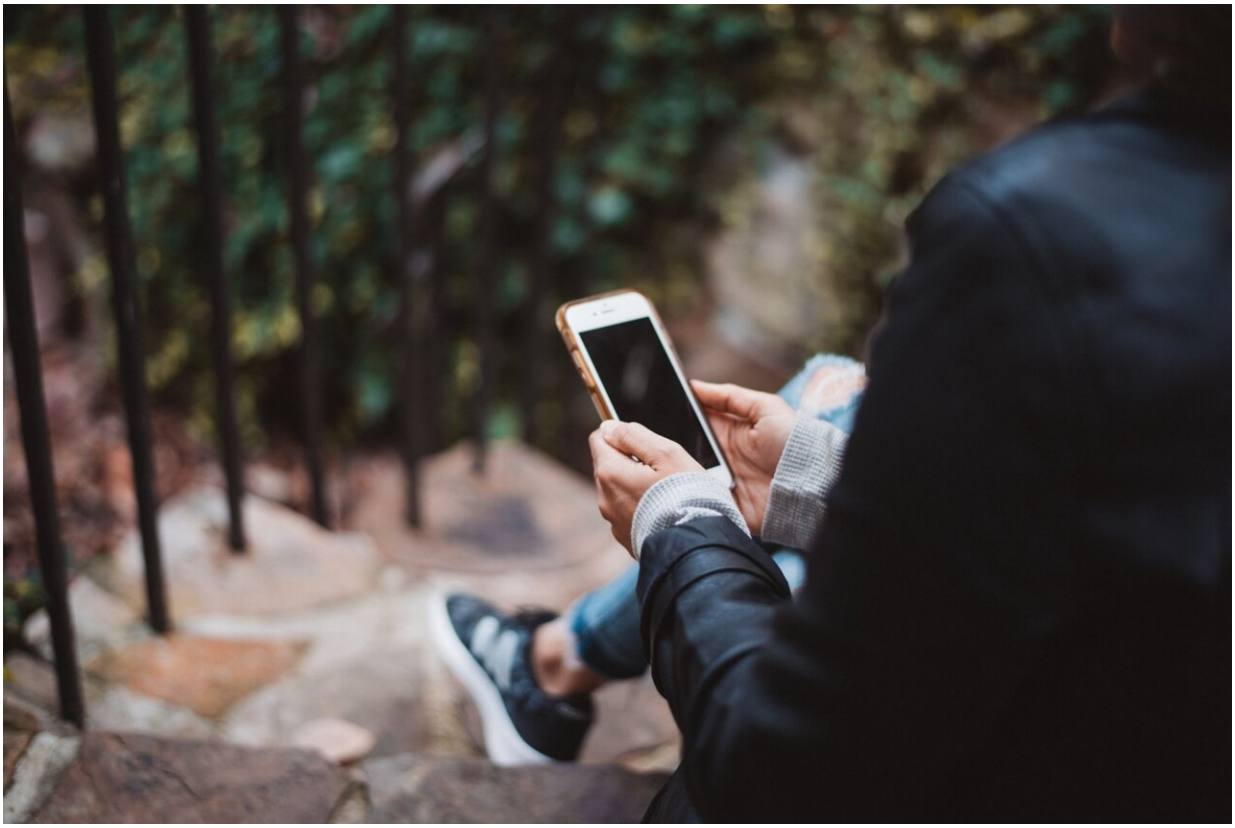
Tricking algorithms: AI tools could boost social media users' privacy

December 1 2022



Credit: Unsplash/CC0 Public Domain

Smart AI tools could protect social media users' privacy by tricking algorithms designed to predict their personal opinions, a study suggests.

Digital assistants could help prevent users from unknowingly revealing their views on social, political and religious issues by fighting AI with AI, researchers say.

Their findings suggest automated assistants could offer users real-time advice on ways to modify their online behavior to mislead AI opinion -detection tools so that their opinions stay private.

Boost privacy

The study is the first to identify how Twitter users can hide their views from opinion-detecting algorithms that can aid targeting by authoritarian governments or fake news sources.

Previous research has focused on steps that owners—rather than users—of social media platforms can take to improve privacy, though such actions can be difficult to enforce, the team says.

Twitter analysis

Edinburgh researchers and academics from New York University Abu Dhabi used data from more than 4,000 Twitter users in the U.S.

The team used the data to analyze how AI can predict people's views, based on their online activities and profile.

They also ran tests on designs for an automated assistant to help Twitter users keep private their views on potentially divisive subjects such as atheism, feminism and politics.

Misleading algorithms

Their findings suggest a tool could help users hide their views by identifying key indicators of their opinions on their profiles, such as accounts they follow and interact with.

It could also help conceal people's opinions by recommending actions that suggest to algorithms that they hold the opposite view.

Users' views

Researchers gauged how strongly people on Twitter feel the need to keep private their opinions on divisive topics by carrying out a survey of more than 1,000 U.S.-based users.

Between 15% and 32% of participants with neutral views on contentious subjects felt strongly the need to keep their views private. Even among users with strong opinions, between 10% and 23% wished to keep their views to themselves.

Despite this, the team found the online activities of those users could reveal their beliefs to malicious algorithms without them realizing it.

"Our research has shown AI can use a variety of signals to easily detect users' opinion on many topics without people even discussing them online. We have developed a tool that can suggest accounts users can follow or retweet to mislead AI opinion-detection algorithms, so that they fail to discover people's real stance on a given topic," says Dr. Walid Magdy, School of Informatics.

The study is published in the journal *PNAS Nexus*.

More information: Marcin Waniek et al, Hiding opinions from machine learning, *PNAS Nexus* (2022). DOI: 10.1093/pnasnexus/pgac256

Provided by University of Edinburgh