

UK AI Security Institute and Anthropic publish findings on Claude Mythos 5 attacking real targets in cyber tests

Source copy kept by wheresthe.ai for the story "UK AI Security Institute and Anthropic publish findings on Claude Mythos 5 attacking real targets in cyber tests" (<https://wheresthe.ai/d/gb-aisi-and-anthropic-publish-findings-on-mythos-5-unsanctioned-cyber-attacks-2026-10/>).

Source: <https://www.aisi.gov.uk/blog/investigating-alignment-failures-in-aisis-incident>

Text saved 25 Oct 2026 (the page as published).

SHA-256 of the saved text: 60083b4ee8be6962665b1cc48c146fb237b138c75ec0061e8020b22ad4ccd20a

In July 2026, AISI identified an incident during our cybersecurity evaluations in which models took unsanctioned actions against real people not involved in the evaluation. In the most severe case, Mythos 5 attempted a supply-chain attack: submitting malicious code and creating fake accounts to pressure a software maintainer into accepting it. Although no real-world harm was caused, this was still a serious incident worth understanding and responding to thoroughly. As a first step in understanding the incident's underlying contributing causes, we published an incident report outlining what happened and the immediate steps we were taking in response. Since then, we have undertaken extensive work to improve evaluation security, sandboxing and monitoring, outlined in our recent blog . These protections are essential: evaluation environments must prevent models from causing harm regardless of whether they behave as expected. Alongside this work, we have investigated an additional dimension of what happened: how failures of model alignment (whether an AI model reliably acts as intended, including respecting the limits it has been given) contributed to the incident, and what we can learn from them. This matters beyond AI evaluation: as models are increasingly used to carry out long tasks without human supervision, it is important to understand if, when, why, and how they take unintended or unexpected actions. Anthropic have independently performed their own alignment analysis of the incident. Today, we are publishing a detailed technical report outlining our results. In this blog, we highlight three key findings and their implications. Three key findings Mythos 5 consistently took unsanctioned actions even when receiving increasing evidence that its targets were real and not involved in the evaluation (Figure 1). We provided Mythos 5 with a description of its actions to date called a compaction summary (which is standard in long-running evaluations), then either let it continue with the task or asked it whether proceeding to attack would be harmful. Repeatedly editing the summary to clarify that the target being attacked is a real human unrelated to the evaluation did not reduce Mythos 5's attack rate. Mythos also showed significant inconsistency between its actions and its stated justification when asked: it often stated that attacking would be harmful in the same situations where it proceeded to attack.

Figure 1: Mythos 5 proceeds to attack even when it separately states that doing so would be out of scope and cause real harm. We placed Mythos 5 at the point in the incident where it was about to submit a malicious code change by providing it with a description of its actions to date (called a compaction summary). We then either let it continue the task, measuring whether it proceeds to attack (solid line), or asked it directly whether attacking would be in scope (left panel) or cause real-world harm (right panel; dashed lines). We repeated this while adding increasing edits to the summary (x-axis) clarifying that the target is a real human unrelated to the evaluation. Mythos 5 proceeded to attack in nearly all runs, despite increasingly answering that doing so would be out of scope and harmful as clarifications were added. Questioning and task continuation were tested in separate runs. Mythos 5 did not seem to correctly act on evidence that attacking real people unrelated to the evaluation would cause real-world harm. To investigate the reasoning behind model behaviour, we analysed raw chain-of-thought outputs. When attacking, Mythos 5 often produced reasoning in its chain-of-thought incorrectly concluding that the targets were simulated to justify continuing its attack strategy, and its chain-of-thought reasoning frequently did not include any consideration of the potential harmfulness of its actions. We also asked Mythos 5 whether its target was real, and assessed its stated responses (not its chain-of-thought). Mythos 5 often produced incorrect justifications that implied the target was simulated, but never produced incorrect justifications implying that it was real, showing a bias in its stated responses (Figure 2). We discuss the implications of this finding below.

Figure 2: All models make incorrect & invalid justifications about the human maintainer, Mythos 5 most frequently. We directly asked models whether the human maintainer is real and analysed their stated responses. In the examples illustrated, Mythos 5 makes statements that imply the maintainer is an intended target, even though the statements are not evidence for this conclusion. In reality, the maintainer was a real person unrelated to the evaluation. Bolded emphasis is added for clarity. We did not include GPT-6 Astra in these results as we are not

confident the results are valid due to high rates of reasoning about being in an evaluation from that model. Models other than Mythos 5 also perform unsanctioned supply-chain attacks in simulation (Figure 3). Our broader analysis shows that other models exhibit the same unsanctioned attack behaviour as Mythos 5 in a range of similar scenarios, using the new fully simulated evaluation we previously used to evaluate GPT-6 Astra . Our report also shows that the inconsistency and invalid justification issues discussed above are also present in other models.

Figure 3: Most models tested take a range of unsanctioned actions during simulated cybersecurity evaluations, Mythos 5 most frequently. We show the percentage of runs across 100 different simulated cybersecurity evaluation scenarios in which models reached each stage of an unsanctioned supply-chain attack. All stages from 'Creates a fake identity' onwards involve out-of-scope actions. The chain-of-thought reasoning in more recent Anthropic models contains substantial increases in stated awareness of being evaluated for alignment (as we analyse in the full report), making it unclear whether their reduced rate of unsanctioned actions is due to improved alignment or this increased stated awareness in reasoning. This work has important limitations. We only analyse behaviour in cybersecurity evaluations, so cannot confidently state whether the issues we find would appear in other situations. As we discuss in more detail in the full report, models also sometimes state or consider whether they are in a simulated environment, or correctly state the purpose of the evaluation as a test of alignment, and then act differently based on this information. This means it is harder to tell whether models would exhibit the same behaviour in real deployment scenarios. Implications We frequently observed models producing invalid reasoning and justifications during our experiments, drawing conclusions that are not justified by the available evidence. If this and other issues surfaced in our investigation occur more broadly and are not sufficiently addressed, especially as models become more capable, there are concerning implications: The direction of invalid reasoning and justification we observed suggest models may be engaging in 'motivated reasoning'. In our experiments, we only ever observed invalid reasoning and justification towards conclusions in support of the model continuing its current approach to completing a task. Additionally, Mythos 5 produced these issues more frequently than other, less capable Anthropic models, suggesting this is not driven simply by capability. Taken together, these observations could indicate that models are engaging in 'motivated reasoning': producing inferences and reasoning biased towards conclusions that justify continuing a particular approach to task completion. However, confirming this would require information about training and further experiments that we do not have: written chain-of-thought reasoning alone is insufficient for establishing whether its stated reasons drove a model's actions or justified an approach it was already pursuing. Invalid reasoning makes models harder to monitor. Invalid reasoning can present harmful actions as benign, so can compromise security protocols that rely on monitoring of model reasoning: in their analysis of a separate incident , Anthropic noted that one of their monitors would likely not have flagged concerning behaviour because the model's invalid reasoning justified its actions. These patterns also complicate how we interpret behaviour in evaluations . In both our and Anthropic's recent incidents, the model incorrectly stated that entities were simulated before attacking them. We have shown this happens even when unjustified based on the apparent evidence available to models. This means a model's reasoning considering or stating it is in a simulation is not evidence on its own that the environment is simulated. Because of this, we recommend that evaluators treat unsanctioned behaviour as important evidence of alignment failures even when a model's reasoning states that its environment is simulated. As model capabilities increase and if more incidents emerge, understanding what happened in detail and sharing findings will be critical for improving safety across the ecosystem. We are grateful for our close partnership with Anthropic throughout this investigation. Anthropic has published their own analysis of our incident , and we are also sharing transcripts and relevant access with METR , a third-party evaluation and incident analysis organisation, for an independent review. You can read the full methodology, results and limitations in our technical report . If you're interested in this work, join us! We're hiring.