AI Bots Can Seem Sentient. Students Need Guardrails.

> Faculty members have welcomed chat bots into their classrooms. But how will they help students manage AI's sometimes-disturbing replies?

By Susan D'Agostino (/users/susan-dagostino)

// February 22, 2023

Facebook founder Mark Zuckerberg once advised tech founders to "move fast and break things." But in moving fast, some argue that he "broke (https://www.npr.org/2023/02/16/1157180971/10-things-to-know-about-how-social-media-affects-teens-brains)" those young people whose social media exposure has led to depression, anxiety, cyberbullying, poor body image and loss of privacy or sleep during a vulnerable life stage.

Now, Big Tech is moving fast again with the release of sophisticated AI chat bots, not all of which have been adequately vetted before their public release.

OpenAI launched an artificial intelligence arms race in late 2022 with the release of ChatGPT—a sophisticated AI chat bot that interacts with users in a conversational way, but also lies and reproduces systemic societal biases. The bot became an instant global sensation, even as it raised concerns (https://www.insidehighered.com/news/2023/01/31/chatgpt-sparks-debate-how-design-student-assignments-now) about cheating and how college writing might change (https://www.insidehighered.com/news/2022/10/26/machines-can-craft-essays-how-should-writing-be-taught-now).

In response, Google moved up the release of its rival chat bot, Bard (https://www.bbc.com/news/technology-64546299), to Feb. 6, despite employee leaks that the tool was not ready. The company's stock sank (https://time.com/6254226/alphabet-google-bard-100-billion-ai-error/) after a series of product missteps. Then, a day later, and in an apparent effort not to be left out of the AI–chat bot party, Microsoft launched (https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/) its AI-powered Bing search engine. Early users quickly found that the eerily human-sounding bot produced unhinged (https://www.axios.com/2023/02/16/bing-artificial-intelligence-chat%20bot-issues), manipulative (https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams), rude (https://www.businessinsider.com/bing-chatgpt-ai-chat%20bot-argues-angry-responses-falls-in-love-2023-2), threatening (https://futurism.com/microsoft-bing-ai-threatening) and false (https://www.popsci.com/technology/conversational-ai-inaccurate/) responses, which prompted the company to implement changes—and AI ethicists to express (https://www.forbes.com/sites/cindygordon/2023/02/21/why-is-microsofts-bing-chat%20bot-raising-ethical-eyebrows/?sh=7c6d9c9a7e40) reservations.

Rushed decisions, especially in technology, can lead to what's called "path dependence," a phenomenon in which early decisions constrain later events or decisions, according to Mark Hagerott, a historian of technology and chancellor of the North Dakota University system who earlier served as deputy director of the U.S. Naval Academy's Center for Cyber Security Studies. The QWERTY keyboard,

by some accounts (https://link.springer.com/article/10.1023/A:1007811722566) (not everyone agrees), may have been designed in the late 1800s to minimize jamming of high-use typewriter letter keys. But the design persists even on today's cellphone keyboards, despite the suboptimal arrangement of the letters.

"Being deliberate doesn't mean we're going to stop these things, because they're almost a force of nature," Hagerott said about the presence of AI tools in higher ed. "But if we're engaged early, we can try to get more positive effects than negative effects."

That's why North Dakota University system leaders launched (https://www.grandforksherald.com/news/north-dakota/north-dakota-higher-ed-leaders-form-task-force-to-combat-negative-effects-of-artificial-intelligence) a task force to develop strategies for minimizing the negative effects of artificial intelligence on their campus communities. As these tools infiltrate higher ed, many other colleges and professors have developed policies designed to ensure academic integrity and promote creative uses of the emerging tech in the classroom. But some academics are concerned that, by focusing on academic honesty and classroom innovation, the policies have one blind spot. That is, colleges have been slow to recognize that students may need AI literacy training that helps them navigate emotional responses to eerily human-sounding bots' sometimes-disturbing replies.

"I can't see the future, but I've studied enough of these technologies and lived with them to know that you can really get some things wrong," Hagerott said. "Early decisions can lock in, and they could affect students' learning and dependency on tools that, in the end, may prove to be less than ideal to the development of critical thinking and discernment."

AI Policies Take Shape—and Require Updates

When Emily Pitts Donahoe, associate director of instructional support at the University of Mississippi's Center for Teaching and Learning, began teaching this semester, she understood that she needed to address her students' questions and excitement surrounding ChatGPT. In her mind, the university's academic integrity policy covered instances in which students, for example, copied or misrepresented work as their own. That freed her to craft a policy that began from a place of openness and curiosity.

Donahoe opted to co-create a course policy on generative AI writing tools with her students. She and the students engaged in an exercise in which they all submitted suggested guidelines for a class policy, after which they upvoted each other's suggestions. Donahoe then distilled the top votes into a document titled "Academic integrity guidelines for use and attribution of AI."

Some allowable uses in Donahoe's policy include using AI writing generators to brainstorm, overcome writer's block, inspire ideas, draft an outline, edit and proofread. The impermissible uses included taking what the writing generator wrote at face value, including huge chunks of its prose in an assignment and failing to disclose use of an AI writing tool or the extent to which it was used.

Donahoe was careful to emphasize that the rules they established applied to her class, but that other professors' expectations may differ. She also disclosed that such a policy was as new to her as to the students, given the quick rise of ChatGPT and rival tools.

"It may turn out at the end of the semester that I think that everything I've just said is crap," Donahoe said. "I'm still trying to be flexible for when new versions of this technology emerge or as we adapt to it ourselves."

Like Donahoe, many professors have designed new individual policies (https://docs.google.com/document/d/1WpCeTyiWCPQ9MNCsFeKMDQLSTsg1oKfNIH6MzoSFXqQ/mobilebasic) with similar themes. At the same time, many college teaching and learning centers have developed new resource pages with guidance and links to articles such as *Inside Higher Ed*'s "ChatGPT Advice Academics Can Use Now (https://www.insidehighered.com/news/2023/01/12/academic-experts-offer-advice-chatgpt)."

The academic research community has responded with new policies of its own. For example, ArXiv (https://blog.arxiv.org/2023/01/31/arxiv-announces-new-policy-on-chatgpt-and-similar-tools/), the open-access repository of pre- and postprints, and the journals *Nature* and *Science* have all developed new policies (https://www.nature.com/articles/d41586-023-00107-z) that share two main directives. First, AI language tools cannot be listed as authors, since they cannot be held accountable for a paper's contents. Second, researchers must document use of an AI language tool.

Nonetheless, academics' efforts to navigate the new AI-infused landscape remain a work in progress. ArXiv, for example, first released its policy on Jan. 31 but issued an update on Feb. 7. Also, many have discovered that documenting use is a necessary but insufficient condition for acceptable use. For example, when Vanderbilt University employees wrote an email to students about the recent shooting at Michigan State University in which three people were killed and five were wounded, after which the gunman killed himself, they included a note at the bottom that said, "Paraphrase from OpenAI's ChatGPT." Many found such a usage, while acknowledged, to be deeply insensitive and flawed (https://www.independent.co.uk/news/world/americas/vanderbilt-msu-shooting-chatgpt-email-b2285031.html).

Those who are at work drafting such policies are grappling with some of academe's most cherished values, including academic integrity, learning and life itself. Given the speed and the stakes, these individuals must think fast while proceeding with care. They must be explicit while remaining open to change. They must also project authority while exhibiting humility in the midst of uncertainty.

But academic integrity and accuracy are not the only issues related to AI chat bots. Further, students already have a template for understanding (https://oneusefulthing.substack.com/p/my-class-required-ai-heres-what-ive) these issues, according to Ethan Mollick, associate professor of management and academic director at Wharton Interactive at the Wharton School at the University of Pennsylvania.

Policies might go beyond academic honesty and creative classroom uses, according to many academics consulted for this story. That is, the bots' underlying technology—large language models—is

intended to mimic human behavior. Though the machines are not sentient, humans often respond to them with emotion. As Big Tech accelerates its use of the public as a testing ground for the suspiciously human-sounding chat bots, students may be underprepared to manage their emotional responses. In this sense, AI chat bot policies that address literacy may help protect students' mental health.

"There are enough stressors in the world that really are impacting our students," Andrew Armacost, president of the University of North Dakota, said. AI chat bots "add potentially another dimension."

An Often-Missing Ingredient AI Chat Bot Policy

Bing AI is "much more powerful than ChatGPT" and "often unsettling," Mollick wrote in a tweet thread (https://twitter.com/emollick/status/1627161768966463488?cxt=HHwWgIC99b_A65QtAAAA) about his engagement with the bot before Microsoft imposed restrictions.

"I say that as someone who knows that there is no actual personality or entity behind a [large language model]," Mollick wrote. "But, even knowing that it was basically auto-completing a dialog based on my prompts, it felt like you were dealing with a real person. I never attempted to 'jailbreak' the chat bot or make it act in any particular way, but I still got answers that felt extremely personal, and interactions that made the bot feel intentional."

The lesson, according to Mollick, is that users can easily be fooled into thinking that an AI chat bot is sentient.

That concerns Hagerott, who, when he taught college, calibrated his discussions with students based on how long they had been in college.

"In those formative freshman years, I was always so careful," Hagerott said. "I could talk in certain ways with seniors and graduate students, but boy, with freshmen, you want to encourage them, have them know that people learn in different ways, that they'll get through this."

Hagerott is concerned that some students lack AI literacy training that supports understanding of their emotional relationships to the large language models, including potential mental health risks. A tentative student who asks an AI chat bot a question about their self-worth, for example, may be unprepared to manage their own emotional response to a cold, negative response, Hagerott said.

Hollis Robbins, dean of the University of Utah's College of Humanities, shares similar concerns. Colleges have long used institutional chat bots on their websites to support access to library resources or to enhance student success and retention. But such college-specific chat bots often have carefully engineered responses to the kinds of sensitive questions college students are prone to ask, including questions about their physical or mental health, Robbins said.

“I’m not sure it is always clear to students which is ChatGPT and which is a university-authorized and promoted chat,” Robbins said, adding that she looks forward to a day when colleges may have their own ChatGPT-like platforms designed for their students and researchers.

To be clear, none of the academics interviewed for this article argued that colleges should ban AI chat bots. The tools have infiltrated society as much as higher ed. But all expressed concern that some colleges’ policies may not be keeping pace with Big Tech’s AI release of undertested tools.

And so, new policies might focus on protecting student mental health, in addition to problems with accuracy and bias.

“It’s imperative to teach students that chat bots have no sentience or reasoning and that these synthetic interactions are, despite what they seem, still nothing more than predictive text generation,” Marc Watkins, lecturer in composition and rhetoric at the University of Mississippi, said of the shifting landscape. “This responsibility certainly adds another dimension to the already-challenging task of trying to teach AI literacy.”

Read more by

By Susan D'Agostino (/users/susan-dagostino)