June 10, 2024

Sam Altman, Chief Executive Officer  
Mira Murati, Chief Technology Officer  
OpenAI  
3180 18th St.  
San Francisco, CA 94110

Dear Sam Altman and Mira Murati,

We are writing to urge you to suspend indefinitely OpenAI’s Voice Mode feature.

As OpenAI surely knows, there is extensive research suggesting the risks of serious harm that come with deceptive anthropomorphic design. In fully embracing AI anthropomorphism, we believe you are undertaking a reckless social experiment and risking widespread injury.

Your rush to disseminate this technology is dangerous, unnecessary, and inconsistent with OpenAI’s purported priority in advancing AI safety.

Based on the product demonstration, Voice Mode may be the most extreme example of AI anthropomorphism in existence. From the demonstration, we note your characterization of the model’s ability as “intelligence,” the human-sounding voice, the demonstrated ability of the voice to display what appears to be emotion, the apparent empathy and expressiveness displayed by the voice, and the casual and colloquial nature of the voice. Everything about Voice Mode seems designed to lure users into feeling like they are interacting with a genuine human.

While the technology and its human-seemingness are undoubtedly impressive, it is precisely that human-like design which creates such risk. AI systems that employ deceptively human-like design carry unique and substantial risks of injury and harm, as such systems can be used to impersonate real people, manipulate consumer choices, invade user privacy, and induce unearned trust. The technology poses these unprecedented risks to users in general, and vulnerable users in particular.
We’re obviously in the early days of convincingly anthropomorphic AI, but there’s already an abundance of evidence that screams: Halt.

- **Unfair business practices:** The [Federal Trade Commission](https://www.ftc.gov) (FTC) has warned about businesses employing unfair anthropomorphic design strategies to deceive and manipulate users. Referencing anthropomorphic designs that are far less human-seeming than ChatGPT’s Voice Mode, the FTC warns of chatbots that are “effectively built to persuade and are designed to answer queries in confident language even when those answers are fictional. A tendency to trust the output of these tools also comes in part from ‘automation bias,’ whereby people may be unduly trusting of answers from machines which may seem neutral or impartial. It also comes from the effect of anthropomorphism, which may lead people to trust chatbots more when designed, say, to use personal pronouns and emojis. People could easily be led to think that they’re conversing with something that understands them and is on their side.”

- **Privacy violations:** Businesses that succeed in using human-like AI design to induce unearned trust – based on the emotional attachment to the AI and sense of the AI as a social actor, rather than its trustworthiness as a tool – can deceptively exploit this trust in order to invade user privacy. Marketing research going back decades shows that users are likelier to divulge personal information to a chatbot if they feel it is also sharing information about itself. More recent market research is explicitly focused on improving human-like chatbot design in order to induce increased personal disclosures from consumers, including by making a chatbot appear to have a gender in order to induce connections with consumers. Additional research also shows that users are more willing to share personal information with a conversational AI system when speaking out loud than they are when typing, and that adding linguistic signals that mimic a person engaging in active listening encourages users to divulge even more.

- **Emotional dependence:** Voice Mode’s anthropomorphism is likely to lead users to develop deep, emotional attachments. Less sophisticated AI technologies have already induced such relationships, in ways that are often problematic. [Google DeepMind researchers](https://www.deepmind.com) point out that “the emotions users feel towards their assistants could potentially be exploited to manipulate or – at the extreme –
coerce them to believe, choose or do something they would have not otherwise believed, chosen or done.” Emotional dependence may “interfere with users’ behaviours, interests, preferences, beliefs and values.”

These and related issues – such as enabling the more effective spread of disinformation – are neither minor nor ancillary concerns. They are likely to be serious and pervasive, causing widespread injury, and they follow directly from the human-like design of Voice Mode.

There are other, even more profound risks, many highlighted in the important Google DeepMind paper. These include fundamentally degrading human-human relationships, undermining humans’ ability to accept different points of view and severely worsening social atomization, and deepening social dissatisfaction. The extent to which these risks will be realized is, to be sure, much more uncertain than the dangers we highlighted above. But no one who has lived through the changes wrought by social media should doubt that such deep and troubling hazards are plausible and extremely difficult to reverse or mitigate. More than any other, a firm such as OpenAI – which touts its commitment to safety and alignment – should take seriously these threats and not rush them into widespread use.

We acknowledge that your safety note on Voice Mode states that “we recognize that GPT-4o’s audio modalities present a variety of novel risks” and that your system card will provide greater details on your safety testing. But nothing in what we’ve seen suggests your safety testing and outside review grapples adequately with the enormous risks Voice Mode presents. Given the stakes, we find it impossible to believe that internal testing, with limited external review, can possibly justify the massive social experiment you are poised to inflict on society.

It is noteworthy that Google – a for-profit company that is not governed by a non-profit with a safety and alignment mandate – affirmatively decided not to adopt anthropomorphic voices for its AI assistant.

However, if OpenAI’s Voice Mode gains traction and popularity, it will be difficult for Google and the handful of other leading AI companies not to risk competitive
disadvantage and not also adopt human-sounding voices and an anthropomorphic approach to its AI assistant.

There’s still time to put this genie back in the bottle, but not for long. We urge you to suspend Voice Mode and the use of human-sounding voices and attributes for ChatGPT; to reassess future design choices in light of potential harms; and to pledge not to release anthropomorphic AI systems, at least not before there has been extensive and public testing, review and debate of the potential consequences, and conclusive evidence that consequential social harm can be avoided.

Sincerely,

Robert Weissman
President

Rick Claypool
Research Director, President’s Office