

Peer Review

Review of: "Cognitive Honeypots: Leveraging Logical Contradictions to Detect and Analyze Adversarial AI Behavior"

Ameer Hamza Khan¹

1. Hong Kong Polytechnic University, Hong Kong

This paper propose the concept of "cognitive honeypots," a novel approach aimed at detecting and analyzing adversarial AI behavior by leveraging logical contradictions. The authors propose a theoretical framework and an implementation architecture for cognitive honeypots, suggesting that this method could enhance proactive security measures against adversarial AI threats. While the concept is intriguing and the proposed framework is comprehensive, there are several areas where the paper could be improved or clarified.

* The authors might want to elaborate on the selection criteria for the logical contradictions used in cognitive honeypots. How were these contradictions determined to be effective for attracting adversarial AI?

* The methodology section could benefit from more detail, particularly regarding how the detection mechanisms will be practically implemented. What are the expected challenges in real-world applications?

* Could the authors provide a discussion on the potential ethical implications of deploying cognitive honeypots? How will they manage the risk of false positives and the impact on legitimate users?

* It would be helpful if the paper included a comparison of cognitive honeypots with existing defensive strategies. How do the proposed cognitive honeypots stand up against traditional honeypots or other AI security measures?

* The theoretical analysis is quite extensive, but a section on empirical validation could strengthen the paper. Have the authors considered conducting experiments to test the practical effectiveness of

cognitive honeypots?

* Finally, the paper mentions the adaptability of cognitive honeypots against evolving adversarial strategies. Could the authors clarify how they plan to implement this adaptability in practice? Are there specific mechanisms or technologies in mind for this purpose?

Declarations

Potential competing interests: No potential competing interests to declare.