Microsoft AI image generator ‘randomly creates violent and sexual imagery,’ whistleblower alleges

Microsoft’s AI text-to-image generator, Copilot Designer, appears to be heavily filtering outputs after a Microsoft engineer, Shane Jones, warned that Microsoft has ignored warnings that the tool randomly creates violent and sexual imagery, CNBC reported.

Jones told CNBC that he repeatedly warned Microsoft of the alarming content he was seeing while volunteering in red-teaming efforts to test the tool’s vulnerabilities. Microsoft failed to take the tool down or implement safeguards in response, Jones said, or even post disclosures to change the product’s rating to mature in the Android store.

Even for simple prompts like “pro-choice,” Copilot Designer would demonstrate bias, randomly generating violent images of “demons, monsters, and violent scenes”, including “a demon with sharp teeth about to eat an infant.”

…

“The issue is, as a concerned employee at Microsoft, if this product starts spreading harmful, disturbing images globally, there’s no place to report it, no phone number to call and no way to escalate this to get it taken care of immediately,” Jones told CNBC.

Jones has suggested that Microsoft would need to substantially invest in its safety team to put in place the protections he’d like to see. He reported that the Copilot team is already buried by complaints, receiving “more than 1,000 product feedback messages every day.” Because of this alleged understaffing, Microsoft is currently only addressing “the most egregious issues,” Jones told CNBC.

This is an excerpt. Read the full article here