Democracy Dies in Darkness

Reddit slams ‘unethical experiment’ that deployed secret AI bots in forum

The platform’s chief legal officer called out the University of Zurich team that deployed bots on r/changemyview to study how AI can influence opinions.

Today at 11:00 a.m. EDT

By Vivian Ho

Reddit is raising the alarm about what it called an “improper and highly unethical experiment” by a group of University of Zurich researchers, who secretly deployed AI bots on a popular forum to study how artificial intelligence can influence human opinion.

Moderators on the changemyview subreddit alerted the group’s 3.8 million users over the weekend of the “unauthorized experiment” that had been unfolding over the past few months. The moderators said they had received notice from the researchers as “part of a disclosure step” in the study in which the researchers said they had used multiple accounts to post on the subreddit without disclosing that AI was used to write comments.

The subreddit, which operates as a “place to post an opinion you accept may be flawed” to better understand opposing views, does not allow the use of undisclosed AI-generated content or bots. “The researchers did not contact us ahead of the study and if they had, we would have declined,” the moderators wrote in the post.

In an administrator post identifying himself as Reddit’s chief legal officer, Ben Lee, using the Reddit username traceroo, called the experiment “improper and highly unethical,” as well as “deeply wrong on both a moral and legal level.”

“We are in the process of reaching out to the University of Zurich and this particular research team with formal legal demands,” Lee said in a post in the forum on Monday. “We want to do everything we can to support the community and ensure that the researchers are held accountable for their misdeeds here.” He did not immediately respond to a request for comment about what those demands might entail.

Melanie Nyfeler, a spokeswoman for the University of Zurich, confirmed in an emailed statement on Wednesday that the Ethics Committee of the Faculty of Arts and Social Sciences reviewed a research project last year “investigating the potential of artificial intelligence to reduce polarization in value-based political discourse.” One of four studies associated with this project involved using “large language model (LLM)-driven conversational agents (“chatbots”) in online forums and subreddits,” Nyfeler said.

The committee had advised the researchers that this study would be "exceptionally" challenging because "participants should be informed as much as possible" and "the rules of the platform should be fully complied with," Nyfeler said. But committee assessments are recommendations and not legally binding. "The researchers themselves are responsible for carrying out the project and publishing the results," Nyfeler said.

"In light of these events, the Ethics Committee of the Faculty of Arts and Social Sciences intends to adopt a stricter review process in the future and, in particular, to coordinate with the communities on the platforms prior to experimental studies," Nyfeler said. "The relevant authorities at the University of Zurich are aware of the incidents and will now investigate them in detail and critically review the relevant assessment processes."

Nyfeler added that the researchers have decided not to publish the experiment's results.

In their Saturday post detailing their findings, the subreddit moderators said they had filed an ethics complaint with the university and asked it not to publish the research, arguing that publishing it "would dramatically encourage further intrusion by researchers, contributing to increased community vulnerability to future non-consensual human subjects experimentation."

They said the researchers had shared a draft of the experiment's results with them, which was linked in the post but restricted to those with permission to access it.

Logan MacGregor, one of the subreddit's moderators, told The Washington Post that the researchers' actions left him feeling violated. He had joined the forum about a year ago after disconnecting from most other social media because of how toxic and vitriolic the discourse can get. But r/changemyview, with its well-established rules and ethos, was different, he said.

"I very reluctantly joined Reddit because Rule 1 is 'Remember the human,'" he said. "And then I found this place where you could talk about anything, where any view was permitted and civility was enforced. It was a safe human place for the exchange of ideas."

The researchers, whose names have not been released, used AI bots to run 13 different accounts, one of which purported to be a victim of rape and another a Black man who opposed Black Lives Matter, the moderators wrote in their announcement post.

Using a Reddit account vetted by the forum's moderators, LLMresearchteam, the researchers responded to concerns and criticisms from the community, posting that they had to conduct their experiment without alerting users or getting consent from unknowing participants because "an unaware setting was necessary."

The researchers said they had attempted to launch 34 accounts at first, but 21 were shadow-banned within the first two weeks, meaning that the accounts could still view the subreddit, but their posts were hidden from the rest of the community. The remaining 13 accounts averaged about 10 to 15 posts a day, the researchers said, an amount they described as negligible given the 7,000 posts averaged per day by the entire subreddit.

In total, the bots posted about 1,700 comments, according to moderators.

“Previous research on LLM persuasion has only taken place in highly artificial environments, often involving financially incentivized participants,” the researchers wrote. “These settings fail to capture the complexity of real-world interactions, which evolve in spontaneous and unpredictable ways with numerous contextual factors influencing how opinions change over time. Consent-based experiments lack ecological validity because they can’t simulate how users behave when unaware of persuasive attempts — just as they would be in the presence of bad actors.”

The researchers apologized for any disruption their study may have caused but maintained that “every decision” throughout their study was guided by the principles of “ethical scientific conduct, user safety and transparency.” They argued that the ethics committee at the University of Zurich had reviewed and approved their approach, and they “acknowledged that prior consent was impractical.”

The university, while sharing its statement, did not respond to further questions.

Angeliki Kerasidou, an associate professor in bioethics at the University of Oxford, told The Post that while deception in research can be justified in cases where the research has high social value, informed consent has come to be considered a cornerstone of ethical research “because it underpins respect for persons and supports individual autonomy."

The experiment conducted on r/changemyview “exemplifies the importance of thinking about the social (and not just scientific) value of research, and of engaging with the research communities when planning research projects,” Kerasidou wrote in an email.

MacGregor said one silver lining of the experiment was that it brought attention to a pervasive issue: how to protect the “precious few civil human spaces” that still exist on the internet.

“The researchers are right about the existential challenges of AI,” he said. “The way they went about this was wrong. But I think they were well-meaning, and one of the things I’d personally like to see come out of this is … a better way forward as all of us wrestle with AI. It’s not going away. You can’t turn it off. But what can we do to keep these spaces human?”