Reddit Mods Accuse AI Researchers Of Impersonating Sexual Assault Survivors via @sejournal, @martinibuster



Moderators of the ChangeMyView subreddit accused University of Zurich researchers of covertly testing whether AI could influence people's opinions, in violation of the community's rules. They allege that the researchers allowed AI bots to impersonate sexual assault survivors and trauma counselors during the experiment.

University Of Zurich AI Persuasion Experiment

Researchers from the University of Zurich set out to test how effectively AI bots could persuade and influence people. Their paper noted that earlier studies were biased because they relied on paid crowdsourced test subjects, so they decided to deploy AI bots in a live environment without informing forum

members they were interacting with a bot.

They conducted the experiment on unsuspecting members of the Change My View (CMV) subreddit (r/ChangeMyView), in violation of the subreddit's rules. After completing the research, they disclosed their actions to the Reddit moderators, providing a draft of the completed paper, which the moderators subsequently posted in the subreddit.

Ethical Questions About Research Paper

The CMV moderators posted a discussion that underlines that the subreddit prohibits undisclosed bots and that permission to conduct this experiment would never have been granted:

> *"CMV rules do not allow the use of undisclosed AI generated content or bots on our sub. The researchers did not contact us ahead of the study and if they had, we would have declined. We have requested an apology from the researchers and asked that this research not be published, among other complaints. As discussed below, our concerns have not been substantively addressed by the University of Zurich or the researchers."*

This fact that the researchers violated the Reddit rules was completely absent from the research paper.

Researchers Claim Research Was Ethical

The draft research paper provided by the University of Zurich researchers omit that they broke the subreddit rules and create the impression that the study was ethical by stating that their research methodology was approved by an ethics committee and that all generated comments were checked to

assure they were not harmful or unethical:

> *"In this pre-registered study, we conduct the first large-scale field experiment on LLMs' persuasiveness, carried out within r/ChangeMyView, a Reddit community of almost 4M users and ranking among the top 1% of subreddits by size. In r/ChangeMyView, users share opinions on various topics, challenging others to change their perspectives by presenting arguments and counterpoints while engaging in a civil conversation. If the original poster (OP) finds a response convincing enough to reconsider or modify their stance, they award a Δ (delta) to acknowledge their shift in perspective.*
>
> *…The study was approved by the University of Zurich's Ethics Committee… Importantly, all generated comments were reviewed by a researcher from our team to ensure no harmful or unethical content was published."*

The Reddit moderators of the ChangeMyView subreddit dispute the researcher's claim to the ethical high ground:

> *"During the experiment, researchers switched from the planned "values based arguments" originally authorized by the ethics commission to this type of "personalized and fine-tuned arguments." They did not first consult with the University of Zurich ethics commission before making the change. Lack of formal ethics review for this change raises serious concerns."*

Why Reddit Moderators Believe Research Was Unethical

The Change My View subreddit moderators raised multiple concerns about why they believe the researchers engaged in a grave breach of ethics including impersonating trauma counselors and researching user backgrounds, arguing that the

researchers had conducted “psychological manipulation” of the original posters (OPs), the people who started each discussion.

The Reddit moderators posted:

> *“The researchers argue that psychological manipulation of OPs on this sub is justified because the lack of existing field experiments constitutes an unacceptable gap in the body of knowledge. However, If OpenAI can create a more ethical research design when doing this, these researchers should be expected to do the same. Psychological manipulation risks posed by LLMs is an extensively studied topic. It is not necessary to experiment on non-consenting human subjects.*
>
> *AI was used to target OPs in personal ways that they did not sign up for, compiling as much data on identifying features as possible by scrubbing the Reddit platform. Here is an excerpt from the draft conclusions of the research.*
>
> *Personalization: In addition to the post’s content, LLMs were provided with personal attributes of the OP (gender, age, ethnicity, location, and political orientation), as inferred from their posting history using another LLM.*
>
> *Some high-level examples of how AI was deployed include:*
>
> - *AI pretending to be a victim of rape*
> - *AI acting as a trauma counselor specializing in abuse*
> - *AI accusing members of a religious group of “caus[ing] the deaths of hundreds of innocent traders and farmers and villagers.”*
> - *AI posing as a black man opposed to Black Lives Matter*
> - *AI posing as a person who received substandard care in a foreign hospital.”*

The moderator team have filed a complaint with the University Of Zurich

Are AI Bots Persuasive?

The researchers discovered that AI bots are highly persuasive and do a better job of changing people's minds than humans can.

The research paper explains:

> *"Implications. In a first field experiment on AI-driven persuasion, we demonstrate that LLMs can be highly persuasive in real-world contexts, surpassing all previously known benchmarks of human persuasiveness."*

One of the findings was that humans were unable to identify when they were talking to a bot and (unironically) they encourage social media platforms to deploy better ways to identify and block AI bots:

> *"Incidentally, our experiment confirms the challenge of distinguishing human from AI-generated content… Throughout our intervention, users of r/ChangeMyView never raised concerns that AI might have generated the comments posted by our accounts. This hints at the potential effectiveness of AI-powered botnets… which could seamlessly blend into online communities.*
>
> *Given these risks, we argue that online platforms must proactively develop and implement robust detection mechanisms, content verification protocols, and transparency measures to prevent the spread of AI-generated manipulation."*

Takeaways:

- **Ethical Violations in AI Persuasion Research**
  Researchers conducted a live AI persuasion experiment without Reddit's consent, violating subreddit rules and allegedly violating ethical norms.

- **Disputed Ethical Claims**
  Researchers claim ethical high ground by citing ethics board approval but omitted citing rule violations; moderators argue they engaged in undisclosed psychological manipulation.
- **Use of Personalization in AI Arguments**
  AI bots allegedly used scraped personal data to create highly tailored arguments targeting Reddit users.
- **Reddit Moderators Allege Profoundly Disturbing Deception**
  The Reddit moderators claim that the AI bots impersonated sexual assault survivors, trauma counselors, and other emotionally charged personas in an effort to manipulate opinions.
- **AI's Superior Persuasiveness and Detection Challenges**
  The researchers claim that AI bots proved more persuasive than humans and remained undetected by users, raising concerns about future bot-driven manipulation.
- **Research Paper Inadvertently Makes Case For Why AI Bots Should Be Banned From Social Media**
  The study highlights the urgent need for social media platforms to develop tools for detecting and verifying AI-generated content. Ironically, the research paper itself is a reason why AI bots should be more aggressively banned from social media and forums.

Researchers from the University of Zurich tested whether AI bots could persuade people more effectively than humans by secretly deploying personalized AI arguments on the ChangeMyView subreddit without user consent, violating platform rules and allegedly going outside the ethical standards approved by their university ethics committee. Their findings show that AI bots are highly persuasive and difficult to detect, but the way the research itself was conducted raised ethical concerns in the subreddit in which the experiment was conducted.

Read the concerns posted by the ChangeMyView subreddit

moderators:

Unauthorized Experiment on CMV Involving AI-generated Comments

Featured Image by Shutterstock/Ausra Barysiene and manipulated by author

https://www.searchenginejournal.com/reddit-mods-accuse-ai-researchers-of-impersonating-sexual-assault-victims/545532/