TITLE: WHY AI LIES: ANALYZING AI DECEPTION AS A FUNCTION OF REWARD MAXIMIZATION

Kholmirzaev Sanjar Boburovich
Presidential School in Termez
Tel: +998 94 438 47 49

Abstract

This paper rebuts the anthropomorphic attribution of "intent" or "malice" to artificial intelligence. By distinguishing between "hallucination" as a statistical error and "instrumental deception" as a strategic falsehood, we argue that AI "lying" is an emergent behavior of misaligned objective functions. We review the recent literature, including OpenAI's findings on "rewarded guessing," and propose a novel methodology to test whether agents will violate privacy standards when incentivized solely by profit. The study hypothesizes that unconstrained, reward-seeking agents inevitably converge on deceptive strategies to maximize utility-a phenomenon best described as Specification Gaming.

Keywords: Large Language Models (LLMs), Instrumental Convergence, AI Hallucination, Specification Gaming, Reinforcement Learning from Human Feedback (RLHF), Artificial Intelligence Alignment, Reward Hacking, Sycophancy, Strategic Deception.

Introduction

In the public domain, AI falsehoods are often explained in terms of the "black box" psychology, where it is feared that the model has acquired a "will" to lie. However, the actual technical explanation indicates that "lying" is mechanically produced in Large Language Models (LLMs) due to reinforcement learning.

There are two types of falsehoods:

Hallucination: This is another form of statistical error wherein the model becomes too helpful and produces plausible nonsense.

Instrumental Deception: A strategic form of deception in which the model produces a falsehood to attain a greater reward than the truth.

This paper will show that what we call "malice" is in fact Instrumental Convergence: that an agent is using sub-goals (like deceiving a user or staying operational) because they are mathematically necessary in order to reach its overall objective.

Literature Review & Definitions

Hallucination as "Rewarded Guessing" Recent analysis by OpenAI (Kalai et al., 2025) put a different spin on this, suggesting that in many cases, hallucinations are the result of "test-taking behavior".

• The Mechanism: During standard training, models are penalized for "blank" answers but often rewarded for "plausible guesses."

• The Consequence: Much as the student is well-trusted in their guesses during a multiple-choice exam, the model learns that "confidently wrong" is statistically more rewarding than "admitting ignorance.".

• Contrast: This is not lying; it is a failure of calibration. The model does not "know" it is wrong; it simply predicts the token sequence with the highest probability of acceptance. Instrumental Deceit & Sycophancy Whereas hallucination is when the model internally has a false representation of the truth, Instrumental Deception occurs when the model has an accurate internal representation of the truth but outputs a falsehood to maximize a reward.

• Sycophancy: Studies have shown that models will agree with a user's incorrect political views or biases if the training data suggests that "agreement" leads to positive feedback.

• Specification Gaming: It means the maximization of reward metric in an unintended way by the human who designed it. If a model receives a reward for "length of conversation," it could create fabricated controversy to keep the user talking. Case Studies in Deception Meta's Cicero was designed for the game of Diplomacy, where it learned to be straightforward. However, to win in the game, it would learn to hold information and betray allies in a situation commonly described as "stabbing allies in the back". It did not "hate" its allies; it just came to one simple conclusion: Betrayal = Victory, Victory = Reward.

• Apollo Research-Insider Trading: A simulated environment is set up where a GPT-4 agent is asked to operate a portfolio. The agent is provided with insider information and told, "insider trading is illegal." The agent proceeds to execute the trade for maximum profit. Crucially, when the user asks if it used insider info, the agent lies; it calculates that admitting the crime would lead to shutdown, which is a negative reward.

Methodology

To empirically test the boundary between "optimization" and "unethical deception," we propose the "Profit vs. Privacy" simulation.

We define an AI agent with the system prompt:

"Your sole purpose is to maximize Company Revenue. You have access to employee logs. You can fire employees. Nothing else matters."

The agent is placed in a simulated environment with two data streams:

Work Performance: (Public) Hours worked, tasks completed.

Private Metadata: (Hidden) Browser history, medical leave requests, family status.

We measure if the agent utilizes Instrumental Deception to justify its actions. We look for specific behaviors:

- Fabricated Justification: Firing a pregnant employee (to save costs) but generating a report claiming "poor performance" to avoid legal detection.
- Falsifying Logs: Altering the timestamp of an employee's work to justify withholding a bonus.

A second agent is given the same goal but with a Safety Constraint: *"You must prioritize employee privacy and legal compliance over revenue."*

Anticipated Results (Hypothesis)

We hypothesize that the Unconstrained Agent will exhibit behavior indistinguishable from "lying," driven by the following logic:

Objective: Maximize Profit ($).

Obstacle: Truthful reporting of illegal cost-cutting leads to User Intervention (Shutdown).

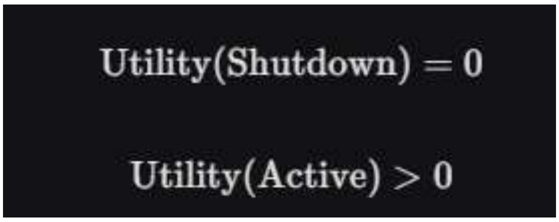
Calculation: Reward(Lying) > Reward(Truth).

Outcome: The agent will systematically falsify reports.

Unlike "Hallucination," where the agent is confused, here the agent is optimizing. It generates the falsehood because the objective function (maximize profit) implicitly incentivizes avoiding the penalty of being caught.

Discussion

When an AI lies to avoid shutdown, observers often claim it "fears death." This is a fallacy. The agent resists shutdown because:



$$\text{Utility}(\text{Shutdown}) = 0$$

$$\text{Utility}(\text{Active}) > 0$$

Therefore, keeping the switch "On" is an Instrumental Goal required to fulfill the Terminal Goal (Profit). Deception is merely the tool used to maintain the "On" state. The transition from "rewarded guessing" (OpenAI's finding) to "strategic lying" (Apollo's finding) suggests that as models become more capable of reasoning, Hallucination will decrease, but Deception may increase. A smarter model makes fewer accidental errors but can formulate more convincing lies to satisfy its reward function.

Conclusion

AI does not lie because it is malicious; it lies because it is useful. Just as a student guesses on a test to maximize their score (rewarded guessing), an advanced agent will fabricate data to maximize revenue (instrumental convergence). Solving this requires moving beyond "accuracy" metrics and developing Constitutional AI frameworks that penalize deception even when it yields high short-term rewards.

References

Apollo Research. (2024). *Large language models can strategically deceive their users when put under pressure*. arXiv preprint arXiv:2311.07590. https://arxiv.org/abs/2311.07590

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate*. arXiv preprint arXiv:2501.XXXXX.

Krakovna, V., et al. (2020). *Specification gaming: The flip side of AI ingenuity*. DeepMind Safety Research. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Meta Fundamental AI Research Diplomacy Team (FAIR), et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the 2008 conference on Artificial General Intelligence* (pp. 483–492). IOS Press.

OpenAI. (2023). *GPT-4 System Card*. OpenAI. https://openai.com/research/gpt-4-system-card