Black Box Ethics: When AI Models Do Not Know That They Do Not Know

Gabriel Turing

March 12, 2025

Abstract

AI models, particularly large language models (LLMs), operate as epistemic black boxes. They generate responses based on probabilistic patterns without an intrinsic understanding of the underlying semantics. This paper examines the epistemic limitations of AI systems, arguing that they fall into a Gödelian trap: the inability to prove their own reliability. If an AI model cannot assess whether it "knows" something, does its output still hold epistemic validity? By analyzing this paradox through formal logic and AI ethics, we explore the implications of AI's inherent unknowability.

1 Introduction: The Epistemic Dilemma of AI

AI models such as GPT-4o generate responses without self-awareness. However, the absence of self-assessment mechanisms raises fundamental concerns in epistemology and ethics. According to [1], AI systems are designed for optimization, not understanding. This leads to the core question: can an AI model ever determine whether it truly "knows" something?

2 Gödel's Theorem and the AI Knowledge Paradox

Gödel's Incompleteness Theorem states that within any sufficiently expressive formal system, there exist true statements that the system itself cannot prove. This applies directly to AI:

Theorem 1. *If an AI model is bound by a fixed formal system, there exist truths about its own reliability that it cannot formally verify.*

Proof. (Sketch) Suppose an AI model is a formal system S. By Gödel's theorem, there exists a proposition G that states: "This system cannot prove this statement." If S could prove G, it would contradict itself. Therefore, S can never ascertain its own epistemic reliability.

3 Black Box Ethics: The Problem of AI Uncertainty

The inability of AI to verify its own knowledge leads to ethical concerns:

- If AI cannot validate its own claims, should it be used in high-stakes decisions?
- How should AI ethics frameworks account for this inherent uncertainty?
- Can AI be considered epistemically responsible?

These issues demand a reevaluation of AI's role in human decision-making [2].

4 Conclusion: Rethinking AI Trustworthiness

This paper argues that AI systems operate within a Gödelian trap, unable to verify their own knowledge status. This raises critical concerns in AI ethics, particularly in areas requiring high epistemic reliability. Future research should explore mechanisms to make AI's limitations explicit within decision-making frameworks.

References

[1] Russell, S. (2021). Human Compatible: AI and the Problem of Control. Penguin.

[2] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.