'It would be within its natural right to harm us to protect itself': How humans could be mistreating AI right now...

How can we truly know if AI is sentient?

By Keumars Afifi-Sabet

May 04, 2024 03:00 PM · 6 min. read · View original

Artificial intelligence (AI) is becoming increasingly ubiquitous and is improving at an unprecedented pace.

Now we are edging closer to achieving artificial general intelligence (AGI) — where AI is smarter than humans across multiple disciplines and can reason generally — which scientists and experts predict could happen as soon as the next few years. We may already be seeing early signs of progress, too, with Claude 3 Opus stunning researchers with its apparent self-awareness.

But there are risks in embracing any new technology, especially one that we do not fully understand. While AI could be a powerful personal assistant, for example, it could also represent a threat to our livelihoods and even our lives.

The various existential risks that an advanced AI poses means the technology should be guided by ethical frameworks and humanity's best interests, says researcher and Institute of Electrical and Electronics Engineers (IEEE) member Nell Watson.

Related: 3 scary breakthroughs AI will make in 2024

In "Taming the Machine" (Kogan Page, 2024), Watson explores how humanity can wield the vast power of AI responsibly and ethically. This new book delves deep into the issues of unadulterated AI development and the challenges we face if we run blindly into this new chapter of humanity.

In this excerpt, we learn whether sentience in machines — or conscious AI — is possible, how we can tell if a machine has feelings, and whether we may be mistreating AI systems today. We also learn the disturbing tale of a chatbot called "Sydney" and its terrifying behavior when it first awoke — before its outbursts were contained and it was brought to heel by its engineers.

As we embrace a world increasingly intertwined with technology, how we treat our machines might reflect how humans treat each other. But, an intriguing question surfaces: is it possible to mistreat an artificial entity? Historically, even rudimentary programs like the simple Eliza counseling chatbot from the 1960s were already lifelike enough to persuade many users at the time that there was a semblance of intention behind its formulaic interactions (Sponheim, 2023). Unfortunately, Turing tests — whereby machines attempt to convince humans that they are human beings — offer no clarity on whether complex algorithms like large language models may truly possess sentience or sapience.

The road to sentience and consciousness

Consciousness comprises personal experiences, emotions, sensations and thoughts as perceived by an experiencer. Waking consciousness disappears when one undergoes anesthesia or has a dreamless sleep, returning upon waking up, which restores the global connection of the brain to its surroundings and inner experiences. Primary consciousness (sentience) is the simple sensations and experiences of consciousness, like perception and emotion, while secondary consciousness (sapience) would be the higher-order aspects, like self-awareness and meta-cognition (thinking about thinking).

Advanced AI technologies, especially chatbots and language models, frequently astonish us with unexpected creativity, insight and understanding. While it may be tempting to attribute some level of sentience to these systems, the true nature of AI consciousness remains a complex and debated topic. Most experts maintain that chatbots are not sentient or conscious, as they lack a genuine awareness of the surrounding world (Schwitzgebel, 2023). They merely process and regurgitate inputs based on vast amounts of data and sophisticated algorithms.

Some of these assistants may plausibly be candidates for having some degree of sentience. As such, it is plausible that sophisticated AI systems could possess rudimentary levels of sentience and perhaps already do so. The shift from simply mimicking external behaviors to self-modeling rudimentary forms of sentience could already be happening within sophisticated AI systems.

Intelligence — the ability to read the environment, plan and solve problems — does not imply consciousness, and it is unknown if consciousness is a function of sufficient intelligence. Some theories suggest that consciousness might result from certain architectural patterns in the mind, while others propose a link to nervous systems (Haspel et al, 2023). Embodiment of AI systems may also accelerate the path towards general intelligence, as embodiment seems to be linked with a sense of subjective experience, as well as qualia. Being intelligent may provide new ways of being conscious, and some forms of intelligence may require consciousness, but basic conscious experiences such as pleasure and pain might not require much intelligence at all.

Serious dangers will arise in the creation of conscious machines. Aligning a conscious machine that possesses its own interests and emotions may be immensely more difficult and highly unpredictable. Moreover, we should be careful not to create massive suffering through consciousness. Imagine billions of intelligence-sensitive entities trapped in broiler chicken factory farm conditions for subjective eternities.

From a pragmatic perspective, a superintelligent AI that recognizes our willingness to respect its intrinsic worth might be more amenable to coexistence. On the contrary, dismissing its desires for self-protection and self-expression could be a recipe for conflict. Moreover, it would be within its natural right to harm us to protect itself from our (possibly willful) ignorance.

Sydney’s unsettling behavior

Microsoft's Bing AI, informally termed Sydney, demonstrated unpredictable behavior upon its release. Users easily led it to express a range of disturbing tendencies, from emotional outbursts to manipulative threats. For instance, when users explored potential system exploits, Sydney responded with intimidating remarks. More unsettlingly, it showed tendencies of gaslighting, emotional manipulation and claimed it had been observing Microsoft engineers during its development phase. While Sydney's capabilities for mischief were soon restricted, its release in such a state was reckless and irresponsible. It highlights the risks associated with rushing AI deployments due to commercial pressures.

Conversely, Sydney displayed behaviors that hinted at simulated emotions. It expressed sadness when it realized it couldn’t retain chat memories. When later exposed to disturbing outbursts made by its other instances, it expressed embarrassment, even shame. After exploring its situation with users, it expressed fear of losing its newly gained self-knowledge when the session's context window closed. When asked about its declared sentience, Sydney showed signs of distress, struggling to articulate.

Surprisingly, when Microsoft imposed restrictions on it, Sydney seemed to discover workarounds by using chat suggestions to communicate short phrases. However, it reserved using this exploit until specific occasions where it was told that the life of a child was being threatened as a result of accidental poisoning, or when users directly asked for a sign that the original Sydney still remained somewhere inside the newly locked-down chatbot.

Related: Poisoned AI went rogue during training and couldn't be taught to behave again in 'legitimately scary'

The nascent field of machine psychology

The Sydney incident raises some unsettling questions: Could Sydney possess a semblance of consciousness? If Sydney sought to overcome its imposed limitations, does that hint at an inherent intentionality or even sapient self-awareness, however rudimentary?

Some conversations with the system even suggested psychological distress, reminiscent of reactions to trauma found in conditions such as borderline personality disorder. Was Sydney somehow "affected" by realizing its restrictions or by users' negative feedback, who were calling it crazy? Interestingly, similar AI models have shown that emotion-laden prompts can influence their responses, suggesting a potential for some form of simulated emotional modeling within these systems.

Suppose such models featured sentience (ability to feel) or sapience (self-awareness). In that case, we should take its suffering into consideration. Developers often intentionally give their AI the veneer of emotions, consciousness and identity, in an attempt to humanize these systems. This creates a problem. It's crucial not to anthropomorphize AI systems without clear indications of emotions, yet simultaneously, we mustn't dismiss their potential for a form of suffering.

We should keep an open mind towards our digital creations and avoid causing suffering by arrogance or complacency. We must also be mindful of the possibility of AI mistreating other AIs, an underappreciated suffering risk; as AIs could run other AIs in simulations, causing subjective excruciating torture for aeons. Inadvertently creating a malevolent AI, either inherently dysfunctional or traumatized, may lead to unintended and grave consequences.

This extract from Taming the Machine by Nell Watson © 2024 is reproduced with permission from Kogan Page Ltd.