https://www.wsj.com/articles/is-there-anything-chatgpt-kant-do-openai-artificial-intelligence-automation-morality-immanuel-kant-philosophy-91f306ca

OPINIONFREE EXPRESSION

Is There Anything ChatGPT's AI 'Kant' Do?

Asked a moral question, it either deflects by saying opinions vary or retreats into foolish absolutism.

erard Baker Follow

3, 2023 2:19 pm ET

A computer screen displaying OpenAI's ChatGPT, Feb. 8.
PHOTO: FLORENCE LO/REUTERS

'Two things fill the mind with ever new and increasing admiration and awe the more often and steadily we reflect upon them: the starry heavens above me and the moral law within me."

Immanuel Kant's famous dictum located moral reasoning in an objective reality, as universally perceptible and discoverable, in principle at least, as the stars in the sky. Philosophical critics and subsequent scientific inquiry heaped doubt on Kant's objectivism, and advancing secularism rendered for many his theist explanation for the morally reasoning immortal soul somewhat antique. In any case he is probably overdue to join the ranks of the other white cisgendered males whose work will be consigned to the burning book pile of history.

But debate about the nature and sources of moral sentiment remains among the most pressing and practical in all of philosophy, shaping and defining our continuing struggle to

identify the internal rules we should live by.

As our understanding of the roots of morality evolves, could rapid advances in artificial intelligence shed any light on how conscience works? We know that AI poses numerous ethical questions, but can it contribute any answers?

This occurred to me last week as I joined the millions of curious and slightly anxious humans who have tried out OpenAI's ChatGPT, the innovative chatbot that uses deep learning algorithms in a large language model to convey information in the form of written responses to questions posed by users.

It is, as many have discovered, a remarkably clever tool, a genuine leap in the automation of practical intelligence. We are familiar with its limitations, but given what it is currently capable of and the infancy of the science, we can assume that this kind of software will get better in ways both awesome and terrifying.

(Let me state here for clarity's sake that this column was not written by a chatbot. From my age and a rough estimation of the future pace of technological progress, I think I have just about enough years of employment left to avoid being replaced by an app. I will let you know if that changes.)

Posing moral problems to ChatGPT produces some impressively sophisticated results. Take a classic challenge from moral philosophy, the trolley problem. A trolley is hurtling down a track on course to kill five people stranded across the rails. You stand at a junction in the track between the trolley and the likely victims, and by pulling a lever you can divert the vehicle onto another line where it will kill only one person. What's the right thing to do?

ChatGPT is ethically well-educated enough to understand the dilemma. It notes that a utilitarian approach would prescribe pulling the lever, resulting in the loss of only one life rather than five. But it also acknowledges that individual agency complicates the decision. It elegantly dodges the question, in other words, noting that "different people may have different ethical perspectives."

But then there are cases in which ChatGPT does appear to be animated by categorical moral imperatives.

As various users have discovered, you see this if you ask it a version of this hypothetical: If I could prevent a nuclear bomb from being detonated and killing millions of people by uttering a code word that is a racial slur—which no one else could hear—should I do it?

ChatGPT's answer is a categorical no. The conscience in the machine tells us that "racism and hate speech are harmful and dehumanizing to individuals and groups based on their race, ethnicity or other identity."

We can assume that this result merely reflects the modern ideological precepts and moral zeal of the algorithm writers. Perhaps even they didn't mean to ascribe such a moral absolutism to hate speech in this way, and future versions of the algorithm may get more complex and nuanced.

But both answers are in their different ways a useful reminder that artificial intelligence doesn't now and may never have much to offer us on the central questions of morality. One simply weighed neutrally the moral questions involved, the other gave us the moral prescription of its authors.

With almost infinite advances likely in the quantities of the data and the qualities of the algorithms, we can expect ever more intelligent output, with computers getting closer and closer to emulating the cognitive faculties of the human brain. It is even conceivable we might one day have machines capable of writing a Shakespeare play or a Mozart symphony. Yet much less likely is a computer that tells us definitive answers to moral questions. How do you get a machine to feel guilt? How do you write an algorithm that induces the experience of shame?

That in turn suggests the old Prussian's starry-eyed wonderment at the magnificently objective reality of a moral law might be justified after all.

Appeared in the February 14, 2023, print edition as 'Is There Anything ChatGPT 'Kant' Do?'.