One potential harmful effect of AI is its use to produce persuasion and propaganda.

Language models can produce propaganda in the form of bots that interact with users on social media. This can be done to push a political agenda or to make fringe views appear more popular than they are.

As of 2025, AI models are capable of generating images or text that can be used for propaganda purposes for a very low cost. Nevertheless, outputs of these models had a limited influence in the various elections that happened around the world in 2024, perhaps because the size of the market for disinformation is constrained by the demand side.[1]

Wei Dai has described what an extreme case of AI persuasion could look like:

> I'm envisioning that in the future there will also be systems where you can input any conclusion that you want to argue (including moral conclusions) and the target audience, and the system will give you the most convincing arguments for it. At that point people won't be able to participate in any online (or offline for that matter) discussions without risking their object-level values being hijacked.

Current AI models (as of 2024) aren't powerful enough to argue persuasively for arbitrary conclusions. However, if AI continues to improve along its current trajectory, it might not be many years before AI is able to write articles and produce other media for propagandistic purposes more effectively than humans can. These could be precisely tailored to individuals by using things like social media feeds and personal digital data.

Recommender systems on content platforms like YouTube, Twitter, and Facebook use machine learning, and the content they recommend can influence the opinions of billions of people. Some research has looked at the tendency for platforms to promote extremist political views and to thereby help radicalize their userbase.

Apart from humans using persuasive AI for their own ends, there's another class of concerns. Future advanced AI, if misaligned, might use its persuasion abilities to gain influence and power for its own ends. This could look like convincing its operators to "let it out of a box" or give it resources, or creating political chaos in order to disable mechanisms that prevent takeover, as in Gwern's short story "It looks like you're trying to take over the world".

Further reading:

- Beth Barnes's report on risks from AI persuasion

Related

- How might a superintelligence socially manipulate humans?

[1] Hugo Mercier has made a similar argument.

- How could a superintelligent AI use the internet to take over the physical world?
- Isn't the real concern AI-enabled totalitarianism?