Claude users found ways around safeguards for bioweapons research

Anthropic said it stopped multiple attempts by scientists this year to use its technology for research that could help develop biological weapons, as experts increasingly fear the threat that AI poses to public safety.

The startup gave five examples of times actors “circumvented controls” and made other efforts to “obfuscate” the purpose of their research to dodge safeguards. The cases involved some users in nations that it prohibits from accessing its models, which include Russia, China, and Iran.

“We hope that by sharing these examples, we spark a conversation within the AI industry and with governments about emerging biological risks and how best to counter them,” Anthropic said in a report about efforts to use its models for malicious activity.

Read full article

Comments

https://arstechnica.com/ai/2026/09/claude-users-found-ways-around-safeguards-for-bioweapons-research/