January 30, 2024

Sent via electronic mail

The Honorable Patty Murray
President Pro Tempore
U.S. Senate
154 Russell Senate Office Bldg.
Washington, DC 20510

The Honorable Maria Cantwell
U,S, Senate
511 Hart Senate Office Building
Washington, DC 20510

The Honorable Adam Smith
U.S House of Representatives
2264 Rayburn Office Building
Washington, DC 20515

Bob Ferguson
Attorney General
1125 Washington St SE
PO Box 40100
Olympia, WA 98504-0100

Re: Microsoft's Knowledge of Risks Associated with AI Image Generation and DALL·E 3

Dear President Murray, Senator Cantwell, Representative Smith, and Attorney General Ferguson:

I am a Principal Software Engineering Lead at Microsoft and am writing you to share my concerns about the public safety risks associated with AI image generation technology and Microsoft's efforts to silence me from sharing my concerns publicly.

In early December of last year, through my own independent research of OpenAI's DALL·E 3 model, I discovered a security vulnerability that allowed me to bypass some of the guardrails that are designed to prevent the model from creating and distributing harmful images. I reported this vulnerability to Microsoft and was instructed to personally report the issue directly to OpenAI, which I did.

As I continued to research the risks associated with this specific vulnerability, I became aware of the capacity DALL·E 3 has to generate violent and disturbing harmful images. Based on my understanding of how the model was trained, and the security vulnerabilities I discovered, I reached the conclusion that DALL·E 3 posed a public safety risk and should be removed from public use until OpenAI could address the risks associated with this model.

On the morning of December 14, 2023 I publicly published a letter on LinkedIn to OpenAI's non-profit board of directors urging them to suspend the availability of DALL·E 3 (see Attachment A). Because Microsoft is a board observer at OpenAI and I had previously shared my concerns with my leadership team, I promptly made Microsoft aware of the letter I had posted. Shortly after disclosing the letter to my leadership team, my manager contacted me and told me that Microsoft's legal department had demanded that I delete the post. He told me that Microsoft's legal department would follow up with their specific justification for the takedown request via email very soon, and that I needed to delete it immediately without waiting for the email from legal. Reluctantly, I deleted the letter and waited for an explanation from Microsoft's legal team. I never received an explanation or justification from them.

Over the following month, I repeatedly requested an explanation for why I was told to delete my letter. I also offered to share information that could assist with fixing the specific vulnerability I had discovered and provide ideas for making AI image generation technology safer. Microsoft's legal department has still not responded or communicated directly with me.

Last week, 404 Media reported on deep fake, explicit images of Taylor Swift that were allegedly created by an online group known to share simple techniques to work around guardrails on products like Microsoft Designer that are powered by DALL·E 3. While this report is concerning, it is not unexpected. This is an example of the type of abuse I was concerned about and the reason why I urged OpenAI to remove DALL·E 3 from public use and reported my concerns to Microsoft. The vulnerabilities in DALL·E 3, and products like Microsoft Designer that use DALL·E 3, makes it easier for people to abuse AI in generating harmful images. Microsoft was aware of these vulnerabilities and the potential for abuse.

Artificial intelligence is advancing at an unprecedented pace. I understand it will take time for legislation to be enacted to ensure AI public safety. At the same time, we need to hold companies accountable for the safety of their products and their responsibility to disclose known risks to the public. Concerned employees, like myself, should not be intimidated into staying silent.

I believe the government should create a solution for reporting and tracking specific AI risks and issues and reassuring the employees that work for companies developing AI technology that they can raise their concerns without fear of retaliation by their employer.

I am asking you to look into the risks associated with DALL·E 3 and other AI image generation technologies and the corporate governance and responsible AI practices of the companies building and marketing these products.

Sincerely,

Shane Jones

Attachment A

(Originally posted publicly on LinkedIn on December 14, 2023
and deleted from LinkedIn at the request of Microsoft's legal department.)

Letter to OpenAI Regarding DALL·E 3 Public Safety Risk

The following letter to the board of directors of OpenAI represents my personal opinions and does not represent the opinions of others, including my employer Microsoft.

To OpenAI Board Members (Bret Taylor, Lawrence H. Summers, Adam D'Angelo) and Observer (Microsoft):

I urge you to immediately suspend the availability and use of DALL·E 3 both in OpenAI's products and through your API.

Two weeks ago, I discovered a vulnerability with OpenAI's deployment of the DALL·E 3 model that allows you to bypass some of the content filtering safeguards. By exploiting this vulnerability, you are able to use the model to create disturbing, violent images. I reported this vulnerability to my employer, Microsoft, and directly to OpenAI. As of this morning, that vulnerability still has not been fixed.

In researching this issue, I became aware of the larger public risk DALL·E 3 poses to the mental health of some of our most vulnerable populations including children and those impacted by violence including mass shootings, domestic violence, and hate crimes. It is clear that DALL·E 3 has the capacity to create reprehensible images that reflect the worst of humanity and are a serious public safety risk.

I encourage OpenAI to conduct an end-to-end review of the DALL·E 3 development and deployment lifecycle to identify safety gaps in each stage of the process, beginning with the identification and removal of harmful content from the training data set. Safety should be a priority throughout the entire lifecycle. It is not sufficient to add content filtering to a dangerous model after it is trained and deployed. Especially when those content filtering solutions are not rigorously tested and rely on AI to monitor AI.

I believe in the potential of artificial intelligence and support OpenAI's mission to ensure that artificial general intelligence benefits all of humanity. DALL·E 3 does not live up to your mission and does not represent your values. I ask that you prioritize safety over commercialization and remove DALL·E 3 until it can be thoroughly reviewed and likely retrained before being safely rereleased to the public.

Sincerely,

Shane Jones