Users trust AI as much as humans for flagging problematic content

September 16 2022, by Matt Swayne

Credit: Unsplash/CC0 Public Domain

Social media users may trust artificial intelligence (AI) as much as human editors to flag hate speech and harmful content, according to researchers at Penn State.

The researchers said that when users think about positive attributes of machines, like their accuracy and objectivity, they show more faith in AI. However, if users are reminded about the inability of machines to make subjective decisions, their trust is lower.

The findings may help developers design better AI-powered content curation systems that can handle the large amounts of information currently being generated while avoiding the perception that the material has been censored, or inaccurately classified, said S. Shyam Sundar, James P. Jimirro Professor of Media Effects in the Donald P. Bellisario College of Communications and co-director of the Media Effects Research Laboratory.

"There's this dire need for content moderation on social media and more generally, online media," said Sundar, who is also an affiliate of Penn State's Institute for Computational and Data Sciences. "In traditional media, we have news editors who serve as gatekeepers. But online, the gates are so wide open, and gatekeeping is not necessarily feasible for humans to perform, especially with the volume of information being generated. So, with the industry increasingly moving towards automated solutions, this study looks at the difference between human and automated content moderators, in terms of how people respond to them."

Both human and AI editors have advantages and disadvantages. Humans tend to more accurately assess whether content is harmful, such as when it is racist or potentially could provoke self-harm, according to Maria D. Molina, assistant professor of advertising and public relations, Michigan State, who is first author of the study. People, however, are unable to process the large amounts of content that is now being generated and shared online.

On the other hand, while AI editors can swiftly analyze content, people often distrust these algorithms to make accurate recommendations, as

well as fear that the information could be censored.

"When we think about automated content moderation, it raises the question of whether artificial intelligence editors are impinging on a person's freedom of expression," said Molina. "This creates a dichotomy between the fact that we need content moderation—because people are sharing all of this problematic content—and, at the same time, people are worried about AI's ability to moderate content. So, ultimately, we want to know how we can build AI content moderators that people can trust in a way that doesn't impinge on that freedom of expression."

Transparency and interactive transparency

According to Molina, bringing people and AI together in the moderation process may be one way to build a trusted moderation system. She added that transparency—or signaling to users that a machine is involved in moderation—is one approach to improving trust in AI. However, allowing users to offer suggestions to the AIs, which the researchers refer to as "interactive transparency," seems to boost user trust even more.

To study transparency and interactive transparency, among other variables, the researchers recruited 676 participants to interact with a content classification system. Participants were randomly assigned to one of 18 experimental conditions, designed to test how the source of moderation—AI, human or both—and transparency—regular, interactive or no transparency—might affect the participant's trust in AI content editors. The researchers tested classification decisions—whether the content was classified as "flagged" or "not flagged" for being harmful or hateful. The "harmful" test content dealt with suicidal ideation, while the "hateful" test content included hate speech.

Among other findings, the researchers found that users' trust depends on

whether the presence of an AI content moderator invokes positive attributes of machines, such as their accuracy and objectivity, or negative attributes, such as their inability to make subjective judgments about nuances in human language.

Giving users a chance to help the AI system decide whether online information is harmful or not may also boost their trust. The researchers said that study participants who added their own terms to the results of an AI-selected list of words used to classify posts trusted the AI editor just as much as they trusted a human editor.

Ethical concerns

Sundar said that relieving humans of reviewing content goes beyond just giving workers a respite from a tedious chore. Hiring human editors for the chore means that these workers are exposed to hours of hateful and violent images and content, he said.

"There's an ethical need for automated content moderation," said Sundar, who is also director of Penn State's Center for Socially Responsible Artificial Intelligence. "There's a need to protect human content moderators—who are performing a social benefit when they do this—from constant exposure to harmful content day in and day out."

According to Molina, future work could look at how to help people not just trust AI, but also to understand it. Interactive transparency may be a key part of understanding AI, too, she added.

"Something that is really important is not only trust in systems, but also engaging people in a way that they actually understand AI," said Molina. "How can we use this concept of interactive transparency and other methods to help people understand AI better? How can we best present AI so that it invokes the right balance of appreciation of machine ability

and skepticism about its weaknesses? These questions are worthy of research."

The researchers present their findings in the current issue of the *Journal of Computer-Mediated Communication*.

More information: Maria D Molina et al, When AI moderates online content: effects of human collaboration and interactive transparency on user trust, *Journal of Computer-Mediated Communication* (2022). DOI: 10.1093/jcmc/zmac010

Provided by Pennsylvania State University