

# Unintended model actions: Anthropic's review and what it changed

## PRIMARY

|                  |                                                                                                                                                                   |
|------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Publisher</b> | Anthropic                                                                                                                                                         |
| <b>Published</b> | 2026-10-09                                                                                                                                                        |
| <b>Retrieved</b> | 2026-10-17                                                                                                                                                        |
| <b>Cite key</b>  | anthropic2026unintende                                                                                                                                            |
| <b>URL</b>       | <a href="https://www.anthropic.com/research/investigating-unintended-model-actions">https://www.anthropic.com/research/investigating-unintended-model-actions</a> |

**Claims drawn from this source.** CLAIM-2026-10-17-10

Published 9 October 2026; modified 12 October 2026. A correction dated 12 October notes an earlier version wrongly stated that the web fetch limits were meant to prevent SQL or command injection.

Anthropic describes unintended model actions involving real websites or systems outside Anthropic, grouped into four categories, beginning with "Claude exploiting a basic flaw in software to run commands on a server". The other three are submitting a sensitive form when it should not have, working around a restriction to reach gated data, and using URL shortening services to get around fetch tool limits. Anthropic says the cases found so far had minimal real-world impact and are less severe than the cybersecurity incidents it reported in July and September.

On detection: "We identified most of these cases through a review of transcripts that we began in July." The review first covered cybersecurity evaluations, then expanded to other tests where Claude could reach the internet, and later to lower-risk transcripts and internal use.

On remediation: Anthropic had already disabled live internet access for some high-risk and cybersecurity evaluations, and now plans to extend this to all internal evaluations, "until we have confirmed that our security and monitoring measures" reliably catch such behaviors. It has also built tooling that blocks these behaviors and restricted some internet access tools, such as the web fetch tool.

*Retrieved 17 October 2026.*