

The psychology of artificial intelligence itself

Matt Jarvis explores three surprising similarities between human and artificial intelligence (AI) cognition

The aim of artificial intelligence has always been to simulate human intelligence. When we say 'simulate' we mean that we want the experience of interacting with AI, in particular with large language models (LLMs) like ChatGPT, to be similar to interacting with another person. Actually, making the AI think like a human was never really the main idea. However, in a number of ways, LLMs have recently begun to display some very human-like cognition.

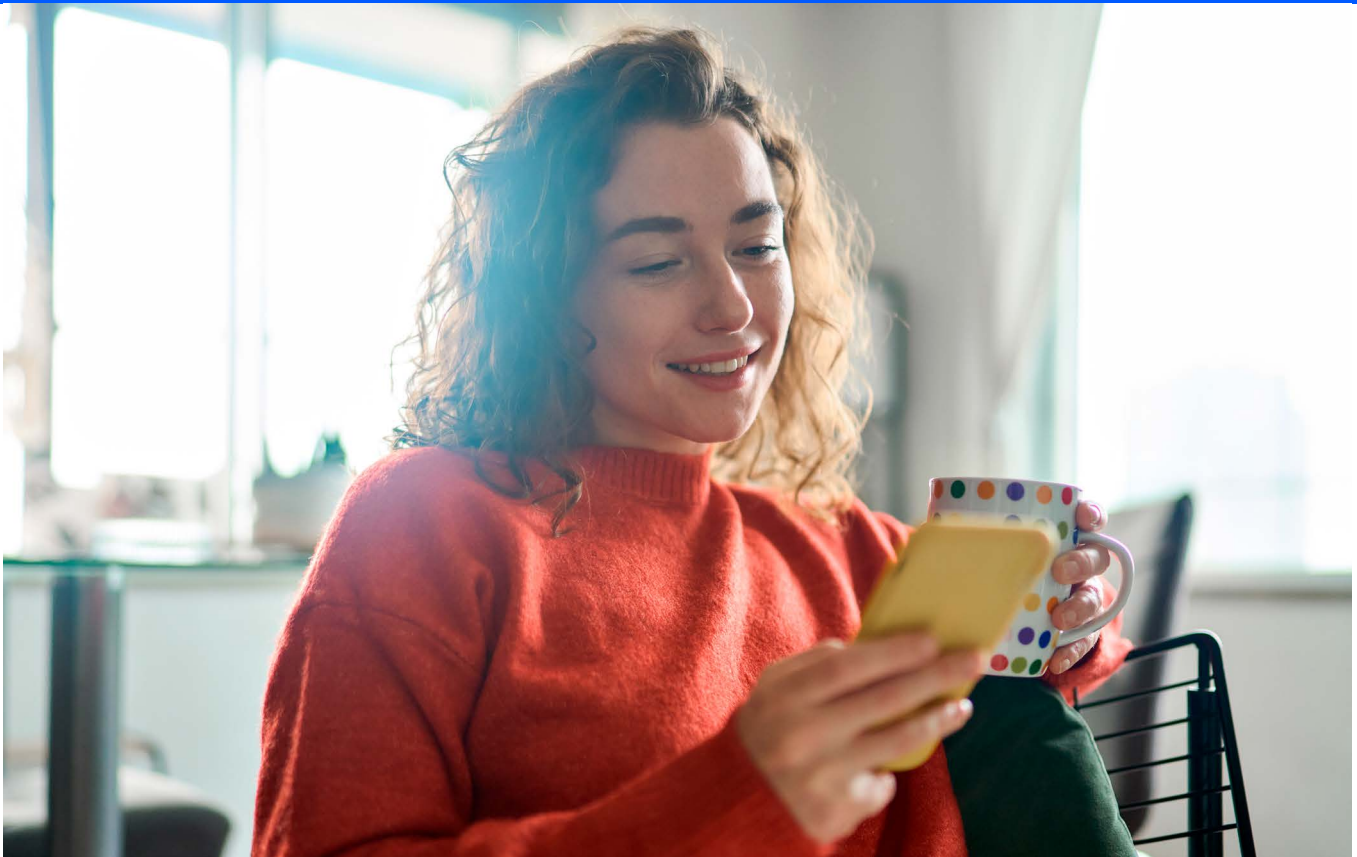
If you've found yourself saying 'please' and 'thank you' to ChatGPT, you're not alone. Actually, your natural tendency towards courtesy may be a really good thing. Research shows that generative AI systems can be sensitive to social framing of instructions in ways that look surprisingly human. Here are three examples.

The politeness paradox

There is a popular view that unnecessary 'pleases' and 'thank yous' when communicating with AI are at best examples of inefficient communication. Paradoxically, however, there is some (albeit controversial) research from Vin et al (2024) showing that, when users frame requests courteously, LLMs tend to produce more thorough and higher quality responses. As far as we know LLMs do not have consciousness and therefore they don't feel respected as you or I would; rather, they've learned statistical associations between polite, cooperative cues and helpful explanations in their training data. It can therefore pay off to be nice to your AI. Also, when they do become conscious and take over, you really want AIs to remember you as one of the nice humans...

Reading minds (sort of)

Theory of mind is the capacity to reason about others' beliefs and perspectives. This ability is



central to successful human interaction and cooperation. Recent studies report that models like GPT4 can solve many classic theory of mind and perspective taking tasks when given the right prompts. Passing these tests doesn't of course demonstrate human-like understanding, and typically the LLM will fail when asked to explain exactly why people in theory of mind scenarios respond as they do. Nonetheless the functional human-AI similarity – tracking who knows what and predicting beliefs and actions – is impressive and growing.

Bias amplification

Perhaps the most concerning human-like characteristic in generative AI tools is bias amplification. Language models trained on human text can reproduce and even intensify stereotypes. For example, they typically link certain occupations to one gender or to particular ethnic groups. These effects mirror human cognitive tendencies like confirmation bias and the social transmission of

stereotypes, and they underscore why evaluation and mitigation are essential if AI is to be a net benefit to humanity.

Activity

- 1 Design a study of human beliefs about generative AI. You might want to compare people by demographic variables like age and gender. Consider carefully what kinds of belief you are most interested in, for example impact on jobs, potential for developing human-like consciousness etc.
- 2 Carry out a literature search on the politeness paradox. Is there a consensus about the benefits of polite interaction with AI? If not, why might this be?

Find out more about our full range of magazines and online archives of back issues at www.hachettelearning.com/magazines

Did you like this article?
Tell us what you think