Recommender systems already influence your daily life through shaping your online experiences. Facebook pushing cute cat videos in your feed to optimize your engagement might sound innocuous, but these algorithms have also steered users on platforms such as YouTube, Netflix, Facebook, Twitter, and TikTok towards content that was harmful to them.

Generative AI is starting to be used to mass-produce textual and visual content to enact disinformation or propaganda campaigns. Some people are worried that the presence of such AIs on social media could be used to run customized campaigns more effectively in the near future, including highly-tailored attempts to manipulate individual people (though we do not know of deceptive AI being used for customized manipulation as of December 2023).[1]

In the future, AI systems could also engage in a range of deceptive behavior, from subtle emotional manipulation (like a car salesman) to outright lying (like a con artist).[2]

As AI becomes more generally capable, could it go beyond matching humans at these kinds of manipulation, and start to substantially outperform us?

Here are some reasons why we might expect that to happen:

- An AI might learn to analyze human psychology deeply enough to understand the hopes and fears of everyone it negotiates with.
- It would likely have an overwhelming advantage in time and attention to bring to bear, due to thinking more quickly than humans or having multiple copies of itself. You could imagine a sufficiently powerful AI as an organization of thousands of experts who build an extensive psychological profile on the person they want to manipulate, and then spend days worth of thinking through their strategy between every breath the human takes.
- Countries, organizations, and individuals could be convinced with threats and promises, or even appeals to morality and empathy.
- More speculatively, a superintelligence might skip conventional persuasion entirely. Humans' apparent susceptibility to subliminal messaging suggests that there could be more effective ways to manipulate people than persuading them with arguments. It seems plausible that something with a better understanding of neuroscience could find and exploit vulnerabilities in human brains. Brains aren't very secure systems, and when you're dealing with something substantially more intelligent than you, it would be wise to expect the unexpected.

As of yet, all of this manipulation by AI has been orchestrated by humans, but future AI might have its own motivations and thus manipulate us for its own ends.

But how can AIs manipulate emotions if they don't have any?

---

[1] One could imagine, for instance, bots on social media that engage with users while pretending to be human to persuade them of a certain political stance.

[2] As of January 2024, LLMs are capable of deception in toy settings but there are no known instances of deceptive AIs being deployed outside of labs.

As a final point: one might think that if AIs don't have emotions, they will lack something essential in the development of social skills needed to convince humans. Popular culture tends to represent AIs as being like nerdy humans: brilliant at technology but clueless about social skills. This need not be true – persuasion and manipulation are just another kind of cognitive skill that AIs could become very good at. The existence of human sociopaths illustrates that it's not necessary to feel particular emotions to be able to manipulate those emotions in others, and similarly, AI could learn to manipulate humans without feeling what humans feel.

Some individuals have already been convinced by AIs that these AIs could feel empathy for them.[3] [4] [5] [6] In all of these cases, it was clear to the users that they were interacting with an AI. If an AI could hide its nature and pretend to be human, it might make manipulating people even easier.

Alternate phrasings

- How can AIs manipulate humans if they don't have any emotions?

Related

- Won't humans be able to beat an unaligned AI since we have a huge advantage …
- How might AGI kill people?
- What is Vingean uncertainty?
- How could a superintelligent AI use the internet to take over the physical world?

[3] ELIZA, the AI psychiatrist from the 1960s, was able to have users open up with primitive technology.
[4] Google's Lambda AI convinced one engineer it was sentient.
[5] Lesswrong user blaked describes their experience forming emotional attachment with an AI in early 2023.
[6] The Replika app created intimacy with users. Many such users were heartbroken when some of these intimacy features were removed from the app.