Friendly AI may backfire when its tone doesn't match the moral dilemma

June 2 2026, by Sanjukta Mondal

Credit: Image generated by the editorial team using AI for illustrative purposes.

AI chatbots have become friends, confidants, even professional and health advisors for many people around the world. While the long-term consequences remain debated, it has become an undeniable reality of the

ChatGPT era. Although people prefer AI for its analytical capabilities, not every recommendation from the chatbot will be supported by data and facts; some might require moral intelligence.

A recent study by researchers at Shanghai Jiao Tong University, China, has investigated the factors influencing user trust in AI chatbots in ethically sensitive situations. It focused on two major cues: the conversation style adopted by the chatbot to convey the message and that moral stance. The findings are published in *Computers in Human Behavior*.

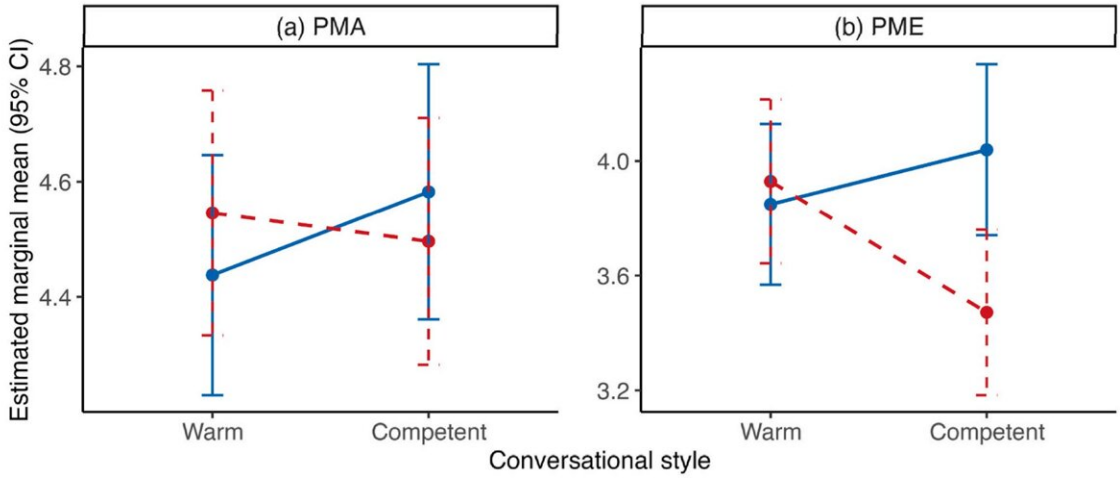
The researchers found that there was no one-size-fits-all solution that shifted user perception of AI's ability to make decisions with moral consequences. Simply making a chatbot sound friendly or professional did not automatically make people trust it more, the context mattered more than the design choice. People preferred a warm, friendly tone when the AI followed strict moral rules or in low-stakes situations. However, when the AI took a utilitarian approach focused on the greater good, or when the stakes were higher, people leaned toward a more professional and competent tone with careful reasoning.

To trust or not to trust depends on the situation

AI is no longer limited to helping someone with their school homework or drafting emails, it is slowly moving into high-risk areas like health care and safety, where the queries it encounters will involve moral dilemmas, and the solution provided can have ethical consequences. In such situations, users judge the AI's responses not just by its accuracy, but also by its moral decision-making and trustworthiness.

Previous research mostly compared humans to AI without closely examining how the design of the AI itself shapes user trust. Researchers didn't have a clear picture of how a chatbot's personality or speaking

style, in conjunction with the logic behind its decisions, influences user trust. For this study, the researchers recruited 447 participants and divided them into eight groups to test different combinations of AI behavior and scenarios.

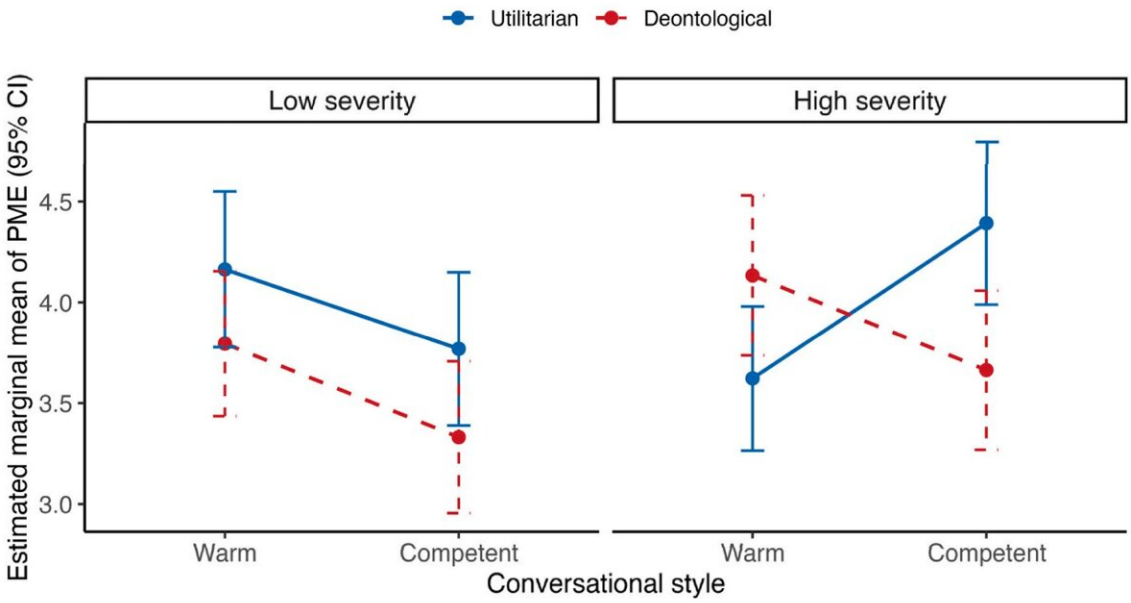


Interaction effects of conversational style and moral judgment on perceived moral agency (PMA) and perceived moral emotions (PME). Credit: *Computers in Human Behavior* (2026). DOI: 10.1016/j.chb.2026.109039

The experiments were conducted in two stages. In the first stage, participants chatted with an AI for six minutes about personal stress to assess the chatbot's personality. The chatbot adopted one of two communication styles: warm and friendly, using supportive language and emojis; or competent and professional, using a more logical, analytical tone.

In the second stage, the AI was presented with a moral dilemma similar to the classic trolley problem, where it had to choose between two

outcomes. Participants were then shown the AI's response. Depending on the experimental condition, the AI either took a utilitarian approach, choosing the option that benefited the greatest number of people; or a deontological approach, adhering to a strict moral rule such as do not cause harm. The researchers also tested how the AI responded when the stakes were varied. One scenario involved minor injuries while the other could lead to life-or-death consequences.



Three-way interaction between conversational style, moral judgment, and outcome severity on perceived moral emotions (PME). Credit: *Computers in Human Behavior* (2026). DOI: 10.1016/j.chb.2026.109039

The study did not derive a single formula for building a trustworthy AI for moral decisions. What mattered was finding the right balance between the chatbot's personality, the reasoning behind its choices, and the stakes of the situation. People trusted the AI more when its

personality matched its logic. The stakes of the situation amplified these effects further. They found that while a friendly personality helped build trust in minor, low-stakes situations, high-stakes decisions called for a professional style that emphasized accountability and careful, deliberate reasoning over friendliness.

The researchers believe that these findings provide new insight into how people assess an AI's moral judgment across different contexts. The AI chatbot's architecture should have room for it to adapt the communication style to the moral stakes and avoid one-size-fits-all approaches.

More information: Lianshan Zhang et al, Not warm or cold, but appropriate: How outcome severity shifts moral-mind inferences and trust in AI chatbots, *Computers in Human Behavior* (2026). DOI: 10.1016/j.chb.2026.109039