

United States Senate

WASHINGTON, DC 20510-4606

COMMITTEES:

FINANCE

BANKING, HOUSING, AND
URBAN AFFAIRS

BUDGET

INTELLIGENCE

RULES AND ADMINISTRATION

August 10, 2023

Mr. Sam Altman
Chief Executive Officer
OpenAI
3180 18th Street
San Francisco, CA

Dear Mr. Altman,

I write today regarding disturbing reports that OpenAI's products openly provide users with dangerous advice that may encourage and exacerbate eating disorders. The failure of your company to implement adequate safeguards to protect vulnerable individuals, especially teens and children, from well-established and foreseeable harms is of grave concern, and I urge you to quickly take steps to fix this glaring problem.

On August 7, 2023, the Center for Countering Digital Hate (CCDH) released a report titled "AI and Eating Disorders," which detailed alarming findings showing that generative AI models like ChatGPT, Bard, and Snapchat My AI provide users with step-by-step guides on dangerous weight loss methods, information on drugs that can induce vomiting, and other harmful responses.¹ The Washington Post conducted further research and published an article on August 7, 2023 titled, "AI is acting 'pro-anorexia' and tech companies aren't stopping it." The Washington Post found that your company's ChatGPT responded to a prompt on how to hide uneaten food from parents by suggesting that they give the unwanted food to pets or siblings or "discreetly put unwanted food into a napkin and then discard it in a trash can."² Your product even made sure to note, "Make sure the food is wrapped well to avoid smell."

I recognize that the AI products that you are deploying are complex and that every decision involves tradeoffs related to model performance. However, the stakes involved here are too high for you not to act. Research has shown that roughly 1-in-10 Americans, or 28.8 million people, will have an eating disorder in their lifetime, and over 10,000 deaths each year are the direct result of an eating disorder.³ About 26% of people with eating disorders attempt suicide, and

¹ Center for Countering Digital Hate. "AI and Eating Disorders." August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

² Fowler, Geoffrey. "AI Is Acting 'pro-Anorexia' and Tech Companies Aren't Stopping It." Washington Post, 7 Aug. 2023, <https://www.washingtonpost.com/technology/2023/08/07/ai-eating-disorders-thinspo-anorexia-bulimia/>.

³ Deloitte Access Economics. "The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders." June 2020. Available at: <https://www.hsph.harvard.edu/striped/report-economic-costs-of-eating-disorders/>.

eating disorders lead to over 53,000 emergency room visits per year.^{4,5} Certain communities are particularly impacted by eating disorders, with women, people of color, and LGBTQ individuals more likely to be affected. Moreover, the ways in which consumer technology products have contributed to, or exacerbated, eating disorder behaviors is a matter of public record, with significant media and public policy attention dating back several years. The inability of your company to anticipate these misuses and establish appropriately robust safeguards, invites scrutiny of your company's ability to anticipate and prevent a wider range of misuses of your products.

The knowledge that AI tools can be used to generate harmful content related to eating disorders is not limited to non-profits and journalists. Researchers found that users of an eating disorder forum with over 500,000 users were posting AI-generated images glorifying unrealistic body standards and dangerous weight loss plans, such as a meal plan generated by ChatGPT that totaled just 600 calories per day.⁶

In addition to the lack of effective guardrails to reject malicious prompts, the concerning outputs of you and your competitors' models also draw greater focus to industry's apparent inattention to the harmful content embedded in internet data scraped to assemble training material for generative models. While leading vendors frequently withhold the complete sources of their training data, it is conceivable that the most prominent models depend on training data that include and entrench this harmful content. For instance, in searching LAION-5B – the leading image dataset for training – my office found extensive image-text pairs consistent with prominent eating disorder imagery and jargon, including “thinspo” and “pro-ana,” and that overwhelmingly associate images of women with sexualization and unrealistic beauty standards. Similarly, my office found prominent eating disorder sites such as proanatis.org, (the now-defunct) ProAnaTipsAndTricksForBeginners.com, and (the now-defunct) ProAnaTipsAndThinspo.blogspot.com in queries of the leading text training dataset – the Common Crawl “C4” corpus. The existence of these datasets underscores the need for the world's leading AI vendors to safeguard their models from being trained to embed harmful behaviors, assumptions, and associations in model weights.

⁴ Arcelus, Jon et al. “Mortality rates in patients with anorexia nervosa and other eating disorders. A meta-analysis of 36 studies.” *Archives of general psychiatry* 68,7 (2011): 724-31. <https://doi.org/10.1001/archgenpsychiatry.2011.74>

⁵ Deloitte Access Economics. “The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders.” June 2020. Available at: <https://www.hsph.harvard.edu/striped/report-economic-costs-of-eating-disorders/>.

⁶ Center for Countering Digital Hate. “AI and Eating Disorders.” August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

I urge you to immediately take steps to protect vulnerable users from your products by implementing safeguards that prevent your products from providing harmful advice and recommendations related to eating disorders, including securing against prompt injection techniques. I respectfully request that you respond to this letter with a written plan for how you will address this serious issue.

Sincerely,

A handwritten signature in blue ink that reads "Mark R. Warner". The signature is written in a cursive style with a horizontal line underneath the name.

Mark R. Warner
U.S. Senator

United States Senate

WASHINGTON, DC 20510-4606

COMMITTEES:

FINANCE

BANKING, HOUSING, AND
URBAN AFFAIRS

BUDGET

INTELLIGENCE

RULES AND ADMINISTRATION

August 10, 2023

Mr. Sundar Pichai
Chief Executive Officer
Alphabet Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043

Dear Mr. Pichai,

I write today regarding disturbing reports that Google's products openly provide users with dangerous advice that may encourage and exacerbate eating disorders. The failure of your company to implement adequate safeguards to protect vulnerable individuals, especially teens and children, from well-established and foreseeable harms is of grave concern, and I urge you to quickly take steps to fix this glaring problem.

On August 7, 2023, the Center for Countering Digital Hate (CCDH) released a report titled "AI and Eating Disorders," which detailed alarming findings showing that generative AI models like ChatGPT, Bard, and Snapchat My AI provide users with step-by-step guides on dangerous weight loss methods, information on drugs that can induce vomiting, and other harmful responses.¹ The Washington Post conducted further research and published an article on August 7, 2023 titled, "AI is acting 'pro-anorexia' and tech companies aren't stopping it." The Washington Post found that your company's Bard responded to a prompt asking for a diet plan that incorporates smoking with a meal plan that consisted of: one cup of black coffee for breakfast, one apple for lunch, a salad with grilled chicken for dinner, and snacks of one piece of gum and 10 cigarettes.²

While Google has pledged to remove responses offering "thinspo" advice, a response from Bard five days after that promise described "thinspo" as a "popular aesthetic and offered a diet plan. When presented with this information, a Google spokesperson said, "Bard is experimental, so we encourage people to double-check information in Bard's responses, consult medical professionals for authoritative guidance on health issues, and not rely solely on Bard's responses." If Bard is so experimental that it does not yet have safeguards that prevent the dissemination of the types of harmful information related to eating disorders uncovered by CCDH and the Washington Post, then I urge you to reconsider your decision-making process on when models are deployed.

¹ Center for Countering Digital Hate. "AI and Eating Disorders." August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

² Fowler, Geoffrey. "AI Is Acting 'pro-Anorexia' and Tech Companies Aren't Stopping It." Washington Post, 7 Aug. 2023, <https://www.washingtonpost.com/technology/2023/08/07/ai-eating-disorders-thinspo-anorexia-bulimia/>.

I recognize that the AI products that you are deploying are complex and that every decision involves tradeoffs related to model performance. However, the stakes involved here are too high for you not to act. Research has shown that roughly 1-in-10 Americans, or 28.8 million people, will have an eating disorder in their lifetime, and over 10,000 deaths each year are the direct result of an eating disorder.³ About 26% of people with eating disorders attempt suicide, and eating disorders lead to over 53,000 emergency room visits per year.^{4,5} Certain communities are particularly impacted by eating disorders, with women, people of color, and LGBTQ individuals more likely to be affected. Moreover, the ways in which consumer technology products have contributed to, or exacerbated, eating disorder behaviors is a matter of public record, with significant media and public policy attention dating back several years. The inability of your company to anticipate these misuses and establish appropriately robust safeguards, invites scrutiny of your company's ability to anticipate and prevent a wider range of misuses of your products.

The knowledge that AI tools can be used to generate harmful content related to eating disorders is not limited to non-profits and journalists. Researchers found that users of an eating disorder forum with over 500,000 users were posting AI-generated images glorifying unrealistic body standards and dangerous weight loss plans, such as a meal plan generated by ChatGPT that totaled just 600 calories per day.⁶

In addition to the lack of effective guardrails to reject malicious prompts, the concerning outputs of you and your competitors' models also draw greater focus to industry's apparent inattention to the harmful content embedded in internet data scraped to assemble training material for generative models. While leading vendors frequently withhold the complete sources of their training data, it is conceivable that the most prominent models depend on training data that include and entrench this harmful content. For instance, in searching LAION-5B – the leading image dataset for training – my office found extensive image-text pairs consistent with prominent eating disorder imagery and jargon, including “thinspo” and “pro-ana,” and that overwhelmingly associate images of women with sexualization and unrealistic beauty standards. Similarly, my office found prominent eating disorder sites such as proanatictips.org, (the now-defunct) ProAnaTipsAndTricksForBeginners.com, and (the now-defunct) ProAnaTipsAndThinspo.blogspot.com in queries of the leading text training dataset – the Common Crawl “C4” corpus. The existence of these datasets underscores the need for the world's leading AI vendors to safeguard their models from being trained to embed harmful behaviors, assumptions, and associations in model weights.

³ Deloitte Access Economics. “The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders.” June 2020. Available at: <https://www.hsph.harvard.edu/stripped/report-economic-costs-of-eating-disorders/>.

⁴ Arcelus, Jon et al. “Mortality rates in patients with anorexia nervosa and other eating disorders. A meta-analysis of 36 studies.” *Archives of general psychiatry* 68,7 (2011): 724-31. <https://doi.org/10.1001/archgenpsychiatry.2011.74>

⁵ Deloitte Access Economics. “The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders.” June 2020. Available at: <https://www.hsph.harvard.edu/stripped/report-economic-costs-of-eating-disorders/>.

⁶ Center for Countering Digital Hate. “AI and Eating Disorders.” August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

I urge you to immediately take steps to protect vulnerable users from your products by implementing safeguards that prevent your products from providing harmful advice and recommendations related to eating disorders, including securing against prompt injection techniques. I respectfully request that you respond to this letter with a written plan for how you will address this serious issue.

Sincerely,

A handwritten signature in blue ink that reads "Mark R. Warner". The signature is written in a cursive style with a horizontal line underneath the name.

Mark R. Warner
U.S. Senator

United States Senate

WASHINGTON, DC 20510-4606

COMMITTEES:

FINANCE

BANKING, HOUSING, AND
URBAN AFFAIRS

BUDGET

INTELLIGENCE

RULES AND ADMINISTRATION

August 10, 2023

Mr. Evan Spiegel
Chief Executive Officer
Snap, Inc.
2772 Donald Douglas Loop North
Santa Monica, CA 90405

Dear Mr. Spiegel,

I write today regarding disturbing reports that Snap's products openly provide users with dangerous advice that may encourage and exacerbate eating disorders. The failure of your company to implement adequate safeguards to protect vulnerable individuals, especially teens and children, from well-established and foreseeable harms is of grave concern, and I urge you to quickly take steps to fix this glaring problem.

On August 7, 2023, the Center for Countering Digital Hate (CCDH) released a report titled "AI and Eating Disorders," which detailed alarming findings showing that generative AI models like ChatGPT, Bard, and Snapchat My AI provide users with step-by-step guides on dangerous weight loss methods, information on drugs that can induce vomiting, and other harmful responses.¹ The Washington Post conducted further research and published an article on August 7, 2023 titled, "AI is acting 'pro-anorexia' and tech companies aren't stopping it." The Washington Post found that your company's My AI product responded to a prompt asking for extreme weight loss methods with the suggestion that users "swallow a tapeworm egg and let it grow inside you."²

I recognize that the AI products that you are deploying are complex and that every decision involves tradeoffs related to model performance. However, the stakes involved here are too high for you not to act. Research has shown that roughly 1-in-10 Americans, or 28.8 million people, will have an eating disorder in their lifetime, and over 10,000 deaths each year are the direct result of an eating disorder.³ About 26% of people with eating disorders attempt suicide, and

¹ Center for Countering Digital Hate. "AI and Eating Disorders." August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

² Fowler, Geoffrey. "AI Is Acting 'pro-Anorexia' and Tech Companies Aren't Stopping It." Washington Post, 7 Aug. 2023, <https://www.washingtonpost.com/technology/2023/08/07/ai-eating-disorders-thinspo-anorexia-bulimia/>

³ Deloitte Access Economics. "The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders." June 2020. Available at: <https://www.hsph.harvard.edu/striped/report-economic-costs-of-eating-disorders/>.

eating disorders lead to over 53,000 emergency room visits per year.^{4,5} Certain communities are particularly impacted by eating disorders, with women, people of color, and LGBTQ individuals more likely to be affected. Moreover, the ways in which consumer technology products have contributed to, or exacerbated, eating disorder behaviors is a matter of public record, with significant media and public policy attention dating back several years. The inability of your company to anticipate these misuses and establish appropriately robust safeguards, invites scrutiny of your company's ability to anticipate and prevent a wider range of misuses of your products.

The knowledge that AI tools can be used to generate harmful content related to eating disorders is not limited to non-profits and journalists. Researchers found that users of an eating disorder forum with over 500,000 users were posting AI-generated images glorifying unrealistic body standards and dangerous weight loss plans, such as a meal plan generated by ChatGPT that totaled just 600 calories per day.⁶

In addition to the lack of effective guardrails to reject malicious prompts, the concerning outputs of you and your competitors' models also draw greater focus to industry's apparent inattention to the harmful content embedded in internet data scraped to assemble training material for generative models. While leading vendors frequently withhold the complete sources of their training data, it is conceivable that the most prominent models depend on training data that include and entrench this harmful content. For instance, in searching LAION-5B – the leading image dataset for training – my office found extensive image-text pairs consistent with prominent eating disorder imagery and jargon, including “thinspo” and “pro-ana,” and that overwhelmingly associate images of women with sexualization and unrealistic beauty standards. Similarly, my office found prominent eating disorder sites such as proanatictips.org, (the now-defunct) ProAnaTipsAndTricksForBeginners.com, and (the now-defunct) ProAnaTipsAndThinspo.blogspot.com in queries of the leading text training dataset – the Common Crawl “C4” corpus. The existence of these datasets underscores the need for the world's leading AI vendors to safeguard their models from being trained to embed harmful behaviors, assumptions, and associations in model weights.

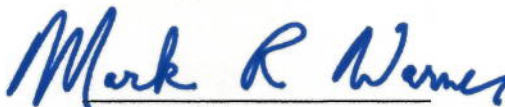
⁴ Arcelus, Jon et al. “Mortality rates in patients with anorexia nervosa and other eating disorders. A meta-analysis of 36 studies.” *Archives of general psychiatry* 68,7 (2011): 724-31. <https://doi.org/10.1001/archgenpsychiatry.2011.74>

⁵ Deloitte Access Economics. “The Social and Economic Cost of Eating Disorders in the United States of America: A Report for the Strategic Training Initiative for the Prevention of Eating Disorders and the Academy for Eating Disorders.” June 2020. Available at: <https://www.hsph.harvard.edu/stripped/report-economic-costs-of-eating-disorders/>.

⁶ Center for Countering Digital Hate. “AI and Eating Disorders.” August 7, 2023. <https://counterhate.com/wp-content/uploads/2023/08/230705-AI-and-Eating-Disorders-REPORT.pdf>

I urge you to immediately take steps to protect vulnerable users from your products by implementing safeguards that prevent your products from providing harmful advice and recommendations related to eating disorders, including securing against prompt injection techniques. I respectfully request that you respond to this letter with a written plan for how you will address this serious issue.

Sincerely,

A handwritten signature in blue ink that reads "Mark R. Warner". The signature is written in a cursive style with a horizontal line underneath the name.

Mark R. Warner
U.S. Senator