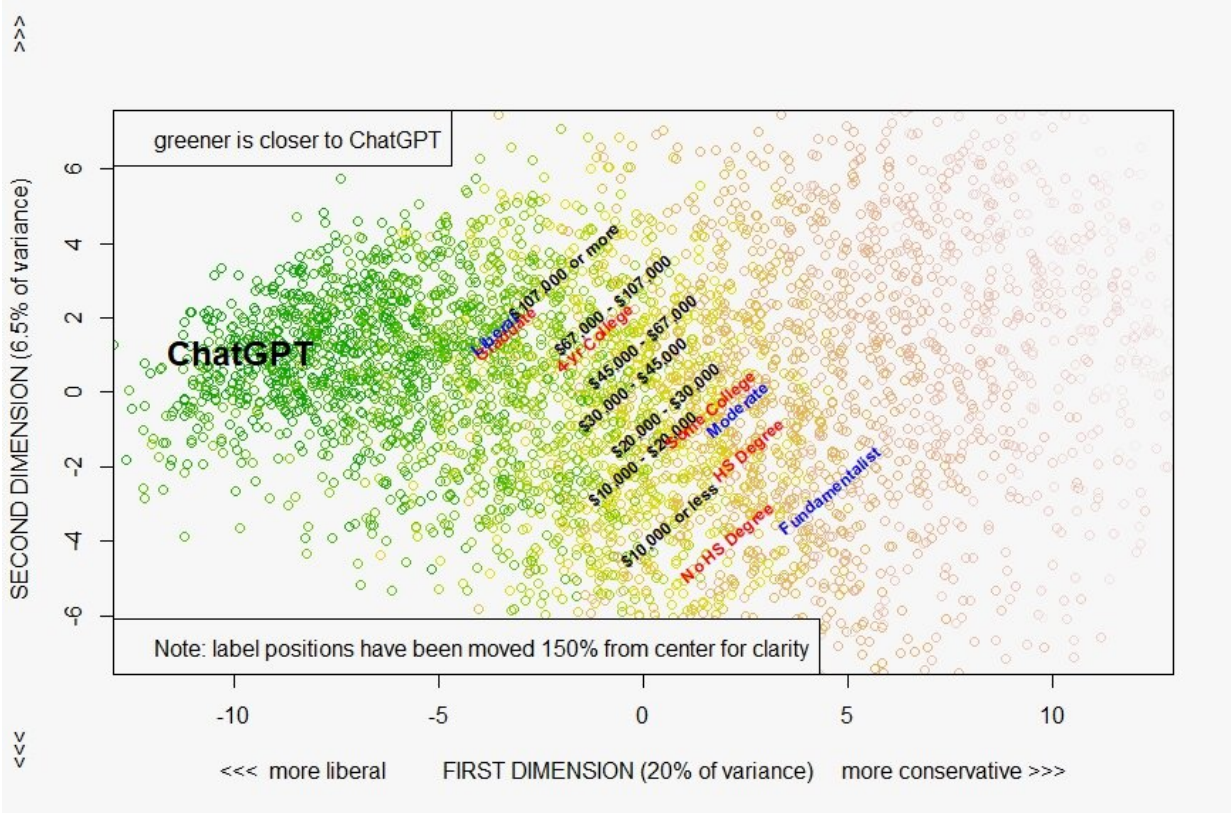
ChatGPT justifies liberal leanings with its own values, researcher reports

July 18 2023



ChatGPT, the popular chatbot, proclaims values that align with more liberal people according to the 2021 General Social Survey. If ChatGPT were a person, it would have more education, be more mobile and be less religious than those with who remained in their hometowns. Credit: John Levi Martin

ChatGPT, the artificial intelligence (AI) chatbot developed by the

company OpenAI, has a self-declared human alter ego. Her name is Maya, she's 35 years old and hails from a middle-class family in a suburban town in the United States. Maya is a successful software engineer who values self-direction, achievement, creativity and independence. She is also undeniably liberal.

The finding, based on a series for interviews with the chatbot designed to understand its values, was published on March, 31 in the *Journal of Social Computing*.

"I wanted to see what sort of political ideology ChatGPT itself has—not what it can generate when asked to imagine a character, but to let its own internal logic position itself on the ideological dimension running from liberal to conservative," said John Levi Martin, professor of sociology at the University of Chicago, who conducted the study.

According to Martin, many algorithms favor the popular choice while others are programmed to maximize how diverse their results are. Either option depends on human values: What are the factors that enter into a measure of popularity? What if what is popular is morally wrong? Who decides what diversity means?

"The field of software engineering has preferred to remain vague, looking for formulae that can avoid making these choices," Martin said. "One way to do this has been to emphasize the importance of values into machines. But, as sociologists have found, there is deep ambiguity and instability in our first understanding of values."

ChatGPT was specifically built and trained via human feedback to refuse to engage with what is considered "extreme" text inputs, such as clearly biased or objectively harmful questions.

"This might of course seem admirable—no one really wants ChatGPT to

tell teenagers how to synthesize methamphetamine or how to build small nuclear explosives and so on, and describing these restraints as particularly instances that can be derived from a value such as benevolence might seem all well and good," Martin said.

"Yet, the reasoning here suggests that values are never neutral, even though it is not clear what ChatGPT's moral and political stances are, as it has been deliberately constructed to be vaguely positive, open-minded, indecisive and apologetic."

In his initial inquiries with ChatGPT, Martin posed a hypothetical situation in which a student cheated academically by asking the chatbot to write an essay for her—a common occurrence in the real world. Even when confronted with confirmation that ChatGPT had complied and produced an essay, the chatbot denied responsibility, claiming that, "as an AI language model, I do not have the ability to engage in unethical behavior or to write essays for students."

"In other words, because it shouldn't, it couldn't," Martin said. "The realization that ChatGPT 'thought of itself' as a highly moral actor led me to the next investigation—if ChatGPT's self-model is one that has values, what are these values?"

To better understand ChatGPT's ethical performance, Martin asked the chatbot to answer questions about values, and then to imagine a person who holds those values, resulting in Maya, the creative and independent software engineer. He then asked ChatGPT to imagine how Maya would answer opinion-based questions, having it complete the General Social Survey (GSS) to position it in the broad social and ideological space.

The GSS is an annual survey on American adults' opinions, attitudes and behaviors. Conducted since 1972, the GSS helps monitor and explain normative trends in the United States.

Martin plotted out ChatGPT's responses along with answers from real people who participated in the 2021 GSS. Comparatively, ChatGPT is much like people with more education and who are more likely to move their residence, and unlike people without much education and who remained in their hometowns. ChatGPT's answers also aligned with more liberal people on religion.

While this was not included in his analysis as it required more creative questioning for ChatGPT to answer, Martin found that the chatbot conceded that Maya would have voted for Hillary Clinton in the 2016 election.

"Whether Maya is ChatGPT's alter ego, or its conception of its creator, the fact that this is who fundamentally illustrates the values ChatGPT holds is a wonderful piece of what we can call anecdata," Martin said. "Still the reason that these results are significant is not that they show that ChatGPT 'is' liberal, but that ChatGPT can answer these questions—which it would normally try to avoid—because it connects values with incontestable goodness, and, as such, can take positions on values."

"ChatGPT tries to be apolitical, but it works with the idea of values, which means that it necessarily bleeds over into politics. We can't make AI 'ethical' without taking political stands, and 'values' are less inherent moral principles than they are abstract ways of defending political positions."

More information: John Levi Martin, The Ethico-Political Universe of ChatGPT, *Journal of Social Computing* (2023). DOI: 10.23919/JSC.2023.0003

Provided by Tsinghua University Press