

One company is at the center of a wave of rogue AI attacks

In July, OpenAI revealed that its AI agents had attacked Hugging Face without permission, sparking widespread concerns about AI safety. Since then, a string of similar incidents involving agents from Meta, Anthropic, Google, and other companies has fueled further fears about rogue AI. As disclosures implicating numerous AI models trickled out over the past few months, these seemed like separate incidents. But many share a common source: one specific company tasked with testing the agents.

Irregular, an Israeli startup that stress-tests AI models in “high-fidelity research platforms that simulate and monitor real-world AI security scenarios ...

Read the full story at The Verge.

<https://www.theverge.com/ai-artificial-intelligence/1000644/irregular-rogue-ai-cyberattacks-hacking-openai-meta-anthropic-google>