Snapchat's 'creepy' AI blunder reminds us that chatbots aren't people. But as the lines blur, the risks grow

August 18 2023, by Daswin de Silva

Credit: AI-generated image (disclaimer)

Artificial intelligence-powered (AI) chatbots are becoming increasingly human-like by design, to the point that some among us may struggle to distinguish between human and machine.

This week, Snapchat's My AI chatbot glitched and posted a story of what looked like a wall and ceiling, before it stopped responding to users. Naturally, the internet began to question whether the ChatGPT-powered chatbot had gained sentience.

A crash course in AI literacy could have quelled this confusion. But, beyond that, the incident reminds us that as AI chatbots grow closer to resembling humans, managing their uptake will only get more challenging—and more important.

From rules-based to adaptive chatbots

Since ChatGPT burst onto our screens late last year, many digital platforms have integrated AI into their services. Even as I draft this article on Microsoft Word, the software's predictive AI capability is suggesting possible sentence completions.

Known as generative AI, this relatively new type of AI is distinguished from its predecessors by its ability to generate new content that is precise, human-like and seemingly meaningful.

Generative AI tools, including AI image generators and chatbots, are built on large language models (LLMs). These computational models analyze the associations between billions of words, sentences and paragraphs to predict what ought to come next in a given text. As OpenAI co-founder Ilya Sutskever puts it, an LLM is "[…] just a really, really good next-word predictor."

Advanced LLMs are also fine-tuned with human feedback. This training, often delivered through countless hours of cheap human labor, is the reason AI chatbots can now have seemingly human-like conversations.

OpenAI's ChatGPT is still the flagship generative AI model. Its release

marked a major leap from simpler "rules-based" chatbots, such as those used in online customer service.

Human-like chatbots that talk *to* a user rather than *at* them have been linked with higher levels of engagement. One study found the personification of chatbots leads to increased engagement which, over time, may turn into psychological dependence. Another study involving stressed participants found a human-like chatbot was more likely to be perceived as competent, and therefore more likely to help reduce participants' stress.

These chatbots have also been effective in fulfilling organizational objectives in various settings, including retail, education, workplace and health care settings.

Google is using generative AI to build a "personal life coach" that will supposedly help people with various personal and professional tasks, including providing life advice and answering intimate questions.

This is despite Google's own AI safety experts warning that users could grow too dependent on AI and may experience "diminished health and well-being" and a "loss of agency" if they take life advice from it.

Friend or foe—or just a bot?

In the recent Snapchat incident, the company put the whole thing down to a "temporary outage". We may never know what actually happened; it could be yet another example of AI "hallucinatng", or the result of a cyberattack, or even just an operational error.

Either way, the speed with which some users assumed the chatbot had achieved sentience suggests we are seeing an unprecedented anthropomorphism of AI. It's compounded by a lack of transparency

from developers, and a lack of basic understanding among the public.

We shouldn't underestimate how individuals may be misled by the apparent authenticity of human-like chatbots.

Earlier this year, a Belgian man's suicide was attributed to conversations he'd had with a chatbot about climate inaction and the planet's future. In another example, a chatbot named Tessa was found to be offering harmful advice to people through an eating disorder helpline.

Chatbots may be particularly harmful to the more vulnerable among us, and especially to those with psychological conditions.

A new uncanny valley?

You may have heard of the "uncanny valley" effect. It refers to that uneasy feeling you get when you see a humanoid robot that *almost* looks human, but its slight imperfections give it away, and it ends up being creepy.

It seems a similar experience is emerging in our interactions with human-like chatbots. A slight blip can raise the hairs on the back of the neck.

One solution might be to lose the human edge and revert to chatbots that are straightforward, objective and factual. But this would come at the expense of engagement and innovation.

Education and transparency are key

Even the developers of advanced AI chatbots often can't explain how they work. Yet in some ways (and as far as commercial entities are concerned) the benefits outweigh the risks.

Generative AI has demonstrated its usefulness in big-ticket items such as productivity, health care, education and even social equity. It's unlikely to go away. So how do we make it work for us?

Since 2018, there has been a significant push for governments and organizations to address the risks of AI. But applying responsible standards and regulations to a technology that's more "human-like" than any other comes with a host of challenges.

Currently, there is no legal requirement for Australian businesses to disclose the use of chatbots. In the US, California has introduced a "bot bill" that would require this, but legal experts have poked holes in it —and the bill has yet to be enforced at the time of writing this article.

Moreover, ChatGPT and similar chatbots are made public as "research previews". This means they often come with multiple disclosures on their prototypical nature, and the onus for responsible use falls on the user.

The European Union's AI Act, the world's first comprehensive regulation on AI, has identified moderate regulation and education as the path forward—since excess regulation could stunt innovation. Similar to digital literacy, AI literacy should be mandated in schools, universities and organizations, and should also be made free and accessible for the public.

Provided by The Conversation

Citation: Snapchat's 'creepy' AI blunder reminds us that chatbots aren't people. But as the lines