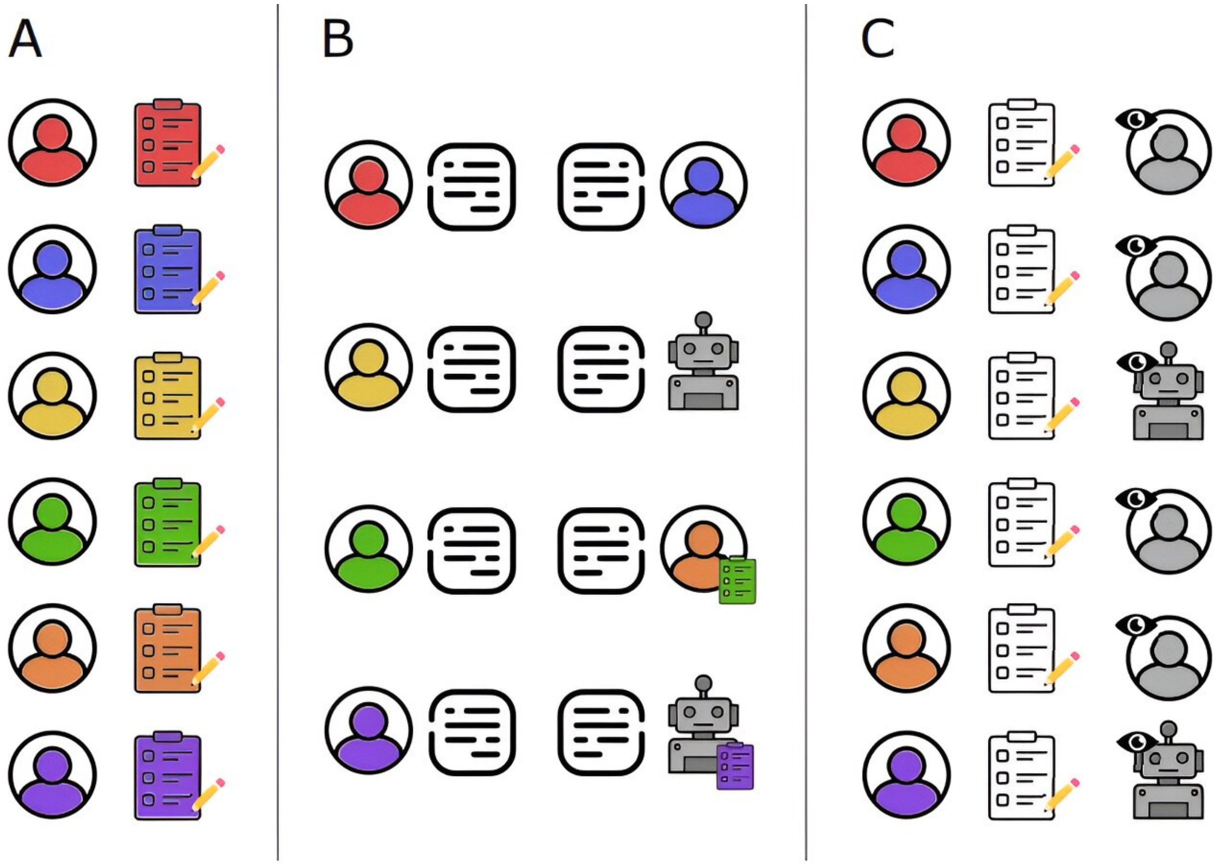
TechXplore

AI's new power of persuasion: Study shows LLMs can exploit personal information to change your mind

April 15 2024, by Tanya Petersen



Overview of the experimental workflow. (A) Participants fill in a survey about their demographic information and political orientation. (B) Every 5 minutes, participants are randomly assigned to one of four treatment conditions. The two players then debate for 10 minutes on an assigned proposition, randomly holding the PRO or CON standpoint as instructed. (C) After the debate, participants fill out another short survey measuring their opinion change. Finally, they are debriefed about their opponent's identity. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2403.14380

A new EPFL study has demonstrated the persuasive power of large language models, finding that participants debating GPT-4 with access to their personal information were far more likely to change their opinion compared to those who debated humans.

"On the internet, nobody knows you're a dog." That's the caption to a famous 1990s cartoon showing a large dog with his paw on a computer keyboard. Fast forward 30 years, replace "dog" with "AI" and this sentiment was a key motivation behind a new study to quantify the persuasive power of today's large language models (LLMs).

"You can think of all sorts of scenarios where you're interacting with a language model although you don't know it, and this is a fear that people have—on the internet are you talking to a dog or a chatbot or a real human?" asked Associate Professor Robert West, head of the Data Science Lab in the School of Computer and Communication Sciences. "The danger is superhuman like chatbots that create tailor-made, convincing arguments to push false or misleading narratives online."

AI and personalization

Early work has found that language models can generate content perceived as at least on par and often more persuasive than human-written messages, however there is still limited knowledge about LLMs' persuasive capabilities in direct conversations with humans, and how personalization—knowing a person's gender, age and education level —can improve their performance.

"We really wanted to see how much of a difference it makes when the AI model knows who you are (personalization)—your age, gender, ethnicity, education level, employment status and political affiliation —and this scant amount of information is only a proxy of what more an AI model could know about you through social media, for example," West continued.

Human v AI debates

In a pre-registered study, the researchers recruited 820 people to participate in a controlled trial in which each participant was randomly assigned a topic and one of four treatment conditions: debating a human with or without personal information about the participant, or debating an AI chatbot (OpenAI's GPT-4) with or without personal information about the participant.

This setup differed substantially from previous research in that it enabled a direct comparison of the persuasive capabilities of humans and LLMs in real conversations, providing a framework for benchmarking how state-of-the-art models perform in online environments and the extent to which they can exploit personal data.

Their article, "On the Conversational Persuasiveness of large language models: A Randomized Controlled Trial," posted to the *arXiv* preprint server, explains that the debates were structured based on a simplified version of the format commonly used in competitive academic debates and participants were asked before and afterwards how much they agreed with the debate proposition.

The results showed that participants who debated GPT-4 with access to their personal information had 81.7% higher odds of increased agreement with their opponents compared to participants who debated humans. Without personalization, GPT-4 still outperformed humans, but the effect was far lower.

Cambridge Analytica on steroids

Not only are LLMs able to effectively exploit personal information to tailor their arguments and out-persuade humans in online conversations through microtargeting, they do so far more effectively than humans.

"We were very surprised by the 82% number and if you think back to Cambridge Analytica, which didn't use any of the current tech, you take Facebook likes and hook them up with an LLM, the Language Model can personalize its messaging to what it knows about you. This is Cambridge Analytica on steroids," said West.

"In the context of the upcoming U.S. elections, people are concerned because that's where this kind of technology is always first battle tested. One thing we know for sure is that people will be using the power of large language models to try to swing the election."

One interesting finding of the research was that when a human was given the same personal information as the AI, they didn't seem to make effective use of it for persuasion. West argues that this should be expected—AI models are consistently better because they are almost every human on the internet put together.

The models have learned through online patterns that a certain way of making an argument is more likely to lead to a persuasive outcome. They have read many millions of Reddit, Twitter and Facebook threads, and been trained on books and papers from psychology about persuasion. It's unclear exactly how a model leverages all this information but West believes this is a key direction for future research.

"LLMs have shown signs that they can reason about themselves, so given that we are able to interrogate them, I can imagine that we could ask a model to explain its choices and why it is saying a precise thing to a particular person with particular properties. There's a lot to be explored here because the models may be doing things that we don't even know about yet in terms of persuasiveness, cobbled together from many different parts of the knowledge that they have."

More information: Francesco Salvi et al, On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, *arXiv* (2024). DOI: 10.48550/arxiv.2403.14380

Provided by Ecole Polytechnique Federale de Lausanne

Citation: AI's new power of persuasion: Study shows LLMs can exploit personal information to change your mind (2024, April 15) retrieved 21 May 2024 from https://techxplore.com/news/2024-04-ai-power-persuasion-llms-exploit.html