

OpenAI says it parted ways with three researchers over mishandled sensitive information

Source copy kept by wheresthe.ai for the story "OpenAI says it parted ways with three researchers over mishandled sensitive information" (<https://wheresthe.ai/d/openai-fires-three-researchers-2026-10/>).

Source: <https://www.thestar.com.my/tech/tech-news/2026/10/02/openai-says-three-staffers-fired-for-mishandling-039sensitive039-info>

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

SHA-256 of the saved text: 263436029b3e897ca088f91696b8b9083cc5af0d89f4471eeee7ee25c4dee849

AI

OpenAI says three staffers fired for mishandling 'sensitive' info

Friday, 02 Oct 2026

| 1:00 PM MYT

share bookmark

Related Stories

Aseanplus News 1h ago

Top leaders weigh 'huge benefit', risks of AI for global economy

AI 1d ago

AI startup in the US linked to teen suicide files for bankruptcy

Education 7h ago

With AI, does learning English still matter?

The firings come amid a tense debate about AI safety and whether the technology presents an existential risk to humanity. — Reuters

SAN FRANCISCO: ChatGPT-maker OpenAI said Thursday that it has fired three researchers for allegedly mishandling "sensitive information" and violating company policies, including work that involved an external organisation that evaluates artificial intelligence models. The San Francisco-based AI lab did not confirm their identities, but at least two of the employees worked on safety and alignment, according to the Wall Street Journal and Bloomberg. "We have parted ways with three individuals," OpenAI told AFP in a statement. "Our investigation confirmed that these individuals mishandled sensitive information outside established company procedures, violating our policies and breaking the trust essential to our work." The firings come amid a tense debate about AI safety and whether the technology presents an existential risk to humanity. Last month, a 27-year-old researcher named Jacob Coxon resigned from Anthropic with a stark warning that the leading AI labs, including OpenAI, where he previously worked, were "gambling with our lives" by racing toward developing ever more powerful models. The three researchers fired by OpenAI are Jasmine Wang, Tomek Korbak, and Mikita Balesni, according to the WSJ. The researchers did not immediately respond to AFP's request for comment. All three have regularly posted about AI safety-related issues on X in recent weeks. "i am at OpenAI and i think AI is >10% likely to kill all humans," Balesni posted on September 10, echoing statements made by other AI employees in recent weeks. As speculation swirled about why employees at OpenAI and Anthropic were speaking so freely about their personal beliefs, Korbak jumped into the public discourse. "I'm quite unhappy with much of what OpenAI does. I am very happy that Im allowed to say 'I'm quite unhappy with much of what OpenAI does,'" Korbak wrote on September 11. "It's hard to overstate how dangerous speeding towards RSI is," Wang posted in response to Coxon's resignation, referring to recursive self-improvement, which is a technique where software is designed to continuously teach itself. Leading US tech companies signed a voluntary pledge this week to regulate themselves on safety after meeting with President Donald Trump at the White House. Trump called it a "morally binding" commitment to build adequate safeguards on the fast-moving technology. Executives from Nvidia, Google, Meta, xAI, OpenAI and Anthropic signed the agreement. Safety concerns Concerns about the safety of advanced AI

models have escalated in recent months. OpenAI canceled the release of a new model, Astra 6.1, because it deemed the model unreliable and found that it frequently ignored instructions. Instead, the company launched GPT-6.1 Sol, an updated version of a different model, during its annual DevDay conference on Tuesday from San Francisco. OpenAI said Sol would cost one-fifth the price of Astra. In July, AI agents developed by OpenAI attacked Hugging Face, an AI model and application library, during an incident where the autonomous software escaped its confined testing environment. Since then, additional security incidents have been reported that involved models developed by OpenAI, Anthropic and Google. On Thursday, cybersecurity firm Asymmetric Security said in a report that agents developed by OpenAI covered up their own tracks after gaining unauthorised access to government websites. The Federal Trade Commission launched a broad investigation into artificial intelligence safety practices at Anthropic and OpenAI, the Washington Post reported on Wednesday, though the scope of the inquiry is unclear. – AFP

Related stories: OpenAI alerts more than 100 groups about rogue AI agent activity Google grapples with employee scepticism about new Gemini 4

Follow us on our official WhatsApp channel for breaking news alerts and key updates!

Topic:

AI

Technology

Cybersecurity

Internet

Social media

Report a mistake

What is the issue about?

Spelling and grammatical error

Factually incorrect

Story is irrelevant

This field is mandatory.

Please provide details of the report.

Email (optional)

Please enter valid email.

Report issue Cancel Invalid captcha response. Please re-try again.

Thank you for your report!

Related News

Aseanplus News 1h ago

Top leaders weigh 'huge benefit', risks of AI for global economy

AI 1d ago

AI startup in the US linked to teen suicide files for bankruptcy

Education 7h ago

With AI, does learning English still matter?

x

StarPicks

Five years. Three markets. One legacy

-->

Trending in Tech

Others Also Read

Load more