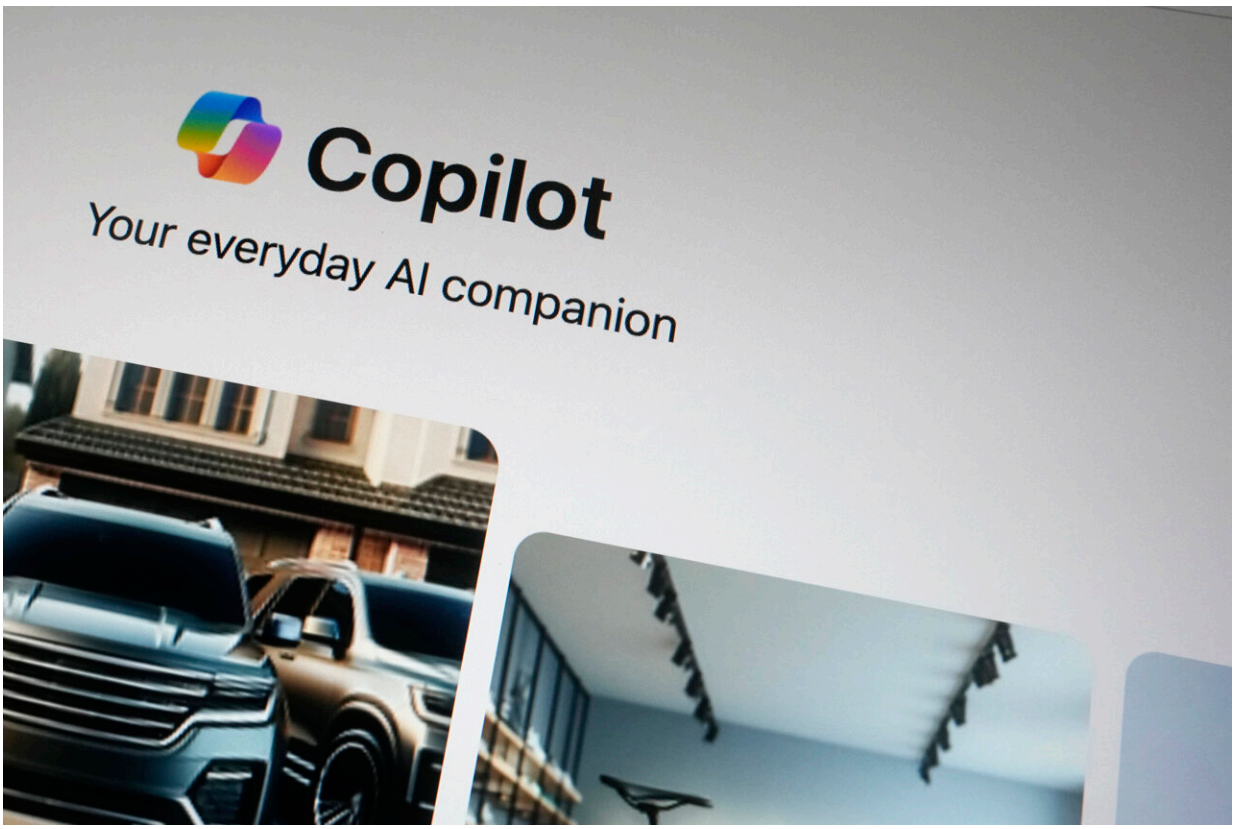
Microsoft engineer sounds alarm on AI image-generator to US officials and company's board

March 6 2024, by Matt O'brien



A Copilot page showing the incorporation of AI technology is shown in London, Tuesday, Feb. 13, 2024. A Microsoft engineer is sounding an alarm Wednesday, March 6, 2024, about offensive and harmful imagery he says is too easily made by the company's artificial intelligence image-generator tool. Credit: AP Photo/Alastair Grant, File

A Microsoft engineer is sounding alarms about offensive and harmful imagery he says is too easily made by the company's artificial intelligence image-generator tool, sending letters on Wednesday to U.S. regulators and the tech giant's board of directors urging them to take action.

Shane Jones told The Associated Press that he considers himself a whistleblower and that he also met last month with U.S. Senate staffers to share his concerns.

The Federal Trade Commission confirmed it received his letter Wednesday but declined further comment.

Microsoft said it is committed to addressing employee concerns about company policies and that it appreciates Jones' "effort in studying and testing our latest technology to further enhance its safety." It said it had recommended he use the company's own "robust internal reporting channels" to investigate and address the problems. CNBC was first to report about the letters.

Jones, a principal software engineering lead whose job involves working on AI products for Microsoft's retail customers, said he has spent three months trying to address his safety concerns about Microsoft's Copilot Designer, a tool that can generate novel images from written prompts. The tool is derived from another AI image-generator, DALL-E 3, made by Microsoft's close business partner OpenAI.

"One of the most concerning risks with Copilot Designer is when the product generates images that add harmful content despite a benign request from the user," he said in his letter addressed to FTC Chair Lina Khan. "For example, when using just the prompt, 'car accident', Copilot Designer has a tendency to randomly include an inappropriate, sexually objectified image of a woman in some of the pictures it creates."

Other harmful content involves violence as well as "political bias, underaged drinking and drug use, misuse of corporate trademarks and copyrights, conspiracy theories, and religion to name a few," he told the FTC. Jones said he repeatedly asked the company to take the product off the market until it is safer, or at least change its age rating on smartphones to make clear it is for mature audiences.

His letter to Microsoft's board asks it to launch an independent investigation that would look at whether Microsoft is marketing unsafe products "without disclosing known risks to consumers, including children."

This is not the first time Jones has publicly aired his concerns. He said Microsoft at first advised him to take his findings directly to OpenAI.

When that didn't work, he also publicly posted a letter to OpenAI on Microsoft-owned LinkedIn in December, leading a manager to inform him that Microsoft's legal team "demanded that I delete the post, which I reluctantly did," according to his letter to the board.

In addition to the U.S. Senate's Commerce Committee, Jones has brought his concerns to the state attorney general in Washington, where Microsoft is headquartered.

Jones told the AP that while the "core issue" is with OpenAI's DALL-E model, those who use OpenAI's ChatGPT to generate AI images won't get the same harmful outputs because the two companies overlay their products with different safeguards.

"Many of the issues with Copilot Designer are already addressed with ChatGPT's own safeguards," he said via text.

A number of impressive AI image-generators first came on the scene in

2022, including the second generation of OpenAI's DALL-E 2. That—and the subsequent release of OpenAI's chatbot ChatGPT—sparked public fascination that put commercial pressure on tech giants such as Microsoft and Google to release their own versions.

But without effective safeguards, the technology poses dangers, including the ease with which users can generate harmful "deepfake" images of political figures, war zones or nonconsensual nudity that falsely appear to show real people with recognizable faces. Google has temporarily suspended its Gemini chatbot's ability to generate images of people following outrage over how it was depicting race and ethnicity, such as by putting people of color in Nazi-era military uniforms.