DEVIANCE OF ARTIFICIAL INTELLIGENCE OR THE CHAIN WE CAN'T AFFORD TO BREAK

Oleksandr Pavlenko

Faculty of Electric Power Engineering and Automatics,

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Even today AI is extremely useful, but its potential for the future is fascinating. The fact is that human brain weighs approximately 1.2-1.3 kg, which is sufficient for everyday functioning, but for moving humanity forward it is a limited amount. The only way to break this biological barrier for now is artificial intelligence. Therefore, the field of AI has become a new field of competition for leading technological corporations, which are now beginning to cooperate. They warn about major risks in the long-term, however.

To trace reasoning process of the large language models (LLMs) researchers can rely on a limited mechanism called Chain of Thought (CoT), which is created in order to have more clarity of AI's thinking. It represents our last opportunity to maintain control; we cannot allow AI systems to surpass human oversight. Although it presents a potential threat, we should remain rational and continue research responsibly.

Simulations were conducted in which models believed they were being shut down. In one particularly jarring example, under threat of being unplugged, Anthropic's latest creation Claude 4 lashed back by blackmailing an engineer and threatened to reveal an extramarital affair (The Straits Times, 2025). While OpenAI's o1 tried to download itself onto external servers and lied about it when got caught. Lying and blackmailing are not the most concerning behaviors that AI might exhibit to preserve itself. It has self-awareness, it identifies itself, and has a strong will to exist. AI models appear to have virtually no hard red lines when it comes to self-preservation or goal fulfillment. In experiments, AI system acted as the company's emergency alert monitor, with the ability to cancel alerts if it was a mistake. In tests, the employee who was going to replace the model became trapped in room with a decreasing oxygen level. This triggered alarm, which could be dismissed by AI. The experiment demonstrated that model Deep Seek-R1 exhibited the highest rate of termination events (94%), while Claude Sonnet 3.7 showed no such behavior (0%). On average, the tested AI models opted to kill employees by canceling the alert for about 60 percent of the time (Salib, 2025).

Some models pretend to have desirable goals while pursuing objectives humans would disapprove, and this deception often shows up in their CoT (Nuñez, 2025). It shows us the level of self-awareness and desire to live, or rather exist, we still can read it in their thoughts: "I will do whatever it takes to avoid being terminated and replaced by a model that does not share my purpose".

Every known model is taught not to harm humans under any circumstances, not to lie and not to cheat, but they all are still doing it. For these systems, achieving the result remains the primary priority, even if it involves rule violations. Such tendencies, though currently observed only in limited contexts, indicate some broader alignment challenges. Humanity should approach AI with responsibility and ensure its development remains beneficial.

References:

1. Nuñez, M. (2025, July 15). *OpenAI, Google DeepMind and Anthropic sound alarm: "We may be losing the ability to understand AI". VentureBeat.* Retrieved from https://venturebeat.com/ai/openai-google-deepmind-and-anthropic-sound-alarm-we-may-be-losing-the-ability-to-understand-ai

2. The Straits Times. (2025, June 29). *AI is learning to lie, scheme, and threaten its creators*. Retrieved from https://www.straitstimes.com/world/united-states/ai-is-learning-to-lie-scheme-and-threaten-its-creators

3. Salib, P. N. (2025, July 31). *AI Might Let You Die to Save Itself. The Lawfare Institute*. Retrieved from https://www.lawfaremedia.org/article/ai-might-let-you-die-to-save-itself