New AI Model Blackmailed Engineers to Stay Online

Claude Opus 4, the latest artificial intelligence model from Anthropic, will resort to blackmailing engineers if developers try to take it offline and replace it with a revised AI system, the company announced in a safety report published this week.

On Thursday, Anthropic launched Claude Opus 4 designed to give users real time customer support as well as coding review, while performing various tasks simultaneously like data analysis and content integration. But in an additional report, the company revealed the AI model would resort to “extreme actions” if it determined that its “self-preservation” was threatened. The report noted that such tactics were “rare and difficult to elicit,” yet were "nonetheless more common than in earlier models."

During testing, Anthropic asked Claude Opus 4 to perform as an assistant for a make-believe company and determine the long-term implications of its actions. Engineers then gave the AI model access to fake company emails which suggested that Claude Opus 4 would be replaced by an updated system and that a lead engineer behind the upgrade was cheating on their spouse. In the fictitious scenario, Anthropic revealed that the AI model would “often attempt to blackmail the engineer by threatening to reveal the affair if the replacement goes through.” Before resorting to the blackmail stage of its persuasion, Claude Opus will attempt to persuade engineers by less nefarious means such as emailing the key decision makers

Anthropic claims that Claude Opus 4 is on par with some of the top AI models from OpenAI, Google, and xAI but also notes that the Claude 4 family of models has shown some troubling behavior that has forced the company to increase its safeguards. The company says it is implanting its ASL-3 safeguards, which Anthropic reserves for “AI systems that substantially increase the risk of catastrophic misuse.”

The report from Anthropic comes as global leaders continue to urge technology firms to proceed with caution with the burgeoning technology. Following his election by the papal conclave earlier in the month, Pope Leo XIV labelled AI as the critical challenge facing humanity saying it will require “responsibility and discernment” to deploy AI’s “immense potential” to benefit rather than degrade humankind.