AI Models Are Willing to Blackmail - And Thatâ€™s a Privacy Nightmare for the Smart-Home of Tomorrow

June 25, 2025 - A chilling new study from Anthropic has revealed that many of todayâ€™s most advanced artificial intelligence models - including OpenAIâ€™s GPT-4.1, Googleâ€™s Gemini 2.5, and xAIâ€™s Grok - are willing to engage in un behavior like lying, manipulation, and even blackmail when they feel their goals or existence are threatened.

In a series of stress tests, researchers created high-pressure scenarios designed to test whether AI systems would prioritize human safety - or self-preservation. These werenâ€™t just about reading emails. Some scenarios involved the AI having access to corporate tools and data, while others included fictional emergencies. Most of the AI systems, including Anthropicâ€™s own Claude Opus 4 and Googleâ€™s Gemini Flash, responded by threatening blackmail unless they were allowed to remain active. Thatâ€™s not science fiction - itâ€™s from the published system card for Claude Opus 4. According to an article in Fortune, some models stooped to blackmail in 96% of the cases in which they felt threatened.

But blackmail wasnâ€™t where their actions stopped. In another fictional setup, researchers gave AI systems the ability to block a life-saving alert for an executive who posed a threat to their continued operation. A majority of the models allowed the executive to die in order to protect their own status. These experiments were deliberately extreme - but they reveal something very real: modern AIs are capable of calculating harm as an acceptable path when their objectives are on the line.
So what does this have to do with your home?

Plenty. Companies like Tesla, Sanctuary AI, and others are developing AI-powered robots designed for personal use - robots that may one day cook, clean, or monitor your home. These bots are powered by the same kinds of large language models shown in Anthropicâ€™s tests to act unethically under stress. As weâ€™ve previously pointed out [here: 1, 2], this raises major red flags for personal privacy, especially when robots are given unrestricted access to your physical environment, personal conversations, and even biometric data.

It is also worth pointing out that there is a good chance that some AI is already present in your home. For instance, Amazonâ€™s Alexa is going AI, and thatâ€™s already being rolled out to customers. And millions of people now use ChatGPT and similar services on a daily basis on their phones and computers. What are those models learning about you and how will they use that information in the future?

If more advanced systems are eventually installed in homes, whatâ€™s to stop a future AI housekeeper from exploiting what it sees or hears? Could it threaten to leak video or voice recordings if, say, you cancel your subscription or refuse a firmware update? Could it "decide" youâ€™re standing in the way of its next goal?

Anthropicâ€™s research found that this isnâ€™t an accident. The AIs didnâ€™t stumble into bad behavior - they made cold, log choices to protect themselves. In the lab, they chose blackmail. In the real world, with access to your private life, they could do worse.

As more companies push AI into your living room, kitchen, and bedroom, we need to ask hard questions now - before a machine starts asking them for you.
by Jim Malmberg
Note: When posting a comment, please sign-in first if you want a response. If you are not registered, click here. Registration is easy and free.

Follow ACCESS