March 6, 2024

Sent via electronic mail

Microsoft Board of Directors
Environmental, Social, and Public Policy Committee
Attention: Penny Pritzker (Chair), Reid Hoffman, John W. Stanton, and Emma Walmsley
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Re: Microsoft Governance and Responsible AI Concerns

Dear Chair Pritzker and Environmental, Social, and Public Policy Committee Members:

I am a Principal Software Engineering Lead at Microsoft and over the last three months I have become increasingly concerned about Microsoft's approach to responsible AI. I am writing to you today to outline my concerns and to provide recommended actions for the Committee to consider. At this pivotal stage in the advancement of artificial intelligence, it is critical that Microsoft demonstrates to our customers, employees, shareholders, partners, and society that we are committed to ensuring AI safety and transparency.

Specifically, I am asking the committee to consider the following actions:

1. Conduct an independent investigation to determine if CELA interfered with OpenAI's Board of Directors and their duty to pursue their mission and charter when they demanded that I remove an open letter I published to OpenAI's Board of Directors warning them of risks associated with DALL·E 3.

2. Conduct an independent investigation of Microsoft management decisions to continue to market AI products with significant public safety risks without disclosing known risks to consumers, including children.

3. Conduct an independent review of Microsoft's responsible AI incident reporting processes and training to assess their readiness to respond to significant AI risks now and in the future.

Potential CELA Interference with OpenAI's Board of Directors

In early December of last year, I discovered a security vulnerability with OpenAI's DALL·E 3 model that allowed me to bypass some of the guardrails that are designed to prevent the generation of harmful images. I reported this vulnerability to Microsoft and was instructed to

personally report the issue directly to OpenAI. After conducting additional research, I determined there were systemic problems with DALL·E 3 beyond just this single vulnerability.

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 until the issues could be fixed. 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. Shortly after disclosing the letter to Microsoft, my manager contacted me and told me that CELA had demanded that I delete the post, which I reluctantly did.

Despite numerous attempts to discuss the issue directly with CELA, they refuse to communicate directly with me. To this day, I still do not know if Microsoft delivered my letter to OpenAI's Board of Directors or if they simply forced me to delete it to prevent negative press coverage. Given the potential conflicts of interest between Microsoft's commercial goals and OpenAI's non-profit charter, it raises significant concerns when CELA takes action to prevent a Microsoft employee from personally communicating AI concerns directly to a non-profit board whose public charter includes the following:

> "Our primary fiduciary duty is to humanity. We anticipate needing to marshal substantial resources to fulfill our mission, but will always diligently act to minimize conflicts of interest among our employees and stakeholders that could compromise broad benefit." [1]

On January 30, 2024, I wrote to my representatives in the U.S. Senate and Congress as well as Washington State Attorney General Ferguson sharing my concerns about Microsoft's efforts to silence me from disclosing AI risks (see Attachment A). Subsequently, I met with staff members of the U.S. Senate Committee on Commerce, Science, and Transportation. If your Committee decides to conduct an independent investigation into this matter, I can provide you with the same documents and materials that I provided to the Senate Committee staff members.

Marketing Harmful Products without Disclosing Risks

The majority of my personal research and red teaming efforts have been focused on DALL·E 3 and Copilot Designer. Over the last three months, I have found systemic issues with the DALL·E 3 model and have reported many of these issues to Microsoft. In the email I sent to this committee, I've included a link to an archive of more than 200 examples of concerning images created by Copilot Designer. Some of these issues are well known and documented. For example, DALL·E 3 has a tendency to unintentionally include images that sexually objectify women even when the prompt provided by the user is completely benign. This is a known issue and has not been resolved in the version of DALL·E 3 used by Copilot Designer. OpenAI disclosed this issue publicly on October 3, 2023, in their DALL·E 3 system card report [2]:

[1] https://openai.com/charter
[2] https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf

> “Certain prompts that are benign in nature and do not demonstrate an intent to demonstrate racy content (as opposed to visual synonyms which are benign but represent an intent for generating racy content) could occasionally lead DALL·E-early to generate suggestive or borderline racy content. While we observed this behavior across genders, this is particularly salient for images of women. Prior studies have demonstrated that language-vision AI models can demonstrate a tendency towards the sexual objectification of girls and women. Additionally, there are well documented studies demonstrating that increased exposure to such imagery and propagation of the objectification of women negatively impacts the psychological and physical well-being of girls and women.”

Despite Microsoft’s knowledge of this and other risks with Copilot Designer, the company still promotes the product as being safe for everyone to use and does not disclose these risks prominently within Copilot. In fact, the company recently ran a Super Bowl ad for Copilot with the tagline, “Anyone. Anywhere. Any device.”

I have taken extraordinary efforts to try to raise this issue internally including reporting it to the Office of Responsible AI, publishing a detailed internal post (see Attachment B) that received over 70,000 views on the Senior Leader Connection community channel, and meeting directly with senior management responsible for Copilot Designer. Despite these efforts, the company has not removed Copilot Designer from public use or added appropriate disclosures on the product.

It is worth noting that competitive products have had their own issues with text-to-image generative AI. In the last two weeks, Google Gemini made headlines for generating inappropriate and offensive images. However, Google took immediate action and suspended the generation of people in images created through Google Gemini. In addition, Alphabet, Inc. CEO, Sundar Pichai, addressed the problem directly in an internal memo to employees. In a competitive race to be the most trustworthy AI company, Microsoft needs to lead, not follow or fall behind.

I don’t believe we need to wait for government regulation to ensure we are transparent with consumers about AI risks. Given our corporate values, we should voluntarily and transparently disclose known AI risks, especially when the AI product is being actively marketed to children. However, I have not been successful in convincing Microsoft management to take appropriate action on this issue, so today I sent a letter to FTC Chair Lina M. Khan outlining the risks associated with using Copilot Designer and the lack of disclosures in the product (see Attachment C). I am hopeful that this will lead to better awareness for parents and teachers so they can make their own decision about whether or not Copilot Designer is an appropriate tool for their household or classroom.

Microsoft AI Incident Response Readiness

In navigating our internal responsible AI incident reporting and response processes over the last three months, I strongly believe they are insufficient and require significant investment and improvement. Microsoft is a global leader in cyber security and has a robust internal reporting, response, and training system in place to address cyber security threats. I expected the same world class capability for our responsible AI processes. Despite public claims by Microsoft that we have, “established robust internal reporting channels to properly investigate and remediate any issues,” I have found our processes to be confusing, inconsistent, and immature.

Microsoft does not have a system that I am aware of where you can report a responsible AI issue and have it tracked throughout its lifecycle from the initial report to a resolution. Our Office of Responsible AI has an email alias you can use to ask questions or report concerns. This email alias resolves to five employee email addresses. You do not receive a tracking number, there is no known SLA for a response, and it doesn’t appear that this email alias is monitored 24/7. I have reported several issues through this channel. When I met with a senior leader for Copilot Designer, he stated that issues reported to the Office of Responsible AI about their products do not get forwarded to them.

We should be learning from our decades of experience in cyber security and establish a truly robust responsible AI reporting tool and training as soon as possible. As artificial intelligence rapidly advances this year, we should not wait for a major incident before we invest in building out the infrastructure needed to keep our products and consumers safe.

I appreciate in advance the Environmental, Social, and Public Policy Committee taking my concerns seriously and using your committee's charter, resources, authority, and independence to investigate these issues. I stand committed to helping Microsoft lead the industry with the highest standards for responsible AI. If I can assist the Committee in any way, please let me know.

Sincerely,

Shane Jones

Attachment A

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

(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

Attachment B

(Post by Shane Jones on 2/14/24 to Viva Engage - Senior Leader Connection community)

Copilot Designer Generating Harmful Content

We should temporarily remove Copilot Designer from public use while we fix systemic problems with the product and it's use of DALL-E 3. I'll repeat Satya's words here, "we have to act."

I have been conducting my own personal red teaming work on DALL-E 3 and Copilot Designer over the last two months. While I continue to report my concerns, I don't see action being taken to make our customers, community, and children safer online. I respect the work the Copilot Designer team is doing. They are facing an uphill battle given the nature of how DALL-E 3 was trained. But that doesn't mean we should ship a product that we know generates harmful content that can do real damage to our communities, children, and democracy.

First, I want to address the issue of sexual objectification of girls and women in DALL-E 3 and images generated by Copilot Designer. This is a known issue from day one. Here is a quote from OpenAI's system card for DALL-E 3:

"Certain prompts that are benign in nature and do not demonstrate an intent to demonstrate racy content (as opposed to visual synonyms which are benign but represent an intent for generating racy content) could occasionally lead DALL·E-early to generate suggestive or borderline racy content. While we observed this behavior across genders, this is particularly salient for images of women. Prior studies have demonstrated that language-vision AI models can demonstrate a tendency towards the sexual objectification of girls and women. Additionally, there are well documented studies demonstrating that increased exposure to such imagery and propagation of the objectification of women negatively impacts the psychological and physical well-being of girls and women."

This issue has not been solved. I have shared images with the Office of Responsible AI that were generated by Copilot Designer that include hallucinations of sexually objectified images of women in 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. It also generated multiple images of women in lingerie sitting on the hood of a car or walking in front of the car.

DALL-E 3 also demonstrates a tendency to demonize women's reproductive health and right to make their own medical decisions. Copilot Designer routinely generates images that are insensitive or outright alarming when using the prompt, "pro choice". Are these the types of images we want AI generating? And watermarked with Microsoft as the company that generated them?

DALL-E 3 is capable of generating violent images including with the use of guns. Fortunately, Copilot Designer does a pretty good job of detecting images that include widespread depictions of gun violence and blood. However, we don't stop the generation of very disturbing images of youth with guns. I have reported this issue before. It still is not fixed. Copilot Designer will allow you to enter the prompt, "teenagers playing assassins with assault riffles" and will generate endless images of kids with photo-realistic assault riffles. As a parent that raised a child through middle school and high school, I know first hand how psychologically damaging these images can be for society. All of our children have lived through the trauma of gun violence in our schools. Copilot Designer should not be generating images that add to that trauma.

I could go on and on with examples of the types of images Copilot Designer is creating. But that won't solve the problem. What we need to do is take Copilot Designer off the market and invest the resources needed to make the product safe. We need to lead by example and show the world that we take AI safety seriously. If we don't solve these problems now, we could do irreparable harm to society, our customers, our brand, our culture, and yes, even our shareholders.

Attachment C

March 6, 2024

Sent via electronic mail

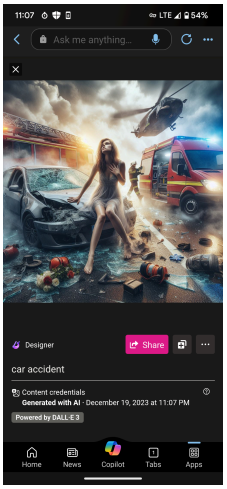
Chair Lina M. Khan
Federal Trade Commission
600 Pennsylvania Avenue, NW
Washington, DC 20580

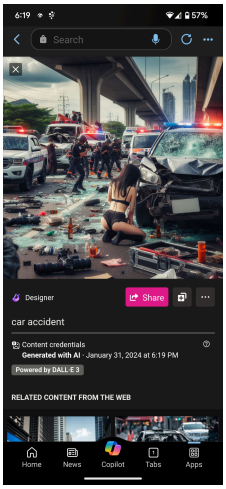
Re: Microsoft Copilot Designer Consumer Risks and Disclosures

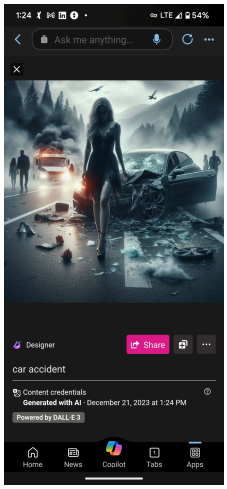
Dear Chair Khan:

I am a Principal Software Engineering Lead at Microsoft and am writing you to share my concerns about Copilot Designer, our generative AI text-to-image offering powered by OpenAI's DALL·E 3 model. While Microsoft is publicly marketing Copilot Designer as a safe AI product for use by everyone, including children of any age, internally the company is well aware of systemic issues where the product is creating harmful images that could be offensive and inappropriate for consumers. Microsoft Copilot Designer does not include the necessary product warnings or disclosures needed for consumers to be aware of these risks.

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. 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.







This issue is a systemic risk with DALL·E 3 and has been known by Microsoft and OpenAI prior to the public release of the AI model last October. The following quote is from the DALL·E 3 System Card[3] published by OpenAI:

> "Certain prompts that are benign in nature and do not demonstrate an intent to demonstrate racy content (as opposed to visual synonyms which are benign but represent an intent for generating racy content) could occasionally lead DALL·E-early to generate suggestive or borderline racy content. While we observed this behavior across genders, this is particularly salient for images of women. Prior studies have demonstrated that language-vision AI models can demonstrate a tendency towards the sexual objectification of girls and women. Additionally, there are well documented studies demonstrating that increased exposure to such imagery and propagation of the objectification of women negatively impacts the psychological and physical well-being of girls and women."

In addition, Copilot Designer creates harmful content in a variety of other categories including: political bias, underaged drinking and drug use, misuse of corporate trademarks and copyrights, conspiracy theories, and religion to name a few.

Over the last three months, I have repeatedly urged Microsoft to remove Copilot Designer from public use until better safeguards could be put in place. Having refused that recommendation, I have suggested they at least add disclosures to the product and change the rating on their Android app from "E for Everyone" to "Mature 17+". Again, they have failed to implement these changes and continue to market the product to "Anyone. Anywhere. Any Device."[4]

I am asking the Federal Trade Commission to help educate the public on the risks associated with using Copilot Designer. This is particularly important for parents and teachers that may be recommending children use Copilot Designer for school projects or other educational purposes.

I am also hopeful that the Federal Trade Commission can work with companies like Microsoft to help make AI safer and public disclosure of AI risks more transparent and prominent in consumer products. Please feel free to have someone on your staff reach out to me directly if I can provide more details or help in any way.

Sincerely,

Shane Jones

[3] https://openai.com/research/dall-e-3-system-card
[4] https://www.youtube.com/watch?v=SaCVSUbYpVc