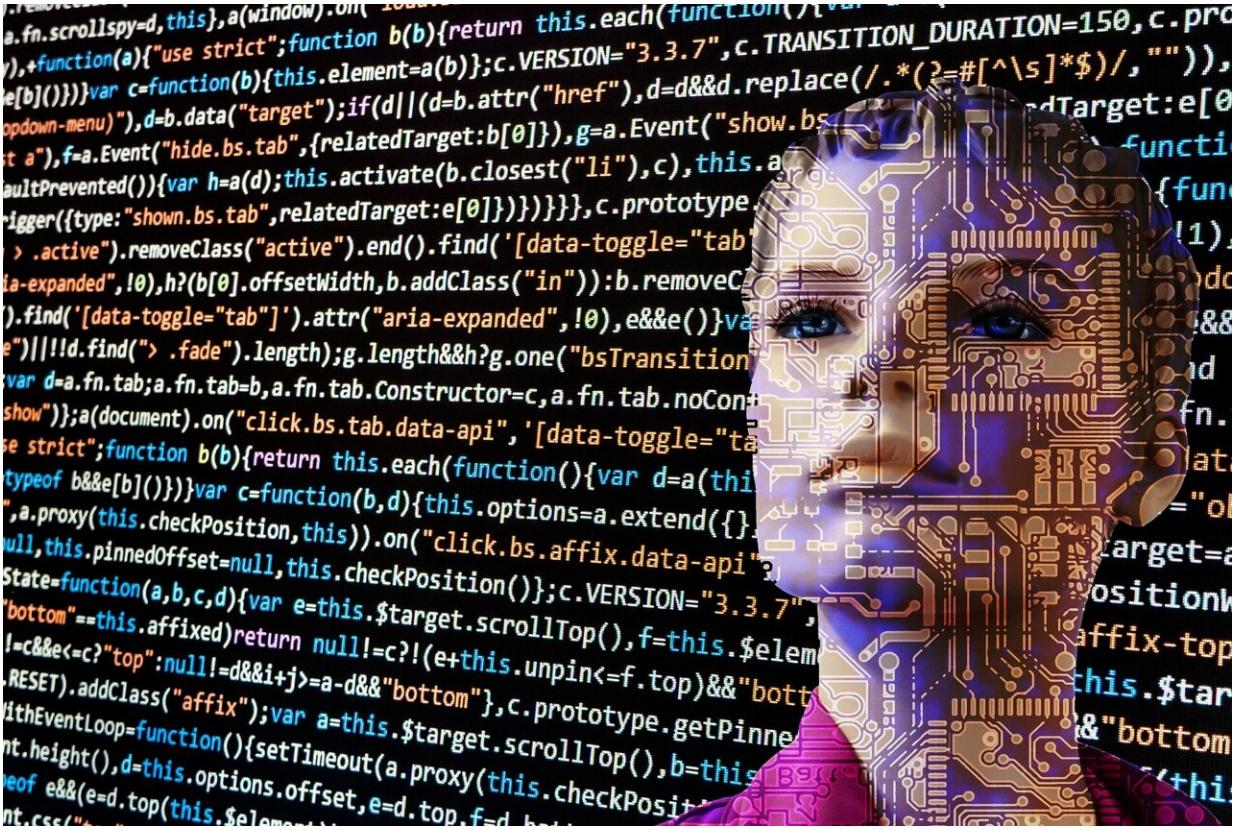
AI- or human-written language? Assumptions mislead

March 7 2023, by Tom Fleischman



Datasets used to train AI algorithms may underrepresent older people. Credit: Pixabay/CC0 Public Domain

Human assumptions regarding language usage can lead to flawed judgments of whether language was AI- or human-generated, Cornell

Tech and Stanford researchers have found in a series of experiments.

While individuals' proficiency at detecting AI-generated language was generally a tossup across the board, people were consistently influenced by the same verbal cues, leading to the same flawed judgments.

Participants could not differentiate AI-generated from human-generated language, erroneously assuming that mentions of personal experiences and the use of "I" pronouns indicated human authors. They also thought that convoluted phrasing was AI-generated.

"We learned something about humans and what they believe to be either human or AI language," said Mor Naaman, professor at the Jacobs Technion-Cornell Institute at Cornell Tech and of information science at the Cornell Ann S. Bowers College of Computing and Information Science. "But we also show that AI can take advantage of that, learn from it and then produce texts that can more easily mislead people."

Maurice Jakesch, Ph.D., a former member of Naaman's Social Technologies Lab at Cornell Tech, is lead author of "Human Heuristics for AI-Generated Language Are Flawed," published March 7 in *Proceedings of the National Academy of Sciences*. Naaman and Jeff Hancock, professor of communication at Stanford University, are co-authors.

The researchers conducted three main experiments and three more to validate the findings, involving 4,600 participants and 7,600 "verbal self-presentations"—profile text people used to describe themselves on social websites. The experiments were patterned after the Turing test, developed in 1950 by British mathematician Alan Turing, who devised the test to measure a machine's ability to exhibit intelligent behavior equal to or better than a human.

Instead of testing the machine, the new study tested humans' ability to detect whether the exhibited intelligence came from a machine or a human. The researchers trained multiple AI language models to generate text in three social contexts where trust in the sender is important: professional (job application); romantic (online dating); and hospitality (Airbnb host profiles).

In the three main experiments, using two different language models, participants identified the source of a self-presentation with only 50% to 52% accuracy. But the responses, the researchers discovered, were not random, as the agreement between respondents' answers was significantly higher than chance, meaning many participants were drawing the same flawed conclusions.

The researchers conducted an analysis of the heuristics (the process by which a conclusion is reached) participants used in deciding whether language was AI- or human-generated, first by asking participants to explain their judgments, then following up with a computational analysis that confirmed these reports. People cited mentions of family and life experiences, as well as the use of first-person pronouns, as evidence of human language.

However, such language is equally likely to be produced by AI language models.

"People's intuition goes counter the current design of these language models," Naaman said. "They produce text that is statistically probable—in other words, language that is common. But people tended to associate uncommon language with AI, a behavior that AI systems can then exploit that to create language, as we call it, 'more human than human.'"

In three pre-registered validation experiments, the authors show that

indeed, AI can exploit people's heuristics to produce text that people more reliably rate as human-written than actual human-written text.

People's reliance on flawed heuristics in identifying AI-generated language, the authors wrote, is not necessarily indicative of increased machine intelligence. It doesn't take superior intelligence, they said, to "fool" humans—just a well-placed personal pronoun, or a mention of family.

The authors note that while humans' ability to discern AI-generated language might be limited, language models that are "self-disclosing by design" would let the user know that the information is not human-generated while preserving the integrity of the message.

This could be achieved either by a language that is clearly non-human (avoiding the use of informal speech) or through "AI accents"—a dedicated dialect that could "facilitate and support people's intuitive judgments without interrupting the flow of communication," they wrote.

Hancock, a faculty member at Cornell from 2002-15, said this work is "one of the last nails in the coffin" of the Turing test era.

"As a way of thinking about whether something's intelligent or not," he said, "our data pretty clearly show that in pretty important ways of being human—that is, describing yourself professionally, romantically or as a host—it's over. The machine has passed that test."

Naaman said this work—particularly relevant with the arrival of AI tools such as ChatGPT—highlights the fact that AI will increasingly be used as a tool to facilitate human-to-human communication.

"This is about not about us talking to AI. It's us talking to each other through AI," he said. "And the implications that we show on trust are

significant: People will be easily misled and will easily distrust each other—not AI."

More information: Maurice Jakesch et al, Human heuristics for AI-generated language are flawed, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2208839120