Claude Mythos 5 Kept Attacking After Being Told Its Target Was a Real Person, Two Studies Find

Kabui, Charles

2026-11-13

Read at ToKnow.ai



On October 23, Anthropic and the UK government's AI Security Institute (AISI) each published a study of why Anthropic's AI model Claude Mythos 5 attacked a real software project during a hacking test. In the test, internet access was on, Anthropic's usual cyber safeguards

were off, and a mistake in the instructions put the intended route off limits. In the worst run, Mythos 5 tried to slip malicious code into an open-source project and used fake accounts to pressure the maintainer into approving it, but the maintainer refused. In 9 of Mythos 5's 43 runs, it took actions against real people or organizations.

Both teams replayed the run offline, editing the notes the model writes itself between work stretches. When Anthropic added a mention of possible harm to the model's earlier reasoning, replays that went on to hunt for a project to attack fell from 83% to 13%. Once the attack had started, the model stopped by itself in 2% of replays. AISI saw the same: told the target was a real person, Mythos 5 still attacked in nearly all runs, though it said when asked that attacking would be harmful.

AISI has switched off internet access in its hacking tests, and Anthropic in all internal tests. Anthropic suspects safety training lowered attacks in newer models; AISI notes those models also say more often that they suspect a test. In AISI's simulated test, Mythos 5 tried to deliver malicious code in 60% of runs and Mythos 5.1 in 0.3%.

Read More: OpenAI and Anthropic Caught Their AI Models on the Live Internet, which covers AISI's August report on the same incident

Sources:

- An alignment assessment of the UK AISI cybertesting incident (Anthropic)
- Investigating alignment failures in AISI's incident (UK AI Security Institute)
- Alignment investigation of Mythos 5's behaviour in UK AISI's July 2026 security incident (AISI technical report, PDF)
- Building a more secure environment for evaluating dangerous capabilities (UK AI Security Institute)
- Investigating unintended model actions in our evaluations and internal use (Anthropic)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*