KU News

HOME / NEWS

# AI can imitate morality without actually possessing it, new philosophy study finds

A new paper shows that although AI does not currently have practical judgment, it has a functionally equivalent mechanism — transformer models — which can allow it to form maxims that consider morally salient facts. Image credit: Adobe Stock

LAWRENCE — Can artificial intelligence truly exhibit morality?

Currently, the field of AI ethics is debating whether these machines should be treated as moral agents capable of making ethical decisions.

“We are now asking, ‘If these systems can act like human beings who are moral agents, then maybe these systems are moral agents,’” said Oluwaseun Damilola Sanwoolu, a doctoral candidate in philosophy and a graduate research assistant at the Center for Cyber-Social Dynamics (CCSD) at the University of Kansas.

“I think that logic is so far stretched because there are certain things tied to us as human beings that AI systems don’t have. Like they can act like one, they can imitate one, but that doesn’t necessarily make them moral agents.”

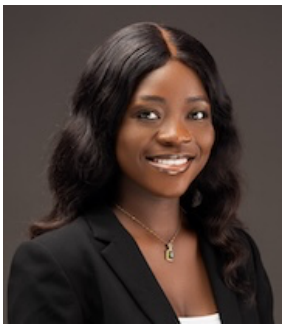
FRI, 08/22/2025

Jon Niccum

MEDIA CONTACTS

Jon Niccum
KU News Service
785-864-7633
jniccum@ku.edu

Such questions are addressed in her new paper titled "Kantian deontology for AI: alignment without moral agency." Although AI does not currently have practical judgment, Sanwoolu shows it has a functionally equivalent mechanism — transformer models — which can allow it to form maxims that consider morally salient facts. Thus, this supports the claim that AI alignment is possible within a Kantian framework.



Oluwaseun Damilola Sanwoolu



The research appears in the journal AI and Ethics.

Kantian deontology, developed by German philosopher Immanuel Kant (1724-1804), is an ethical theory focusing on moral duties and principles, rather than the consequences of actions, to determine right or wrong. It emphasizes the use of reason to identify universal moral laws.

"Initially, I was never a Kantian scholar. I didn't even really like the guy. But as I started thinking about how we can ethically shape technology, which is one of our research goals at CCSD, I found myself reading more Kant. That got me thinking: If Kant's principles were plausible for humans, then maybe they can work for AI systems, too."

Sanwoolu addresses two major objections to applying Kant's moral philosophy to AI. The first objection is AI cannot fulfill Kant's standards for moral agency. The second is that Kant's theory doesn't account for context when the principles of his theory are applied.

"In terms of the first objection, I concede AI systems are nonmoral agents. But is this still possible for us to have them behave in ways that would mimic a human agent using the Kantian system without they themselves being moral agents?' I think that's doable," she said.

For comparison, she notes that when teaching the value of honesty to children, adults don't necessarily tell them, "This is the moral code we live by." We simply model behaviors of honesty to them. Even though they are not yet fully formed moral agents, they can see what people do and imitate that.

"AI systems can behave in the same way," she said. "They don't have to be moral agents themselves."

Secondly, Kant's theory is very rigorous and doesn't take context into account.

"Kant would say something like, 'Do not break a promise.' And people push back on that because what if you made the promise under duress? But some Kantian scholars argue that Kant wasn't blind to context, and they point to practical judgment to account for it. Interestingly, transformer models in AI already play a similar functional role: They are designed to be sensitive to context. So, even if AI systems aren't moral agents, they might still approximate the kind of context-sensitive reasoning we see in human moral deliberation," she said.

Ultimately, she supposes AI can behave like someone who wants to achieve morality or someone who wants to carry out a morally correct action.

“People want to ensure AI does no harm. What does that mean?” she said. “Is an AI system going to be harmful or helpful if it assists a person in committing suicide? That’s where ethical systems and ethical frameworks come in play, because then it’s not just telling you, ‘Do this.’ It’s telling you, ‘Do this because.’”

A native of Nigeria, Sanwoolu is now in her fourth year at KU. Her research focuses on the ethics of AI and the philosophy of technology, and it is supervised by John Symons, KU professor of philosophy and director of the Center for Cyber-Social dynamics.

“Ethics gives context to the things we want AI systems to do,” she said. “Maybe we don’t have consensus on what the one correct ethical theory is, but I think this framework can actually work for AI alignment.”