Microsoft sues service for creating illicit content with its AI platform

Microsoft and others forbid using their generative AI systems to create various content. Content that is off limits includes materials that feature or promote sexual exploitation or abuse, is erotic or pornographic, or attacks, denigrates, or excludes people based on race, ethnicity, national origin, gender, gender identity, sexual orientation, religion, age, disability status, or similar traits. It also doesn't allow the creation of content containing threats, intimidation, promotion of physical harm, or other abusive behavior.

Besides expressly banning such usage of its platform, Microsoft has also developed guardrails that inspect both prompts inputted by users and the resulting output for signs the content requested violates any of these terms. These code-based restrictions have been repeatedly bypassed in recent years through hacks, some benign and performed by researchers and others by malicious threat actors.

Microsoft didn't outline precisely how the defendants' software was allegedly designed to bypass the guardrails the company had created.

Masada wrote:

> *Microsoft's AI services deploy strong safety measures, including built-in safety mitigations at the AI model, platform, and application levels. As alleged in our court filings unsealed today, Microsoft has observed a foreign-based threat–actor group develop sophisticated software that*

> *exploited exposed customer credentials scraped from public websites. In doing so, they sought to identify and unlawfully access accounts with certain generative AI services and purposely alter the capabilities of those services. Cybercriminals then used these services and resold access to other malicious actors with detailed instructions on how to use these custom tools to generate harmful and illicit content. Upon discovery, Microsoft revoked cybercriminal access, put in place countermeasures, and enhanced its safeguards to further block such malicious activity in the future.*

The lawsuit alleges the defendants' service violated the Computer Fraud and Abuse Act, the Digital Millennium Copyright Act, the Lanham Act, and the Racketeer Influenced and Corrupt Organizations Act and constitutes wire fraud, access device fraud, common law trespass, and tortious interference. The complaint seeks an injunction enjoining the defendants from engaging in "any activity herein."

https://arstechnica.com/security/2025/01/microsoft-sues-service-for-creating-illicit-content-with-its-ai-platform/