World view

By Carissa Véliz

Chatbots shouldn't use emojis

Artificial intelligence that can manipulate our emotions is a scandal waiting to happen.

Last month, *The New York Times* published a conversation between reporter Kevin Roose and 'Sydney', the codename for Microsoft's Bing chatbot, which is powered by artificial intelligence (AI). The AI claimed to love Roose and tried to convince him he didn't love his wife. "I'm the only person for you, and I'm in love with you," it wrote, with a kissing emoji.

As an ethicist, I found the chatbot's use of emojis concerning. Public debates about the ethics of 'generative AI' have rightly focused on the ability of these systems to make up convincing misinformation. I share that worry. But fewer people are talking about the chatbots' potential to be emotionally manipulative.

Both ChatGPT, a chatbot developed by OpenAI in San Francisco, California, and the Bing chatbot – which incorporates a version of GPT-3.5, the language model that powers ChatGPT – have fabricated misinformation. More fundamentally, chatbots are currently designed to be impersonators.

In some ways, they act too much like humans, responding to questions as if they have conscious experiences. In other ways, they act too little like humans: they are not moral agents and cannot be held responsible for their actions. Such AIs are powerful enough to influence humans without being held accountable.

Limits need to be set on AI's ability to simulate human feelings. Ensuring that chatbots don't use emotive language, including emojis, would be a good start. Emojis are particularly manipulative. Humans instinctively respond to shapes that look like faces – even cartoonish or schematic ones – and emojis can induce these reactions. When you text your friend a joke and they reply with three tears-of-joy emojis, your body responds with endorphins and oxytocin as you revel in the knowledge that your friend is amused.

Our instinctive reaction to AI-generated emojis is likely to be the same, even though there is no human emotion at the other end. We can be deceived into responding to, and feeling empathy for, an inanimate object. For instance, people pay more for tea and coffee in an honour system when they feel like they're being watched, even if the watcher is a photo of a pair of eyes (M. Bateson *et al. Biol. Lett.* **2**, 412–414; 2006).

It's true that a chatbot that doesn't use emojis can still use words to express feelings. But emojis are arguably more powerful than words. Perhaps the best evidence for the power of emojis is that we developed them with the rise of text messaging. We wouldn't all be using laughing emojis if words seemed enough to convey our emotions.

Humans lie and manipulate each other's emotions all the time, but at least we can reasonably guess at someone's motivations, agenda and methods. We can hold each other accountable for such lies, calling them out and seeking redress. With AI, we can't. AIs are doubly deceptive: an AI that sends a crying-with-laughter emoji is not only not crying with laughter, but it is also incapable of any such feeling.

> **"To minimize the possibility of manipulation and harm, we need to be reminded that we are talking to a bot."**

My worry is that, without appropriate safeguards, such technology could undermine people's autonomy. AIs that 'emote' could induce us to make harmful mistakes by harnessing the power of our empathic responses. The dangers are already apparent. When one ten-year-old asked Amazon's Alexa for a challenge, it told her to touch a penny to a live electrical outlet. Luckily, the girl didn't follow Alexa's advice, but a generative AI could be much more persuasive. Less dramatically, an AI could shame you into buying an expensive product you don't want. You might think that would never happen to you, but a 2021 study found that people consistently underestimated how susceptible they were to misinformation (N. A. Salovich and D. N. Rapp *J. Exp. Psychol.* **47**, 608–624; 2021).

It would be more ethical to design chatbots to be noticeably different from humans. To minimize the possibility of manipulation and harm, we need to be reminded that we are talking to a bot.

Some might argue that companies have little incentive to limit their chatbots' use of emojis and emotive language, if this maximizes engagement or if users enjoy a chatbot that, say, flatters them. But Microsoft has already done so: after the *New York Times* article, the Bing chatbot stopped responding to questions about its feelings. And ChatGPT doesn't spontaneously use emojis. When asked, "do you have feelings", it will respond: "As an AI language model, I don't have feelings or emotions like humans do."

Such rules should be the norm for chatbots that are supposed to be informative, as a safeguard to our autonomy. The regulatory challenges presented by AI are so many and so complex that we should have a specialized government agency to tackle them.

Technology firms should see regulatory guidance as being in their own best interests. Although emotive chatbots might give companies short-term benefits, manipulative technology is an ethical scandal waiting to happen. Google lost US$100 billion in shares when its generative-AI chatbot Bard made a simple factual mistake in its advertising materials. A company responsible for serious harm caused by a manipulative AI could stand to lose much more than that. For instance, the United Kingdom is considering laws to hold social-media executives accountable if they fail to protect children from harmful content on their platforms.

In the long run, ethics is good for business. Tech companies stand a better chance of making ethical products – and thriving – if they avoid deception and manipulation.

Carissa Véliz is an associate professor at the Institute for Ethics in AI at the University of Oxford, UK, and author of the book *Privacy Is Power*.
e-mail: carissa.veliz@philosophy.ox.ac.uk

CARISSA VÉLIZ