et al., 2023). To further account for inferences about *unobserved* behavior after the fact (e.g., forensics), recent work has extended inverse planning to incorporate causal models of how actions may leave physical traces in the environment (Jacobs et al., 2021; Jara-Ettinger & Schachner, 2024; Jin et al., 2024; Lopez-Brau et al., 2022; Pelz et al., 2020; Wu et al., 2024).

In our framework, we characterize agents who pursue their goals without deception as *naïve suspects*. When a *naïve detective* makes inferences about what happened, they engage in simple inverse planning with naïve suspects in mind. However, a naïve detective can be exploited by a *sophisticated suspect* that acts deceptively. Recent work exploring social behaviors such as storytelling and intervening on others' emotions has cast these behaviors as *inverse inverse planning*, where reasoners optimize over a model of an audience or observer (who is in turn executing inverse planning) to meet certain presentational goals (Chandra et al., 2024; Collins et al., 2024; Yoon et al., 2020).

Inverse inverse planning models provide a possible account of how a sophisticated suspect might plant evidence to make another agent seem guilty to a naïve detective. The daunting task for the *sophisticated detective* is to perform yet another inversion by considering the plausibility of behavior intended to deceive the careless observer. Formulating the problem in this way speaks to the inferential complexity of such reasoning, yet as the case of O. J. Simpson revealed, our intuitive theories of deception allow juries and members of the general public to evaluate this possibility. In this project, we develop and test a recursive simulation model to capture how people engage in deceptive action planning and reason about such behavior.

Experimental Paradigm

Our experimental paradigm tests this theoretical framework by examining how people plan and reason about deceptive actions in a controlled setting. We designed an environment featuring agents in a household setting, where one of two agents

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

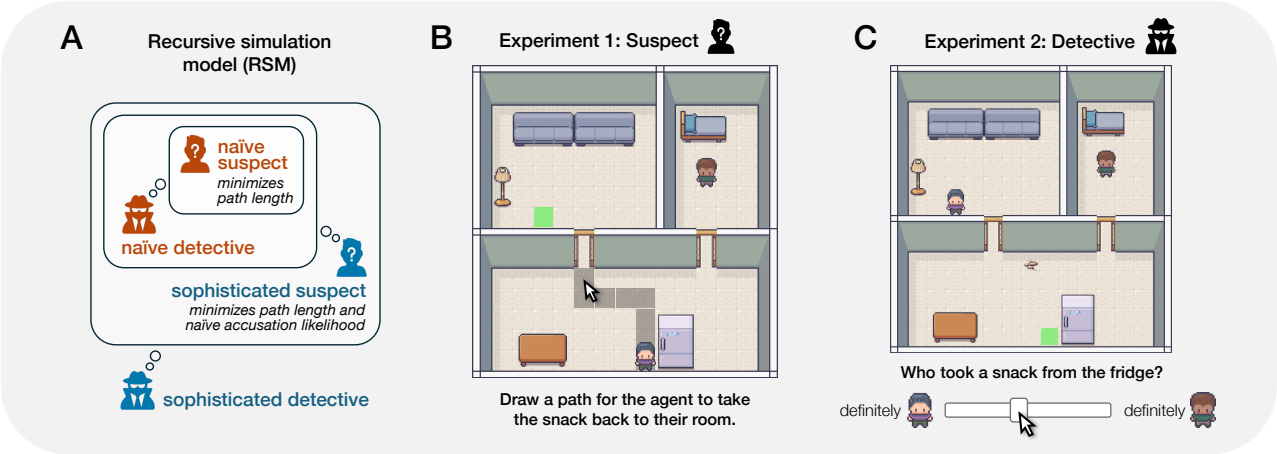

Figure 1: **Model and experiment overview.** **(A)** In the recursive simulation model (RSM), a naïve detective reasons about naïve suspects who only plan efficient paths to and from the fridge. Sophisticated suspects balance both path length and the likelihood of being accused by a naïve detective. Sophisticated detectives reasons about sophisticated suspects. **(B)** In Experiment 1, participants drew paths for the agents and were instructed to either simply get a snack (naïve), or to steal someone else's snack without being caught (sophisticated). **(C)** In Experiment 2, participants were asked to infer which agent took the snack based on the evidence in each scene, given context about whether the agents were naïve or sophisticated.

walks to the fridge in the kitchen to grab or steal a snack (Figure 1). The agents leave crumbs behind on their way back, which provides a clue about their identity (see Lopez-Brau et al., 2022). Our recursive simulation model plans deceptive behavior with a motive to frame the other agent based on visual evidence, and inverts this planning process to draw inferences about which agent took the snack. We test the generative component of our model in Experiment 1, where participants act as suspects, and the inference component of our model in Experiment 2, where participants act as detectives.

Recursive Simulation Model

Our model consists of three key components: (1) a path planning mechanism for suspects, (2) an evidence generation process, and (3) an inference mechanism for detectives. We model our environment as a grid world where agents may step in any of the four cardinal directions, but cannot move through walls or furniture. In each trial scenario, two suspects, A and B, are initially located in different rooms. Suspects plan paths to retrieve a snack from the kitchen, and detectives draw inferences about which suspect most likely did so after the fact. We use level-k reasoning (Camerer et al., 2004; Wright, 2010) to capture the recursive relationship between suspects and detectives (Figure 1A): naïve suspects are only concerned with path efficiency, and naïve detectives assume as much. Sophisticated suspects choose paths to deceive a naïve detective, and sophisticated detectives draw inferences about sophisticated suspects.

The recursive simulation model (RSM) simulates a suspect $s \in \{A, B\}$ at reasoning level k as a rational planner who samples paths p between their initial and target locations according to the cost function:

$$C(s^k, p) = w \cdot \text{len}(p) + (1 - w) \cdot \text{accusation}(s^{k-1}) \quad (1)$$

where $w \in [0,1]$ is a weight parameter that balances path efficiency against the probability of being accused, and $\text{len}(p)$ represents the normalized path length. The accusation term $\text{accusation}(s^{k-1})$ represents the probability of being accused by a detective at level $k-1$ if that path were taken.[1] The parameter w captures the trade-off between path efficiency and accusation likelihood, with lower values of w leading to potentially more deceptive behavior at the cost of taking longer paths. The costs are passed through a softmax function controlled by a temperature parameter τ. Each suspect's path model induces a probability distribution over evidence E—in this case, possible locations of a dropped crumb:

$$\Pr(E = e \mid s^k) = \sum_p \frac{\mathbb{1}_{e \in P}}{\text{len}(p)} \cdot \exp\left(- \frac{C(s^{k-1}, p)}{\tau} \right) \quad (2)$$

This models a uniform probability of leaving evidence at any point along the path from the fridge back to the suspect's room. We approximated this distribution by sampling 1000 paths for each suspect and only considering all simple paths (paths that do not repeat tiles). To make inferences over possible suspects given evidence E, the model applies Bayes' rule to compute the accusation likelihood:

$$\text{accusation}(s^k) = \Pr(s^k \mid E = e) \propto \Pr(E = e \mid s^k) \cdot \Pr(s^k) \quad (3)$$

In this way, generative plans and accusation inferences are computed recursively through the levels. We start with a detective at level $k = 0$ who accuses each suspect equally of having taken the snack, regardless of the evidence.

[1] Path lengths were rescaled to $[0,1]$ and accusation likelihoods were rescaled to $[-0.5, 0.5]$.

Naïve suspect A suspect at level $k = 1$ can simplify their costs to $C(s^1, p) = w \cdot \text{len}(p)$, and becomes equivalent to a simple planner that takes shortest paths.

Naïve detective The detective at level $k = 1$ is an inverse planner (Lopez-Brau et al., 2022) who assumes the agents are naïve suspects, and computes accusation likelihoods by simulating how they would act by sampling with $C(s^1, p)$.

Sophisticated suspect The suspect at level $k = 2$ is an inverse inverse planner (Chandra et al., 2023) who reasons about how their paths might be interpreted by a naïve detective. Their costs take into account the naïve detective's accusation likelihoods computed for s^1. This leads to more strategic planning, such as taking a longer path back from the fridge that may leave evidence closer to the other suspect's location, which would lead the naïve detective to accuse the wrong suspect.

Sophisticated detective Finally, the detective at level $k = 2$ evaluates evidence by simulating each agent as a sophisticated suspect. They recognize that suspects may want to minimize their odds of being accused by taking paths that create evidence which might incriminate the other suspect.

Together, the RSM reasons about potentially deceptive behavior by using recursive theory of mind to simulate other agents' actions and make judgments based on these simulations. We test the RSM against alternative accounts that don't differentiate between naïve and sophisticated suspects, and ones that use heuristics instead of simulating behavior, allowing us to quantify the impact of these processes for explaining human judgments. We also compare participants' responses to those of a vision-language model, GPT-4o (OpenAI, 2024), given similar prompting as human participants. Recent work exploring the capabilities of multimodal models suggests that they exhibit fairly robust visual scene understanding (e.g., object identification, scene descriptions), but may fail in tasks that require more sophisticated physical or causal reasoning about the elements of the scene (Buschoff et al., 2024; Chen et al., 2024; Li et al., 2025). In this way, GPT-4o performance can offer insights into the complexity of planning and social inference over a visual scene required in the current task, while also contributing to our understanding of how well such models can mimic human behavior.

Experiment 1: Suspects

All experiments reported here were pre-registered on OSF.[2] In Experiment 1, participants acted as naïve or sophisticated suspects choosing a path to the fridge and back (Figure 1B).

Methods

Participants 100 participants (*age*: median = 37, SD = 12; *gender*: 39 female, 56 male, 5 non-binary) were recruited on Prolific and compensated $12/hour. All participants were native English speakers residing in the US. Participants were randomly assigned to either the *naïve* or *sophisticated* condition, with $N = 50$ in each.

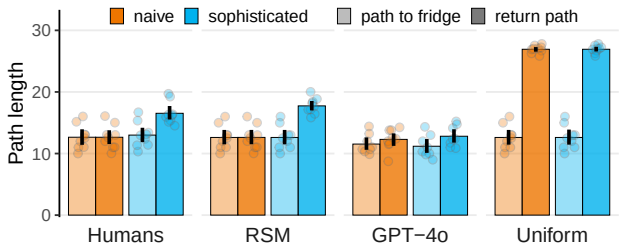

Figure 2: **Experiment 1 path lengths**. Length of paths produced by participants and predicted by models in Experiment 1. Bars show means for each condition and path type, error bars are bootstrapped 95% confidence intervals, and individual points show trial means.

Procedure Participants were first introduced to the apartment setting and the task. In the naïve suspect condition, they were asked to plan a path for one agent to retrieve a snack from the kitchen and return to their room. In the sophisticated suspect condition, participants were instead asked to help one of the agents steal a third roommate's snack. They were told that the agent was bound to drop crumbs on their way back with the food, but would not know exactly when or where, and that they wanted to avoid leaving obvious evidence of having taken the snack. Participants completed a practice trial and a required comprehension check before proceeding to the main experiment. In each trial, participants were shown an image and asked to draw a path from an agent's starting location to the fridge, and then from the fridge back to the agent's starting location, by clicking on adjacent floor tiles in the image. They were not allowed to repeat tiles or move through furniture or walls. The experiment took an average of 13.2 minutes (SD = 6.7) to complete.

Design We designed nine scenes of apartments with varied layouts and locations of rooms, furniture, and agents. Participants were asked to draw paths to and from the fridge for both agents in all nine scenes, yielding a total of 36 trials.

Models To compare model predictions to participants' responses, we sampled 50 paths from the RSM for each trial. The weight w and softmax temperature τ in the RSM were fit by minimizing the Earth Mover's Distance (EMD; Rubner et al., 2000) between the distributions of path locations produced by the model and participants. EMD accounts for both the spatial distance between locations and the probability mass of different paths. We also evaluated a uniform simulation model that samples all paths with equal probability, irrespective of length or accusation likelihood. This model does not differentiate between naïve and sophisticated levels of reasoning. Finally, we compared participants' responses to GPT-4o, which was provided with the same images for each trial along with a text prompt that mirrored the instructions given to participants in each condition. The

[2]Data, materials, and links to pre-registrations can be found here: https://github.com/cicl-stanford/recursive_deception

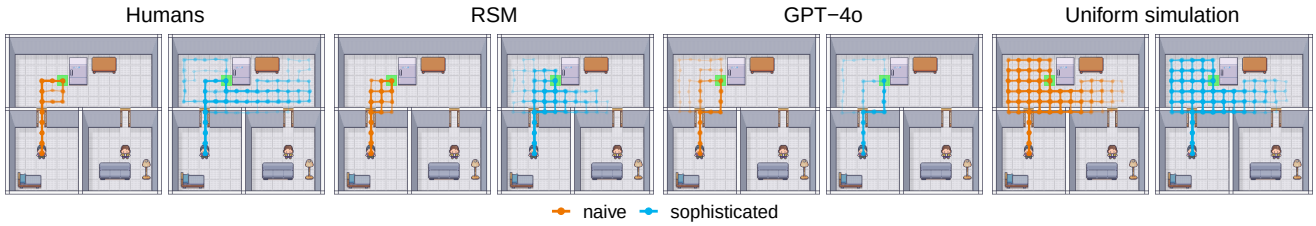

Figure 3: **Experiment 1 results for an example trial**. Paths returning from the fridge (initial tile highlighted in green) back to the agent's room generated by participants, the RSM, GPT-4o, and the uniform model for a select trial. Paths consisted of successive adjacent tiles and were not allowed to repeat tiles or pass through furniture or walls.

prompt also included a JSON scene graph that listed the positions and dimensions of the agents, furniture, and walls in that trial. We used a temperature of 0.7 and zero-shot chain-of-thought prompting (Kojima et al., 2023; Yang et al., 2024). We queried GPT-4o 50 times per trial and condition.

Results

We compared the length and distribution of paths produced by participants and models in the naïve and sophisticated suspect conditions. Of GPT-4o's responses, 7% were non-paths (e.g., refusals to answer) and 65% were invalid paths (e.g., moved through walls). We only included the remaining 28% of responses that contained valid paths in the analysis. The RSM used path cost weights $w = 0.7$ for both conditions and softmax temperatures $\tau_{\text{naïve}} = 0.01$ and $\tau_{\text{sophisticated}} = 0.05$.

Path length The RSM predicts that sophisticated suspects will take more roundabout paths when returning from the fridge compared to naïve suspects, since crumbs dropped along direct paths make suspects more likely to be accused by a naïve detective. Therefore, average return path lengths should differ between naïve and sophisticated suspects. For both participants and the RSM, return paths were longer in the sophisticated condition than in the naïve condition, while paths to the fridge were similar in length (Figure 2). Neither GPT-4o nor the uniform model showed this pattern. As a quantitative test, we fit a linear mixed effects model to participants' path lengths using condition (dummy coded as *naïve* $= 0$, *sophisticated* $= 1$) as a fixed effect, with random intercepts to account for variation across participants and trials. We found that condition was a credible positive predictor of path length for participants (posterior mean: 2.15; 95% highest density interval: [1.36, 2.95]) and the RSM (4.50 [4.18, 4.81]), but not for GPT-4o (-0.16 [-0.48, 0.18]).

Path distributions In addition to path length, we examined how closely the distribution of participants' paths aligned with those produced by different models. In the naïve condition, participants often took efficient paths, while in the sophisticated condition, they tended to wander around the kitchen in hopes of avoiding dropping crumbs in an incriminating spot. The RSM qualitatively captured these different patterns (Figure 3). In contrast, GPT-4o tended to predict shortest paths in both conditions. The uniform model, meanwhile, often generated paths that did not match either naïve or sophisticated agents.

To measure the similarity between participant and model path distributions, we computed the EMD between the probability distribution of locations visited (Figure 4). We fit a linear mixed effects model to predict average EMD values using condition (naïve or sophisticated) and model type as fixed effects. The RSM was found to have a credibly smaller EMD compared to GPT-4o (difference: -0.37; 95% HDI: [-0.65, -0.10]) and the uniform model (-0.93 [-0.65, 1.19]).

Experiment 2: Detectives

In Experiment 1, we found that participants chose different paths depending on whether they had deceptive intentions. This pattern was uniquely captured by the RSM. In Experiment 2, we investigate how people make inferences about suspects from physical evidence (Figure 1C). We compare the inferences of naïve and sophisticated detectives given different information about the suspects' level of reasoning.

Methods

Participants 100 participants (*age*: median = 37, SD = 12; *gender*: 67 female, 32 male, 1 non-binary) were recruited and compensated $12/hour. All participants were native English speakers in the US. They were randomly assigned to either the *naïve* or *sophisticated* condition, with $N = 50$ in each.

Procedure The procedure was initially very similar to Experiment 1. After identical instructions and comprehension check questions, participants completed six trials selected from Experiment 1 as a suspect in the same condition. These trials were meant to familiarize them with the generative process of suspect planning before making detective inferences. Participants then completed a second comprehension check.

In each of the main trials, participants saw an image showing the two agents at their starting locations and a pile of crumbs left somewhere in the kitchen. In the naïve detective condition, they were simply asked to determine which agent had most likely retrieved a snack. In the sophisticated detective condition, they were told that one of the agents had stolen a snack. The agents knew they would leave evidence behind, although they would not know exactly where, and wanted to avoid leaving evidence that incriminated them-

selves. Participants were asked to judge which agent they thought took the snack from the fridge, responding on a continuous slider ranging from "definitely agent *A*" to "definitely agent *B*", with the midpoint labeled "uncertain". The experiment took an average of 11 minutes (SD = 4.9) to complete.

Design The same nine layouts from Experiment 1 were used to generate three trials each for a total of 27 unique trials. Crumb locations for each trial were selectively sampled to produce a wide range of predicted judgments across conditions. In particular, model predictions for some of the trials exhibited high certainty about one of the suspects, while others were more uncertain.

Models As in Experiment 1, we compared participants' responses to those of the RSM, GPT-4o, and the uniform simulation model. The RSM parameters w and τ were estimated by minimizing squared error between model predictions and mean participant judgments. GPT-4o was given similar prompts as in Experiment 1, but was instead asked to produce a judgment about which agent took the snack on the same numerical scale as the slider shown to participants.

In addition, we compared these models to several alternatives that replace or modify key components of the RSM. The first is a variant of the RSM based on empirical data from Experiment 1. This model computes the likelihood of each suspect leaving evidence at a particular location (Equation 2) using the distribution of paths produced by participants in place of simulated paths, allowing us to assess how detective predictions differ when using simulated compared to actual participant-generated paths. Meanwhile, to test whether simulating suspects at the correct level of recursion is in fact a key factor of detectives' predictions, we also considered a "mismatched" empirical model, in which naïve detective predictions are computed based on sophisticated human suspect paths and sophisticated detectives use naïve human suspect paths. Finally, to understand whether participants need to simulate path-planning behavior in this task, we evaluated a heuristic model that relies on directly observable features

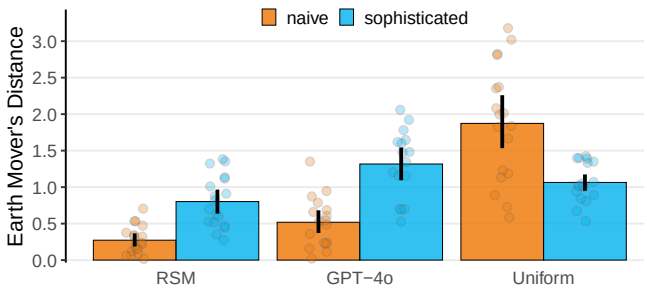

Figure 4: **Experiment 1 model comparison**. The Earth Mover's Distance (EMD) between suspect paths generated by participants and various models. Lower numbers indicate more similar distributions to humans. Bars show means aggregated over all trials, error bars are bootstrapped 95% confidence intervals, and small points show individual trials.

to infer the accusation likelihood of each agent rather than simulations of suspect behavior. The heuristic model fits a linear regression with four features—the Euclidean distance between each agent and the fridge, and between each agent and the evidence—to participants' judgments on each trial.

Results

GPT-4o gave valid numerical responses in 47% of all queries; only those responses were included in our analyses. The RSM used path cost weights $w_{\text{naïve}} = 0.8$, $w_{\text{sophisticated}} = 0.5$, and softmax temperatures $\tau_{\text{naïve}} = 0.05$, $\tau_{\text{sophisticated}} = 0.2$.

Inference (un)certainty Participants in the detective conditions were asked to infer which agent took the snack based on the location of the crumbs. We hypothesized that participants would exhibit more uncertainty in their inferences in the sophisticated condition compared to the naïve condition because sophisticated suspects are expected to produce less diagnostic paths (see Figure 3). We fit a linear mixed effects model to predict the magnitude of participants' slider responses, where larger magnitudes indicate more certainty about a particular agent. The fixed effect of condition was dummy coded as *naïve* = 0, *sophisticated* = 1, while accounting for random variation in responses across trials and participants. The model showed a credible negative effect of condition on slider response (posterior mean: -10.28; 95% HDI: $[-13.57, -7.13]$), indicating that sophisticated participants' judgments were closer to zero, the "uncertain" midpoint of the response slider.

Model comparison We compared participants' responses in each trial and condition to model predictions (Figure 5). Table 1 shows Akaike Information Criterion (AIC) scores for each model fit to mean responses across trials in each condition. Across both conditions, the RSM captured participants' judgments well in terms of correlation, RMSE, and AIC. The uniform simulation and heuristic models also performed comparably. The empirical model based on paths drawn in Experiment 1 did not perform well, likely due to idiosyncrasies in the paths that suspect participants drew (they tended to avoid zig-zagging, for example, which resulted in sparser path distributions). The mismatched empirical model

Table 1: **Experiment 2 model comparison**. AIC scores for all detective models, computed using $n = 2$ parameters for the RSM, $n = 5$ for the heuristic model, and $n = 0$ for all others. Lower scores indicate better model fits (**best**, second best).

Model	Naïve	Sophisticated
RSM	206.22	203.02
Empirical	239.43	241.10
Mismatched empirical	247.85	245.34
GPT-4o	232.45	212.51
Uniform simulation	**200.85**	203.24
Heuristic	207.16	**184.87**

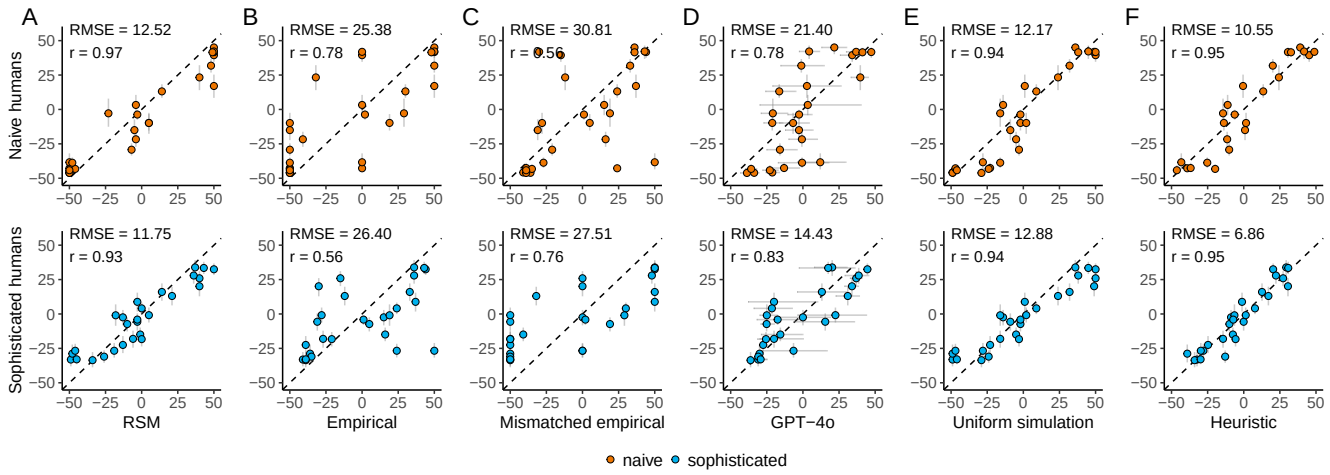

Figure 5: **Experiment 2 results**. Participants' mean detective inferences in each condition (top row = naïve, bottom row = sophisticated) compared to predictions from **(A)** the RSM, **(B)** a model using empirical path data from Experiment 1, **(C)** same model as (B) but with switched data from the two conditions, **(D)** GPT-4o, **(E)** the uniform simulation model, and **(F)** the heuristic model. Error bars are 95% bootstrapped CIs, RMSE = root mean squared error, and r = Pearson correlation coefficient.

generally performed worse; though naïve suspects were comparable to sophisticated suspects for predicting sophisticated human detectives, sophisticated suspects were worse at predicting naïve human detectives. A possible explanation for this is that participants in the sophisticated detective condition were generally more uncertain, so the exact location of the evidence mattered less for their judgments. GPT-4o also did not perform as well, suggesting that inference about deceptive visual evidence may still be a challenging task for vision-language models.

General Discussion

In this paper, we investigated how people plan actions to mislead others, and how they make inferences about the behavior of misleading actors. We developed a computational model that recursively simulates suspects choosing actions that leave behind evidence for detectives, who in turn invert the suspects' planning model to infer the source of the evidence. We tested our model in a variety of scenarios asking participants to act as suspects planning actions with or without deceptive intent (Experiment 1), or detectives inferring the most likely actor given evidence of a naïve or intentionally misleading suspect's behavior (Experiment 2).

Our results suggest that people are adept at acting as deceptive suspects, but may struggle to reason about them. As suspects, participants deviated from the shortest path and planned routes that were consistent with framing the other apartment resident when the goal was not to be accused of stealing food. However, as detectives, participants exhibited uncertainty about which of agent was the likely thief when taking into account the potential for misleading actions by each agent. One possible explanation is that simulating potentially disingenuous agents in the current task may be cognitively demanding; the RSM requires several layers of recursion to reason as a sophisticated detective. Participants' simulations may have been noisy at this depth, or they may have preferred simpler reasoning strategies that did not involve simulation at all. Consistent with this, alternative models that relied on uniform simulations and simple visual heuristics performed similarly to the RSM as an account of participants' sophisticated detective inferences. These findings align with previous research on deception, which suggest that lying can be cognitively demanding (Vrij et al., 2006; Zuckerman, 1981), and that people perform at chance when detecting lies in many situations (Bond & DePaulo, 2006; Levine, 2010) or exhibit more accurate detection when using simple cues and heuristics (Verschuere et al., 2023).

Several models were highly correlated with participants' inferences in Experiment 2. Future work should aim to tease these approaches apart. For example, by manipulating room layouts and furniture positions, superficial cues like the closeness of the evidence to a suspect may not be as diagnostic anymore. In addition, by increasing the costs of misidentification or the rewards of successful forensics, people might rely on more sophisticated simulation-based reasoning. Prior work has found that manipulating the cost of accusation impacts the behavior of both liars and lie detectors (Oey & Vul, 2024; Oey et al., 2023), suggesting that deceptive behavior and inferences about deception rely on general purpose utility calculations (Jara-Ettinger et al., 2016, 2020).

Finally, our paradigm allows us to incorporate additional types of evidence, such as the sounds produced by an agent's actions. We would expect suspects to act differently based on what sources of evidence they believe a detective has access to. The current work contributes towards a comprehensive account of the ways people plan and reason about deceptive behavior, offering insights into the broader sophistication and flexibility of human social reasoning.

Acknowledgments

E.B. was supported by NSF SBE Postdoctoral Research Fellowship #2404706. T.G. was supported by grants from Stanford's Human-Centered Artificial Intelligence Institute (HAI) and Cooperative AI.

References

Alon, N., Schulz, L., Barnby, J. M., Rosenschein, J. S., & Dayan, P. (2024). Detecting and deterring manipulation in a cognitive hierarchy. *arXiv preprint arXiv:2405.01870*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Buschoff, L. M. S., Akata, E., Bethge, M., & Schulz, E. (2024). Visual cognition in multimodal large language models.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games*. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Chandra, K., Li, T.-M., Tenenbaum, J., & Ragan-Kelley, J. (2023). Acting as Inverse Inverse Planning.

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2024). Storytelling as inverse inverse planning. *Topics in Cognitive Science*, *16*(1), 54–70.

Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., & Liu, X. (2024). Egoplan-bench: Benchmarking multimodal large language models for human-level planning.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

Jacobs, C., Lopez-Brau, M., & Jara-Ettinger, J. (2021). What happened here? Children integrate physical reasoning to infer actions from indirect evidence. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., & Schachner, A. (2024). Traces of our past: The social representation of the physical world.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jin, E., Huang, Z., Fränken, J.-P., Liu, W., Cha, H., Brockbank, E., Wu, S., Zhang, R., Wu, J., & Gerstenberg, T. (2024). MARPLE: A Benchmark for Long-Horizon Inference [http://arxiv.org/abs/2410.01926]. *Advances in Neural Information Processing Systems*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023, January). Large Language Models are Zero-Shot Reasoners.

Levine, T. R. (2010). A Few Transparent Liars Explaining 54% Accuracy in Deception Detection Experiments. *Communication Yearbook*, *34*(1), 41–61.

Li, Z., Wang, H., Liu, D., Zhang, C., Ma, A., Long, J., & Cai, W. (2025). Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities.

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*, *151*(9).

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346.

Oey, L. A., & Vul, E. (2024). Accurate approximations about the truth from literally false messages. *Computational Brain & Behavior*, *7*(1), 23–36.

OpenAI. (2024). *OpenAI o1 System Card* (tech. rep.).

Pelz, M., Schulz, L., & Jara-Ettinger, J. (2020). The signature of all things: Children infer knowledge states from static images.

Rubner, Y., Tomasi, C., Guibas, L. J., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 99–121.

Schulz, L., Alon, N., Rosenschein, J., & Dayan, P. (2023). Emergent deception and skepticism via theory of mind. *First Workshop on Theory of Mind in Communicating Agents*.

Tan, Z. Y., Jara-Ettinger, J., & Berke, M. (2024). Reasoning about knowledge in lie production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, *22*.

Verschuere, B., Lin, C.-C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E., Goor, T., Loewy, L., Appiah, O., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, *7*.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*, 141–142.

Wright, J. (2010). Beyond equilibrium: Predicting human behaviour in normal form games. *Proceedings of the Behavioral and Quantitative Game Theory on Conference on Future Directions*.

Wu, S. A., Brockbank, E., Cha, H., Fränken, J.-P., Jin, E., Huang, Z., Liu, W., Zhang, R., Wu, J., & Gerstenberg, T. (2024). Whodunnit? Inferring what happened from multimodal evidence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0).

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 3375–3382).

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2024, April). Large Language Models as Optimizers.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.

Zuckerman, M. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, *14*, 1–59.