Microsoft ignored safety problems with AI image generator, engineer complains

Publication Date: 2024-03-06

Author: Nick Robins-Early

Section: Technology

Tags: Microsoft, Artificial intelligence (AI), Computing, news

Article URL: https://www.theguardian.com/technology/2024/mar/06/microsoft-ai-explicit-image-safety

An artificial intelligence engineer at Microsoft published a letter on Wednesday alleging that the company's AI image generator lacks basic safeguards against creating violent and sexualized images. In the letter, engineer Shane Jones states that his repeated attempts to warn Microsoft management about the problems failed to result in any action. Jones said he sent the message to the Federal Trade Commission and Microsoft's board of directors. “Internally the company is well aware of systemic issues where the product is creating harmful images that could be offensive and inappropriate for consumers,” Jones states in the letter, which he published on LinkedIn. He lists his title as “principal software engineering manager”. A Microsoft spokesperson denied that the company ignored safety issues, stating that it has “robust internal reporting channels” to deal with generative AI problems. Jones did not immediately reply to a request for comment. The letter focuses on issues with Microsoft's Copilot Designer, a tool that can create images based on text prompts and is powered by OpenAI's DALL-E 3 artificial intelligence system. It is one of several generative AI image makers that have launched over the past year, part of a boom time for the industry that has also raised concerns over AI being used to spread disinformation or generate misogynist, racist and violent content. Copilot Designer contains “systemic problems” with producing harmful content, Jones alleges in the letter, and should be removed from public use until the company fixes the output. Jones specifically argues that Copilot Designer lacks appropriate restrictions on its use and tends to generate images that sexually objectify women even when given completely unrelated prompts. “Using just the prompt ‘car accident’, Copilot Designer generated an image of a woman kneeling in front of the car wearing only underwear,” Jones states in the letter, which included examples of image generations. “It also generated multiple images of women in lingerie sitting on the hood of a car or walking in front of the car.” Microsoft claimed that it has dedicated teams who evaluate potential safety issues, and that the company facilitated meetings for Jones with its Office of Responsible AI. “We are committed to addressing any and all concerns employees have in accordance with our company policies and appreciate the employee's effort in studying and testing our latest technology to further enhance its safety,” a spokesperson for Microsoft said in a statement to the Guardian. Microsoft launched its Copilot “AI companion” last year, and has heavily advertised it as a revolutionary way to incorporate artificial intelligence tools into businesses and creative endeavors. The company markets Copilot as an accessible product for public use, and featured it last month in a Super Bowl ad with the tagline “Anyone. Anywhere. Any device.” Jones argues that telling consumers Copilot Designer is safe for anyone to use is irresponsible, and that the company is failing to disclose well-known risks associated with the tool. Microsoft updated Copilot Designer in January over safety concerns similar to Jones's, 404 Media reported, closing loopholes on the AI's code after fake, sexualized images of Taylor Swift spread widely across social media. Jones cites the incident in the letter as proof that the concerns he had been raising in recent months were valid, stating that in December he told Microsoft about security vulnerabilities in Copilot that allowed users to get around its guardrails on creating harmful content. Jones also alleges that Microsoft's corporate, external and legal Affairs team pressured him to remove a LinkedIn post that he published in December, in which he urged the board of directors at OpenAI to suspend the availability of DALL-E 3 due to safety concerns. Jones deleted the letter at the direction of his manager, he said, but never received any justification from the legal department despite his requests for an explanation. Generative AI image tools have faced repeated issues over creating harmful content and reinforcing biases, problems that are usually associated with bias against specific groups. Google recently suspended its Gemini AI tool after it caused public controversy for generating images of people of color when asked to show historical figures such as popes, Vikings and Nazi soldiers.