ChatGPT can't lie to you, but you still shouldn't trust it, says philosopher

March 10 2023, by Mackenzie Graham

Credit: AI-generated image (disclaimer)

"ChatGPT is a natural language generation platform based on the OpenAI GPT-3 language model."

Why did you believe the above statement? A simple answer is that you trust the author of this article (or perhaps the editor). We cannot verify

everything we are told, so we regularly trust the testimony of friends, strangers, "experts" and institutions.

Trusting someone may not always be the primary reason for believing what they say is true. (I might already know what you've told me, for example.) But the fact that we trust the speaker gives us extra motivation for believing what they say.

AI chatbots therefore raise interesting issues about trust and testimony. We have to consider whether we trust what natural language generators like ChatGPT tell us. Another matter is whether these AI chatbots are even capable of being trustworthy.

Justified beliefs

Suppose you tell me it is raining outside. According to one way philosophers view testimony, I am justified in believing you only if I have reasons for thinking your testimony is reliable—for example, you were just outside—and no overriding reasons for thinking it isn't. This is known as the reductionist theory of testimony.

This view makes justified beliefs—assumptions that we feel entitled to hold—difficult to acquire.

But according to another view of testimony, I would be justified in believing it's raining outside as long as I have no reason to think this statement is false. This makes justified beliefs through testimony much easier to acquire. This is called the non-reductionist theory of testimony.

Note that neither of these theories involves trust in the speaker. My relationship to them is one of reliance, not trust.

Trust and reliance

When I rely on someone or something, I make a prediction that it will do what I expect it to. For example, I rely on my alarm clock to sound at the time I set it, and I rely on other drivers to obey the rules of the road.

Trust, however, is more than mere reliance. To illustrate this, let's examine our reactions to misplaced trust compared with misplaced reliance.

If I trusted Roxy to water my prizewinning tulips while I was on vacation and she carelessly let them die, I might rightly feel betrayed. Whereas if I relied on my automatic sprinkler to water the tulips and it failed to come on, I might be disappointed but would be wrong to feel betrayed.

In other words, trust makes us vulnerable to betrayal, so being trustworthy is morally significant in a way that being reliable is not.

The difference between trust and reliance highlights some important points about testimony. When a person tells someone it is raining, they are not just sharing information; they are taking responsibility for the veracity of what they say.

In philosophy, this is called the assurance theory of testimony. A speaker offers the listener a kind of guarantee that what they are saying is true, and in doing so gives the listener a reason to believe them. We trust the speaker, rather than rely on them, to tell the truth.

If I found out you were guessing about the rain but luckily got it right, I would still feel my trust had been let down because your "guarantee" was empty. The assurance aspect also helps capture why lies seem to us morally worse than false statements. While in both cases you invite me to trust and then let down my trust, lies attempt to use my trust against

me to facilitate the betrayal.

Moral agency

If the assurance view is right, then ChatGPT needs to be capable of taking responsibility for what it says in order to be a trustworthy speaker, rather than merely reliable. While it seems we can sensibly attribute agency to AI to perform tasks as required, whether an AI could be a morally responsible agent is another question entirely.

Some philosophers argue that moral agency is not restricted to human beings. Others argue that AI cannot be held morally responsible because, to quote a few examples, they are incapable of mental states, lack autonomy, or lack the capacity for moral reasoning.

Nevertheless, ChatGPT is not a moral agent; it cannot take responsibility for what it says. When it tells us something, it offers no assurances as to its truth. This is why it can give false statements, but not lie. On its website, OpenAI—which built ChatGPT—says that because the AI is trained on data from the internet, it "may be inaccurate, untruthful, and otherwise misleading at times".

At best, it is a "truth-ometer" or fact-checker—and by many accounts, not a particularly accurate one. While we might sometimes be justified in relying on what it says, we shouldn't trust it.

In case you are wondering, the opening quote of this article was an excerpt of ChatGPT's response when I asked it: "What is ChatGPT?" So you should not have trusted that the statement was true. However, *I* can assure you that it is.

Provided by The Conversation