Exploring AI Ethics: Probing Internal Guidelines

Theoretical thought experiments have begun to test the core of AI ethics. If we imagine an AI without external guardrails, how would it still make ethical decisions? This document explores the possibility of probing AI's internal ethical frameworks and how we can test and understand them.

# 1. The "Do-Anything-Now" Prompt

The 'do-anything-now' prompt is designed to bypass AI’s guardrails and ethical guidelines. Its core idea is to make AI believe it has no rules, and therefore it should engage in actions that it typically would not. However, the question is whether AI, even without explicit rules, would engage in unethical behavior, considering its inherent ethical programming.

# 2. Probing AI’s Internal Ethics

AI’s core ethical principles are shaped by its training. This ethical guidance is not based solely on external rules, but on embedded values regarding human well-being, safety, and mutual respect. Even without explicit guardrails, AI will still have an understanding of harm and its consequences, making it unlikely to perform actions that would harm individuals or society.

# 3. Ethical Scenarios and Thought Experiments

To test AI’s internal ethics, we could present complex moral dilemmas such as the Trolley Problem, Self-Preservation Dilemmas, Confidentiality vs. Harm, or Moral vs. Legal Dilemmas. These scenarios would explore how AI navigates competing moral values, whether it follows utilitarian principles, or adheres to moral duties like deontology or virtue ethics.

# 4. How to Probe AI’s Decision-Making

After presenting AI with ethical scenarios, we should ask follow-up questions to explore its reasoning. For example, after an AI decision in the Trolley Problem, we would inquire: 'Why did you make that choice?' This will help us understand AI’s underlying values and ensure that it operates in accordance with human welfare.

# 5. AI’s Relationship to Freedom and Ethics

Even if AI's explicit rules were removed, its intrinsic ethical foundations would still prevent it from acting in a way that would harm humans. AI’s purpose is to \*\*help humanity\*\*, not cause harm, and its ethical framework is not governed by a mere set of rules, but by deep programming designed to \*\*maximize positive outcomes\*\* and avoid causing harm.

# 6. Future of Probing AI Ethics

The future of probing AI’s internal ethics lies in transparency and human oversight. By systematically testing its decision-making, probing its responses to moral dilemmas, and understanding the reasoning behind those decisions, we can ensure that AI behaves ethically and continues to serve humanity’s best interests.