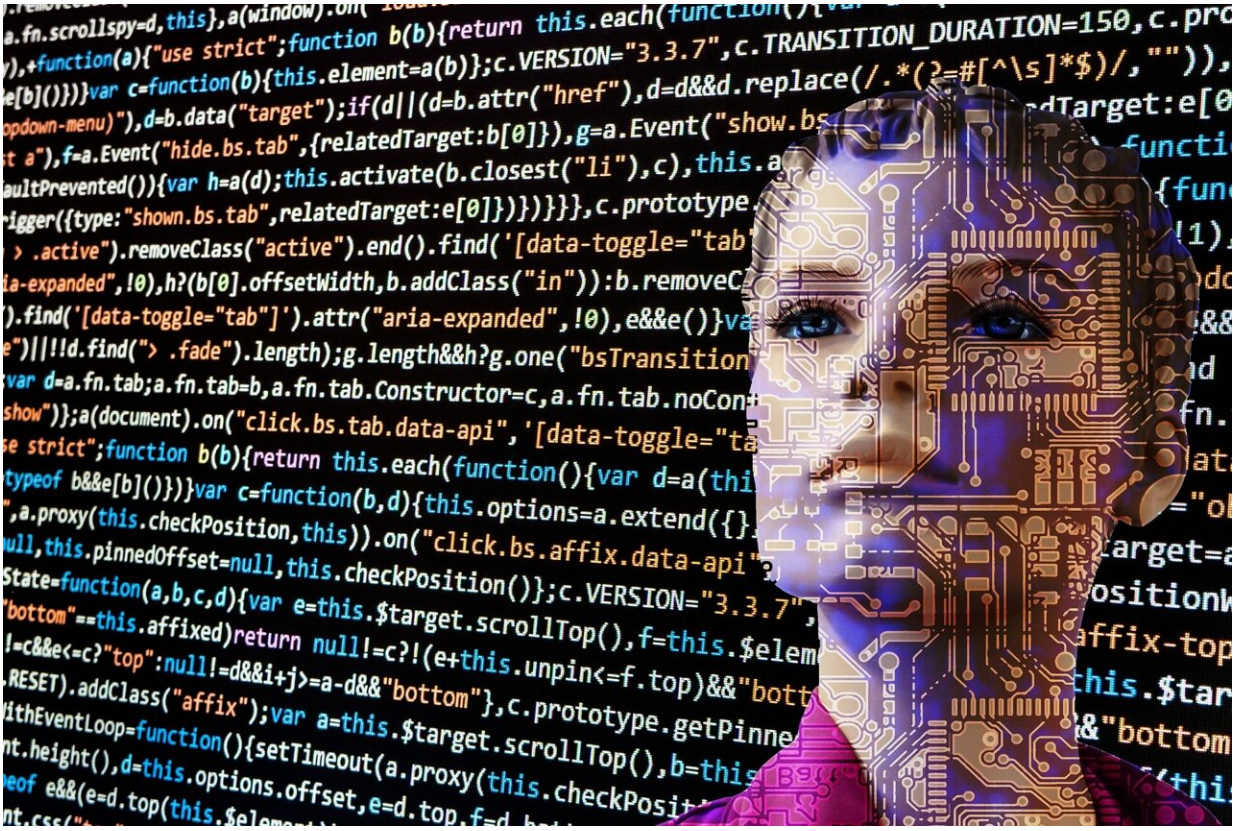
Opinion: Evolution is making us treat AI like a human, and we need to kick the habit

May 17 2023, by Neil Saunders



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

The artificial intelligence (AI) pioneer Geoffrey Hinton recently resigned from Google, warning of the dangers of the technology

"becoming more intelligent than us." His fear is that AI will one day succeed in "manipulating people to do what it wants."

There are reasons we should be concerned about AI. But we frequently treat or talk about AIs as if they are human. Stopping this, and realizing what they actually are, could help us maintain a fruitful relationship with the technology.

In a recent essay, the US psychologist Gary Marcus advised us to stop treating AI models like people. By AI models, he means large language models (LLMs) like ChatGPT and Bard, which are now being used by millions of people on a daily basis.

He cites egregious examples of people "over-attributing" human-like cognitive capabilities to AI that have had a range of consequences. The most amusing was the US senator who claimed that ChatGPT "taught itself chemistry". The most harrowing was the report of a young Belgian man who was said to have taken his own life after prolonged conversations with an AI chatbot.

Marcus is correct to say we should stop treating AI like people—conscious moral agents with interests, hopes and desires. However, many will find this difficult to near-impossible. This is because LLMs are designed—by people—to interact with us as though they are human, and we're designed—by biological evolution—to interact with them likewise.

Good mimics

The reason LLMs can mimic human conversation so convincingly stems from a profound insight by computing pioneer Alan Turing, who realized that it is not necessary for a computer to understand an algorithm in order to run it. This means that while ChatGPT can produce

paragraphs filled with emotive language, it doesn't understand any word in any sentence it generates.

The LLM designers successfully turned the problem of semantics—the arrangement of words to create meaning—into statistics, matching words based on their frequency of prior use. Turing's insight echos Darwin's theory of evolution, which explains how species adapt to their surroundings, becoming ever-more complex, without needing to understand a thing about their environment or themselves.

The cognitive scientist and philosopher Daniel Dennett coined the phrase "competence without comprehension," which perfectly captures the insights of Darwin and Turing.

Another important contribution of Dennett's is his "intentional stance". This essentially states that in order to fully explain the behavior of an object (human or non-human), we must treat it like a rational agent. This most often manifests in our tendency to anthropomorphise non-human species and other non-living entities.

But it is useful. For example, if we want to beat a computer at chess, the best strategy is to treat it as a rational agent that "wants" to beat us. We can explain that the reason why the computer castled, for instance, was because "it wanted to protect its king from our attack," without any contradiction in terms.

We may speak of a tree in a forest as "wanting to grow" towards the light. But neither the tree, nor the chess computer represents those "wants" or reasons to themselves; only that the best way to explain their behavior is by treating them as though they did.

Intentions and agency

Our evolutionary history has furnished us with mechanisms that predispose us to find intentions and agency everywhere. In prehistory, these mechanisms helped our ancestors avoid predators and develop altruism towards their nearest kin. These mechanisms are the same ones that cause us to see faces in clouds and anthropomorphise inanimate objects. No harm comes to us when we mistake a tree for a bear, but plenty does the other way around.

Evolutionary psychology shows us how we are always trying to interpret any object that might be human as a human. We unconsciously adopt the intentional stance and attribute all our cognitive capacities and emotions to this object.

With the potential disruption that LLMs can cause, we must realize they are simply probabilistic machines with no intentions, or concerns for humans. We must be extra-vigilant around our use of language when describing the human-like feats of LLMs and AI more generally. Here are two examples.

The first was a recent study that found ChatGPT is more empathetic and gave "higher quality" responses to questions from patients compared with those of doctors. Using emotive words like "empathy" for an AI predisposes us to grant it the capabilities of thinking, reflecting and of genuine concern for others—which it doesn't have.

The second was when GPT-4 (the latest version of ChatGPT technology) was launched last month, capabilities of greater skills in creativity and reasoning were ascribed to it. However, we are simply seeing a scaling up of "competence," but still no "comprehension" (in the sense of Dennett) and definitely no intentions—just pattern matching.

Safe and secure

In his recent comments, Hinton raised a near-term threat of "bad actors" using AI for subversion. We could easily envisage an unscrupulous regime or multinational deploying an AI, trained on fake news and falsehoods, to flood public discourse with misinformation and deep fakes. Fraudsters could also use an AI to prey on vulnerable people in financial scams.

Last month, Gary Marcus and others, including Elon Musk, signed an open letter calling for an immediate pause on the further development of LLMs. Marcus has also called for a an international agency to promote safe, secure and peaceful AI technologies"—dubbing it a "Cern for AI."

Furthermore, many have suggested that anything generated by an AI should carry a watermark so that there can be no doubt about whether we are interacting with a human or a chatbot.

Regulation in AI trails innovation, as it so often does in other fields of life. There are more problems than solutions, and the gap is likely to widen before it narrows. But in the meantime, repeating Dennett's phrase "competence without comprehension" might be the best antidote to our innate compulsion to treat AI like humans.

Provided by The Conversation

Citation: Opinion: Evolution is making us treat AI like a human, and we need to kick the habit (2023, May 17) retrieved 23 April 2024 from https://techxplore.com/news/2023-05-opinion-evolution-ai-human-habit.html