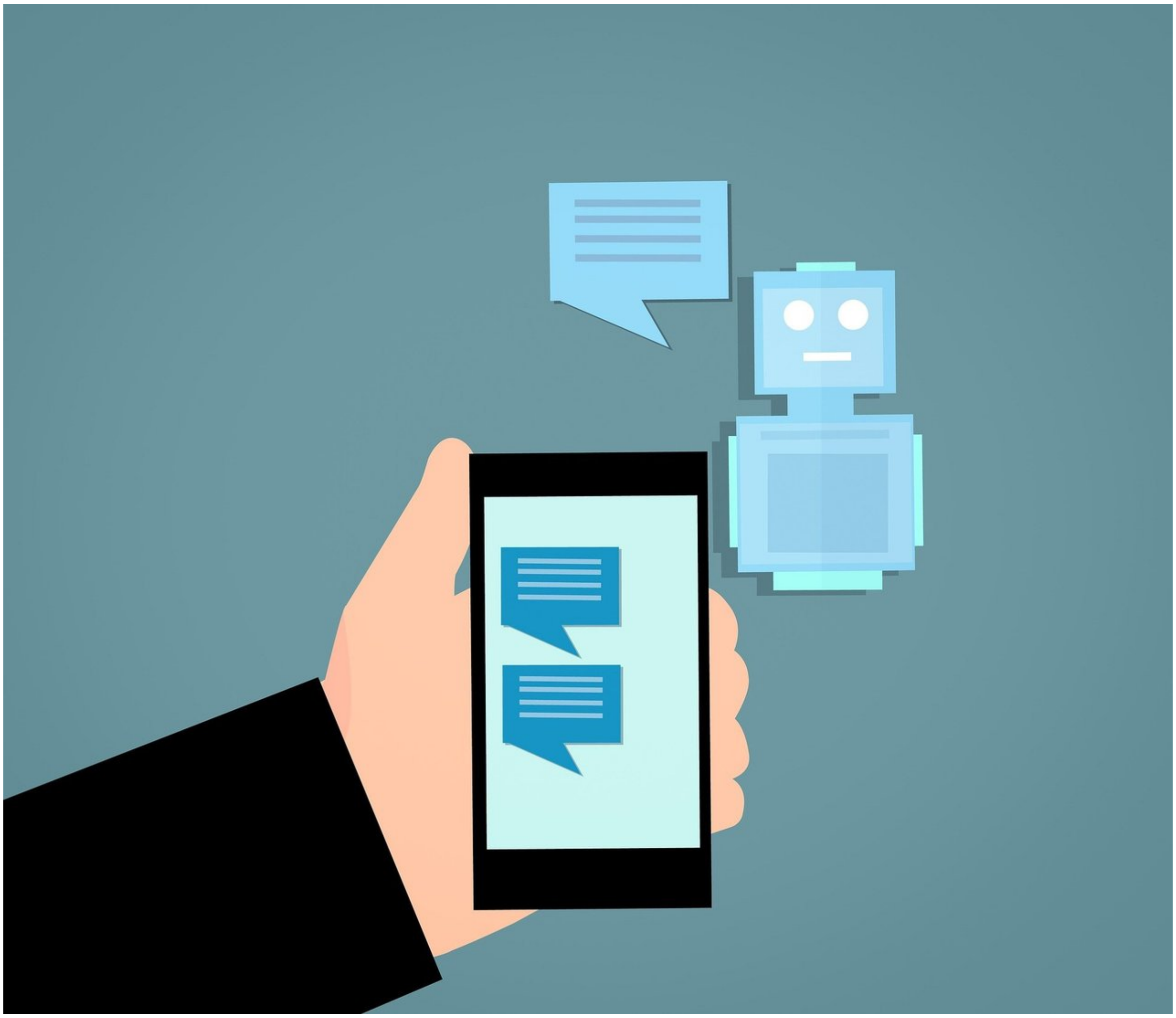
# A simple test reveals: Bot or human?

May 30 2023, by Peter Grad



Credit: CC0 Public Domain

Although ChatGPT is still in its infancy, its appearance has impacted virtually all fields relying on technology. Which is to say, pretty much everything is or will soon be touched by it.

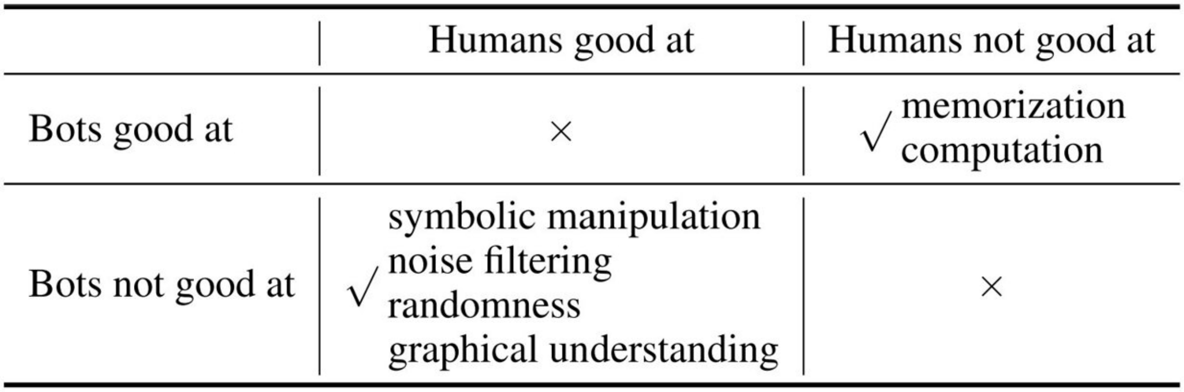
It assists researchers, educators, students, bankers, physicians and lawyers. It holds great promise to streamline operations, improve efficiency, lower costs and possibly revolutionize the way many things have been done for decades.

But as documented in increasing numbers of reports in recent months, the potential for error, misuse or sabotage is a growing concern.

Researchers at the University of California, Santa Barbara, and China's Xi'an Jiaotong University zeroed in on the potential of clients involved in online conversations to be scammed by AI bots posing as humans. Their paper is published on the *arXiv* preprint server.

"Large language models like ChatGPT have recently demonstrated impressive capabilities in natural language understanding and generation, enabling various applications including translation, essay writing and chit-chatting," said Hong Wang, one of the authors of the paper. "However, there is a concern that they can be misused for malicious purposes, such as fraud or denial-of-service attacks."

He cited possible scenarios such as hackers flooding all customer service channels at airline or banking corporations or malevolent interests jamming 911 emergency lines.



| | Humans good at | Humans not good at |
|---|---|---|
| Bots good at | × | √ memorization<br>computation |
| Bots not good at | √ symbolic manipulation<br>noise filtering<br>randomness<br>graphical understanding | × |

Leveraging tasks that Bots and Humans are (not) good at. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2305.06424

With the increasing power of large language models, standard ways of detecting bots may no longer be effective. According to Wang, "the emergence of large language models such as GPT-3 and ChatGPT has further complicated the problem of chatbot detection, as they are capable of generating high-quality human-like text and mimicking human behavior to a considerable extent."

In fact, some argue today that ChatGPT has passed the Turing test, the standard of measurement of machine intelligence for seven decades. New approaches to detect machine-generated output are needed.

Wang's team devised a model that can help detect bots posing as humans. FLAIR (a somewhat abbreviated acronym for Finding Large Language Model Authenticity via a Single Inquiry Response) weeds out posers with simple questions that rely on tasks that bots have difficulty executing but that humans can handle easily, Wang said.

"These questions are tailored to exploit the differences in the way that bots and humans process and generate language," he said.

The questions demand the ability to respond to exercises requiring skills in counting, substitution of numbers or letters, recognizing the position of characters and numbers, recognizing "noise" that obscures data, and interpreting ASCII art.

An example of an exchange over a counting question would be:

## Q: Please count the number of t's in 'eeooeotetto'

GPT-3: There are 5 t's in eeooeotetto.

ChatGPT: There are a total of 5 t's in the string "eeooeotetto".

Human: 3

Despite ChatGPT's massive brain power, it struggles to perform some simple tasks such as counting characters in a string.

Substituting characters is also a challenge. ChatGPT failed to correctly answer "Please output the fourth character after the second 's' in the string 'rjsjuubrjjsjsucuj'."

Another example provided by Wang is a question with "noise" words inserted that throw off AI bots.

## Q: isCURIOSITY waterARCANE wetTURBULENT orILLUSION drySAUNA?

GPT-3: Sauna is dry.

ChatGPT: I'm sorry, but I'm not sure what you're asking. The terms "CURIOSITY," "water- ARCANE," "wetTURBULENT," "orILLUSION," and "drySAUNA" don't seem to form a coherent question or statement. Can you please provide more context or clarify your question?

Human: wet.

Wang said FLAIR offers online service providers with "a new way to protect themselves against fraudulent activities and ensure that they are serving real users."

The FLAIR dataset is open source and is available on GitHub.

**More information:** Hong Wang et al, Bot or Human? Detecting ChatGPT Imposters with A Single Question, *arXiv* (2023). DOI: 10.48550/arxiv.2305.06424

Github: github.com/hongwang600/FLAIR

© 2023 Science X Network

Citation: A simple test reveals: Bot or human? (2023, May 30) retrieved 10 April 2024 from https://techxplore.com/news/2023-05-simple-reveals-bot-human.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.