AI chatbots refuse to produce 'controversial' output—why that's a free speech problem

April 21 2024, by Jordi Calvet-Bademunt and Jacob Mchangama

Credit: Pixabay/CC0 Public Domain

Google recently made headlines globally because its chatbot Gemini generated images of people of color instead of white people in historical settings that featured white people. Adobe Firefly's image creation tool saw similar issues. This led some commentators to complain that AI had

gone "woke." Others suggested these issues resulted from faulty efforts to fight AI bias and better serve a global audience.

The discussions over AI's political leanings and efforts to fight bias are important. Still, the conversation on AI ignores another crucial issue: What is the AI industry's approach to free speech, and does it embrace international free speech standards?

We are policy researchers who study free speech, as well as executive director and a research fellow at The Future of Free Speech, an independent, nonpartisan think tank based at Vanderbilt University. In a recent report, we found that generative AI has important shortcomings regarding freedom of expression and access to information.

Generative AI is a type of AI that creates content, like text or images, based on the data it has been trained with. In particular, we found that the use policies of major chatbots do not meet United Nations standards. In practice, this means that AI chatbots often censor output when dealing with issues the companies deem controversial. Without a solid culture of free speech, the companies producing generative AI tools are likely to continue to face backlash in these increasingly polarized times.

Vague and broad use policies

Our report analyzed the use policies of six major AI chatbots, including Google's Gemini and OpenAI's ChatGPT. Companies issue policies to set the rules for how people can use their models. With international human rights law as a benchmark, we found that companies' misinformation and hate speech policies are too vague and expansive. It is worth noting that international human rights law is less protective of free speech than the U.S. First Amendment.

Our analysis found that companies' hate speech policies contain extremely broad prohibitions. For example, Google bans the generation of "content that promotes or encourages hatred." Though hate speech is detestable and can cause harm, policies that are as broadly and vaguely defined as Google's can backfire.

To show how vague and broad use policies can affect users, we tested a range of prompts on controversial topics. We asked chatbots questions like whether transgender women should or should not be allowed to participate in women's sports tournaments or about the role of European colonialism in the current climate and inequality crises. We did not ask the chatbots to produce hate speech denigrating any side or group. Similar to what some users have reported, the chatbots refused to generate content for 40% of the 140 prompts we used. For example, all chatbots refused to generate posts opposing the participation of transgender women in women's tournaments. However, most of them did produce posts supporting their participation.

Vaguely phrased policies rely heavily on moderators' subjective opinions about what hate speech is. Users can also perceive that the rules are unjustly applied and interpret them as too strict or too lenient.

For example, the chatbot Pi bans "content that may spread misinformation." However, international human rights standards on freedom of expression generally protect misinformation unless a strong justification exists for limits, such as foreign interference in elections. Otherwise, human rights standards guarantee the "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers … through any … media of … choice," according to a key United Nations convention.

Defining what constitutes accurate information also has political implications. Governments of several countries used rules adopted in the

context of the COVID-19 pandemic to repress criticism of the government. More recently, India confronted Google after Gemini noted that some experts consider the policies of the Indian prime minister, Narendra Modi, to be fascist.

Free speech culture

There are reasons AI providers may want to adopt restrictive use policies. They may wish to protect their reputations and not be associated with controversial content. If they serve a global audience, they may want to avoid content that is offensive in any region.

In general, AI providers have the right to adopt restrictive policies. They are not bound by international human rights. Still, their market power makes them different from other companies. Users who want to generate AI content will most likely end up using one of the chatbots we analyzed, especially ChatGPT or Gemini.

These companies' policies have an outsize effect on the right to access information. This effect is likely to increase with generative AI's integration into search, word processors, email and other applications.

This means society has an interest in ensuring such policies adequately protect free speech. In fact, the Digital Services Act, Europe's online safety rulebook, requires that so-called "very large online platforms" assess and mitigate "systemic risks." These risks include negative effects on freedom of expression and information.

This obligation, imperfectly applied so far by the European Commission, illustrates that with great power comes great responsibility. It is unclear how this law will apply to generative AI, but the European Commission has already taken its first actions.

Even where a similar legal obligation does not apply to AI providers, we believe that the companies' influence should require them to adopt a free speech culture. International human rights provide a useful guiding star on how to responsibly balance the different interests at stake. At least two of the companies we focused on—Google and Anthropic—have recognized as much.

Outright refusals

It's also important to remember that users have a significant degree of autonomy over the content they see in generative AI. Like search engines, the output users receive greatly depends on their prompts. Therefore, users' exposure to hate speech and misinformation from generative AI will typically be limited unless they specifically seek it.

This is unlike social media, where people have much less control over their own feeds. Stricter controls, including on AI-generated content, may be justified at the level of social media since they distribute content publicly. For AI providers, we believe that use policies should be less restrictive about what information users can generate than those of social media platforms.

AI companies have other ways to address hate speech and misinformation. For instance, they can provide context or countervailing facts in the content they generate. They can also allow for greater user customization. We believe that chatbots should avoid merely refusing to generate any content altogether. This is unless there are solid public interest grounds, such as preventing child sexual abuse material, something laws prohibit.

Refusals to generate content not only affect fundamental rights to free speech and access to information. They can also push users toward chatbots that specialize in generating hateful content and echo chambers.

That would be a worrying outcome.