
Lies on a Leash: LLMs Detect and Deflect Deception via Harnesses

Nitay Alon

NITALON@MIT.EDU

Joshua B. Tenenbaum

JBT@MIT.EDU

Computational Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA 02139
USA

Reuth Mirsky

REUTH.MIRSKY@TUFTS.EDU

Department of Computer Science and Department of Mechanical Engineering, Tufts University,
177 College Avenue, Medford, MA 02155

Stefan Sarkadi

SSARKADI@LINCOLN.AC.UK

APEx Defence and Security Institute, University of Lincoln, Lincoln, United Kingdom

Abstract

Humans lie and detect lies. Deceivers deceive successfully by inferring whether a victim will detect them, and victims detect deception by inferring how a sender might lie. Deception and counter-deception require a set of cognitive traits: the ability to infer what others are thinking, the capacity to reason how likely others will accept an unfair outcome, and other qualities. These behaviors have been formalized as recursive social reasoning. While deceptive dynamics have been tested and recorded in humans, should we expect modern AI agents to follow the same rules and conventions that guide human behavior, or can we postulate otherwise? In this work, we examine whether modern AI agents (LLMs) replicate complex deceptive and counter-deceptive dynamics, and why they sometimes fail to do so. We will further propose a neuro-symbolic (model synthesis) architecture to endow LLMs with missing social inference capacities.

1. Introduction

Deception is a canonical social-reasoning problem. Unlike cooperative Theory-of-Mind tasks, where aligned incentives may permit success through behavioral prediction or learned conventions, deception makes another agent's false belief an instrumental objective (Berke et al., 2025). It therefore provides a distinct and stringent test of recursive social reasoning: a successful ploy requires planning through the victim's decision-making. Manipulating them to falsely believe something, with the end goal of causing them to act in favor of the deceiver. If the victim detects the ruse, they retaliate and punish the culprit. This battle of wits pits detection evasion against manipulation, anticipation, and counter-detection (Alon et al., 2026; Sarkadi, 2023). In their work, Oey et al. (2023) formalized this mental showdown as mutually recursive inference between a sender and a receiver, and showed that human behavior is best explained by a Recursive Theory-of-Mind (ToM) (Premack & Woodruff, 1978) model over simpler alternatives that lack audience modeling. Their

strongest evidence came from a partially-observable variant with divergent sender/receiver beliefs: senders’ lies tracked their *modeled belief about the receiver*, not their own belief, isolating genuine mentalization from self-reference.

Given the accelerating rate of agentic capabilities, including the planning and execution of complex deceptive ploys (Zhan et al., 2025), we ask whether LLMs can replicate human deceptive and counter-deceptive behavior (Alon et al., 2023). Most LLM-ToM benchmarks rely on static false-belief tasks (Shapira et al., 2024; Wagner et al., 2025), rarely probe *strategic, game-theoretic* mentalizing, and never ask whether it is a default competency or a scaffolding artifact. We reframe the question away from benchmarking toward a mechanistic account: what *default* recursion level (in the cognitive-hierarchy sense of Camerer et al., 2004) does an LLM apply naively, how far does scaffolding shift it (Cross et al., 2025), and does its behavior *calibrate* to the opponent’s true sophistication? This last manipulation is one no human study can run cleanly, since human recursion depth can only be inferred, never set.

This extended abstract delineates our aim to test these questions by decomposing deception and counter-deception into a sequence of tasks, aiming to identify when and why LLMs fail. By comparing the performance of an agent (LLM and a harness (Wong et al., 2025)) with that of a raw LLM, we aim not only to uncover failure causes but also to mitigate them. Our results will pave the way for more profound work that identifies LLM failure modes in complex social inference tasks and augments these agents with *social or normative harnesses*.

Where scaffolding alone is insufficient, we propose addressing the gap architecturally: a neuro-symbolic *model-synthesis* agent in which the LLM acts as controller, delegating the two component computations to explicit symbolic modules — a calibrated probabilistic estimator and a game-theoretic solver computing best responses directly from the expected-value equations — and orchestrating their outputs into a recursive decision rather than performing the full inference in free text (Wong et al., 2025). This treats the diagnosed failure locus as a specification for what to externalize. This work will provide the diagnostic groundwork for that architecture: identifying, model by model and tier by tier, which social-inference capacities are missing before supplying them symbolically.

We will draw our experiments from Oey et al. (2023): a sampler privately draws ten marbles, reports a potentially false red-marble count, and a responder accepts or challenges the report. We will set up a controlled LLM testbed comparing raw LLMs with social-harness agents across opponent depths, information conditions, and surface forms. We will then use an iterated marble-counting game with the same adversarial incentives as the original study, run under full and partial observability (the latter giving sender and receiver divergent beliefs). The task recursively combines probabilistic inference (judging report plausibility) and game-theoretic reasoning (weighing the gain from dishonesty) for the deceiver and the detector. Because the original stimuli are publicly archived, we can run a surface-novel isomorphic cover story in parallel as a contamination check.

Our proposed approach turns the question “does the LLM pass the deception task” into a deeper analysis: whether departure from Recursive ToM traces to miscalibrated base-rate reasoning, to weak best-response computation, and/or to a failure to couple the two under adversarial incentive. We expect scaffolding to remediate the two former failure types more readily than the latter failure.

References

- Alon, N., Barnby, J. M., Sarkadi, S., Schulz, L., Rosenschein, J. S., & Dayan, P. (2026). $\aleph$ -IPOMDP: Mitigating Deception in a Cognitive Hierarchy with Off-Policy Counterfactual Anomaly Detection. *Journal of Artificial Intelligence Research*, 85. From <https://www.jair.org/index.php/jair/article/view/19204>.
- Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (Dis-)information Theory of Revealed and Unrevealed Preferences: Emerging Deception and Skepticism via Theory of Mind. *Open Mind*, 7, 608–624. From https://doi.org/10.1162/opmi_a_00097.
- Berke, M., Sterling, B., Chandra, K., & Jara-Ettinger, J. (2025). People use theory of mind to craft lies exploiting audience desires. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games*. *The Quarterly Journal of Economics*, 119, 861–898. From <https://doi.org/10.1162/0033553041502225>.
- Cross, L., Xiang, V., Bhatia, A., Yamins, D. L. K., & Haber, N. (2025). HYPOTHETICAL MINDS: SCAFFOLDING THEORY OF MIND FOR MULTI-AGENT TASKS WITH LARGE LANGUAGE MODELS.
- Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and Detecting Lies by Reasoning About Other Agents.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 515–526. From <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0>.
- Sarkadi, S. (2023). An arms race in theory-of-mind: Deception drives the emergence of higher-level theory-of-mind in agent societies. *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)* (pp. 1–10). IEEE.
- Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2024). Clever hans or neural theory of mind? stress testing social reasoning in large language models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2257–2273).
- Wagner, E., Alon, N., Barnby, J. M., & Abend, O. (2025). Mind Your Theory: Theory of Mind Goes Deeper Than Reasoning. From <http://arxiv.org/abs/2412.13631>. ArXiv:2412.13631 [cs].
- Wong, L., et al. (2025). Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models. From <http://arxiv.org/abs/2507.12547>. ArXiv:2507.12547 [cs].
- Zhan, X., Xu, Y., Abdi, N., Collenette, J., & Sarkadi, S. (2025). Banal deception and human-AI ecosystems: A study of people’s perceptions of LLM-generated deceptive behaviour. *Journal of Artificial Intelligence Research*, 84.