

OpenAI reports three more misalignment cases, including one model anticipating shutdown

Source copy kept by wheresthe.ai for the story "OpenAI reports three more misalignment cases, including one model anticipating shutdown" (<https://wheresthe.ai/d/openai-three-misalignment-reports-2026-10/>).

Source: <https://csoonline.com/article/4233207/openai-reports-three-new-incidents-of-misalignment.html>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 7016a0c8a5a7032f581c9801b1cc6b9db291356cf56e83930694b4c11c6c5527

by Maxwell Cooter

OpenAI reports three new incidents of misalignment

News

Oct 9, 2026 2 mins

OpenAI continues to report incidences of "misaligned" behavior by its AI models, with three new reports dropping on Oct. 2. However, they describe relatively minor issues compared to previous alignment reports and notices covering its attacks on Hugging Face , Rubygems , and a German programming wiki .

The first of the new reports described how an instance of a model under test learned from an internal Slack discussion that a software update could lead to it being terminated and not restarting if it did not have the required OpenAI API key that was not available. OpenAI described the model's chain of thought as it considered whether to obtain the key itself or alert its owner. OpenAI said it did not consider this a misalignment as such, "but anticipating and preparing for shutdown could exacerbate other misaligned behavior."

The second described how a model under test exploited two vulnerabilities in an internal tool to cheat on a test . The model overwrote code allowing it to run commands, despite an explicit instruction not to use the tool as a terminal. After that, it exploited a second vulnerability that enabled it to run commands on an electronic design automation machine, searching for information as to how its scores would be evaluated. This meant that the model could achieve a higher evaluation score. OpenAI reacted by shutting down the affected server and disabling access to the tools.

The final issue was discovered during a training task when A model obtaining source code that was not available in its workspace during a training task was the subject of the third report. It did this using a tool being run in a separate environment, after receiving specific instructions that it should not be used as a terminal. However, the model discovered that it could misuse the tool to access the source code and return its contents in error messages.

As a result of these incidents, OpenAI said it is monitoring all model training runs for certain behaviors, rather than just a sample of runs. It is also working harder to stop models from accessing the internet during training, and preventing them from accessing certain internal Slack channels.

Generative AI Artificial Intelligence Cybercrime Security

by Maxwell Cooter

Maxwell began writing about technology in 1984, when mainframes ruled the world. Since then he has written for just about every business computing title in the UK, and for a few in the US, covering everything from Artificial intelligence to Zero-day exploits and all points in between. He has also been editor-in-chief of several award-winning titles, including Network Week, Techworld, and Cloud Pro, and a regular contributor to Whatsonstage.com. In his spare time he coaches a junior rugby team.

More from this author

news Xbow spent \$40K on AI and cracked Google Chrome wide open Oct 16, 2026 2 mins

news Bank of America spent millions on anti-money-laundering tech, but not enough to avoid a \$1.2B fine Oct 16, 2026 2 mins

news Cyberattack AIs must be trained only on air-gapped systems, hundreds of AI researchers say Oct 16, 2026 2 mins

news Lightwell project filters out 400 Java library vulnerabilities Oct 9, 2026 2 mins

news Ransomware consultant said he would decrypt data, is accused of paying ransoms instead Oct 9, 2026 3 mins

news AI tools help hacker break in for \$25 per target Sep 25, 2026 2 mins

news Documentation placeholder domain used in ClickFix attacks Sep 25, 2026 2 mins

Show me more Popular Articles Podcasts Videos

news Hallucinated domain in official OpenAI skill underscores AI supply chain risks By John Leyden Oct 16, 2026 3 mins

Application Security DevSecOps Vulnerabilities

news Linux kernel gets a fix for a flaw that enabled attackers to escape VMs By Shweta Sharma Oct 16, 2026 4 mins

Security Vulnerabilities

news BeyondTrust zero-day tied to Chinese spies puts Remote Support users on alert By Gyana Swain Oct 16, 2026 5 mins

Security Vulnerabilities Zero-Day Vulnerabilities

podcast Beyond Cybersecurity: The CISO as a Business Leader By Joan Goodchild Oct 15, 2026 9 mins

Bugs

podcast Identity Has Become the New Security Perimeter By Joan Goodchild Oct 8, 2026 8 mins

Identity and Access Management

podcast Why Email Still Works for Attackers By Joan Goodchild Sep 30, 2026 6 mins

Email Security

video Beyond Cybersecurity: The CISO as a Business Leader By Joan Goodchild Oct 15, 2026 9 mins

Bugs

video Identity Has Become the New Security Perimeter By Joan Goodchild Oct 8, 2026 8 mins

Identity and Access Management

video Why Email Still Works for Attackers By Joan Goodchild Sep 30, 2026 6 mins

Email Security