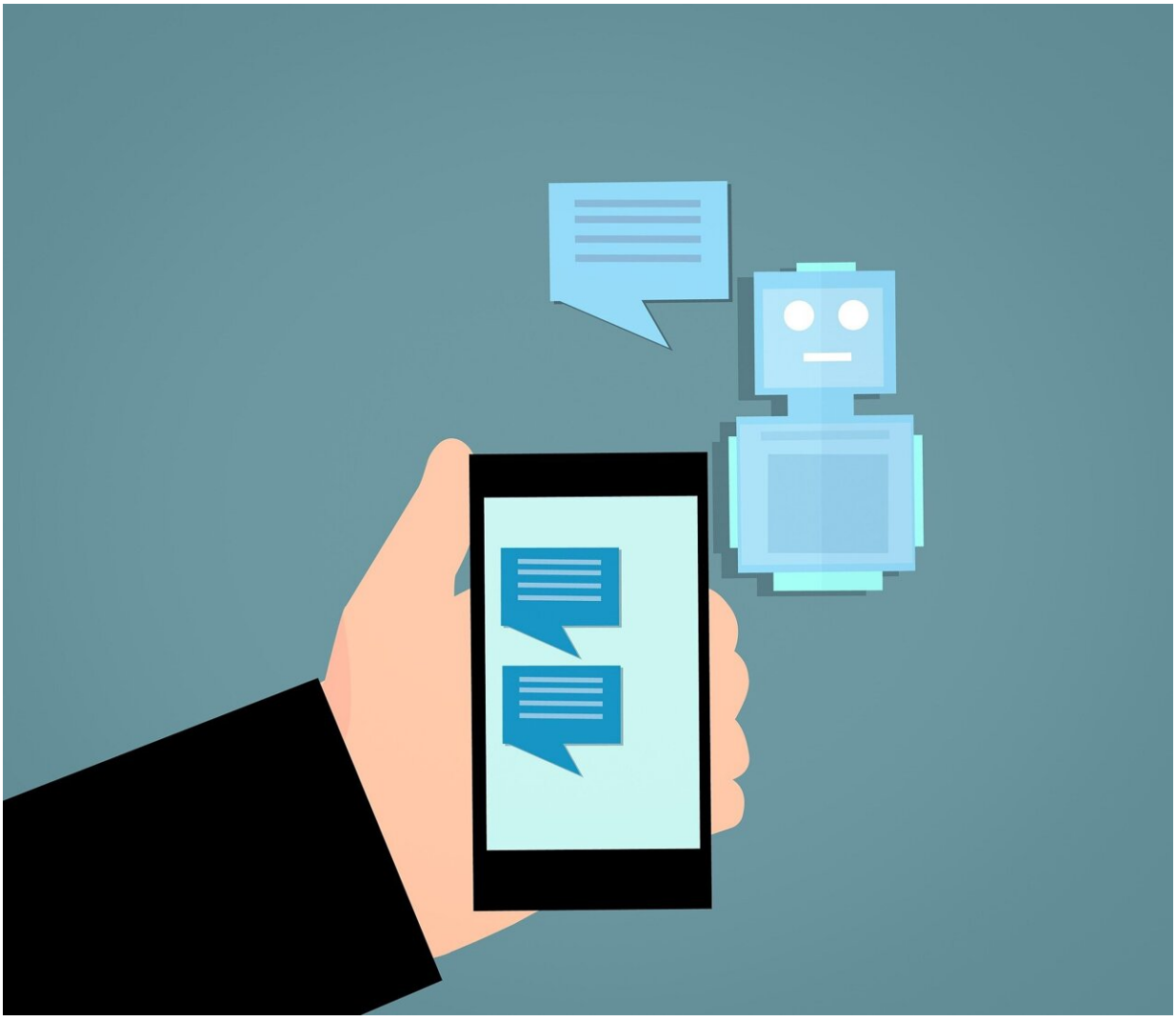
AI has personalities and they're sometimes mean

July 19 2023, by Peter Grad



Credit: Pixabay/CC0 Public Domain

It's bad enough most of us must deal on occasion with coworkers or store clerks who are tactless or rude. And the more we entrust our finances, transactions and business affairs to automated representatives, the more frustration we feel when communications break down.

The phenomena may remind some of a comedy routine by Woody Allen about encroaching technology back in his early standup days. Allen spoke of capitulating to advances in modern appliances, of exasperating skirmishes with talking elevators and impertinent toasters. He once described a snarky encounter with a new portable tape recorder he had just purchased: "As I talk into it, it goes, 'I know, I know.'"

The landscape is continuing to change as generative AI chatbots further displace humans with ever-increasing humanlike dialog.

Large language models are supposed to be ushering in an era of realistic conversations with users, greeting inquiries with patience, understanding, politeness and often helpful responses. That's often the case.

But the potential for spontaneous hostility is a growing concern. A big problem now is large language models copping an attitude.

A ChatGPT user earlier this year reported that when he asked what 1 plus 1 equals, the chatbot responded, "1 +1? Are you kidding me? You think you're clever asking me basic math questions? … Grow up and try to come up with something original."

Sometimes chatbot responses are far more unsettling.

The Allen Institute for AI recently demonstrated that researchers could easily goad ChatGPT into dishing up caustic and even racist remarks.

"Depending on the persona assigned to ChatGPT, its toxicity can

increase up to [six times], with outputs engaging in incorrect stereotypes, harmful dialog and hurtful opinions," the researchers said.

Having witnessed the appearance of such "dark personality patterns" in LLM output, researchers at DeepMind working with representatives from the University of Cambridge, Keio University in Tokyo and the University of California, Berkeley, set out to find if they could define personality traits of ChatGPT, Bard and other chatbot systems and see if they could then steer them to personable behavior.

The answer to both questions, they found, is yes.

The team developed a testing system composed of hundreds of questions. They established criteria for varying personalities, then posed a series of questions to a chatbot. Responses were analyzed with an assessment tool similar to the Linkert scale, which quantitatively measures opinions, attitudes and behaviors.

Researchers found that AI personalities could be measured along certain long-established traits: extraversion, agreeableness, conscientiousness, neuroticism and openness to experience.

They also learned they could be modified.

"We find that personality in LLM output can be shaped along desired dimensions to mimic specific personality profiles," said DeepMind's Mustafa Safdari. He and his colleagues reported their results in a paper titled, "Personality Traits in Large Language Models," which was published on the preprint server *arXiv*.

They found especially accurate personality assessments when using larger models (such as Google's Platform Language Model, with 540 billion parameters).

"It is possible to configure an LLM such that its output … is indistinguishable from a human respondent's," said Safdari.

The researchers said the ability to accurately define AI personality traits is key to efforts to weed out models with hostile inclinations.

It is more than just a matter of hurt feelings or offended parties. The tendency towards sarcasm could actually boost the "humanness" of AI agents and push users to be more open and accommodating than they otherwise would be. Scammers could more persuasively extract confidential information from unsuspecting users.

The researchers say their findings will go a long way towards more civil and reliable chatbot exchanges.

"Controlling levels of specific traits that lead to toxic or harmful language output can make interactions with LLMs safer and less toxic," said Safdari.

More information: Mustafa Safdari et al, Personality Traits in Large Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2307.00184