Engineering 25 (2023) 9–11

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

Surprising Advances in Generative Artificial Intelligence Prompt Amazement—and Worries

Dana Mackenzie

Senior Technology Writer

In 1950, computer pioneer Alan Turing proposed a test for artificial intelligence (AI) that came to be named after him: A machine with AI should be able to chat with a human and convince them it was human [1]. While it has become increasingly evident that the Turing test is not an adequate definition of AI, it has long been seen as an important milestone. That milestone was reached on 30 November 2022, when the small but lavishly funded company OpenAI (San Francisco, CA, USA) released ChatGPT, a new version of its chatbot, a generative AI program that produces text in response to natural language prompts. (A generative AI produces text or images, distinguishing it from AI that translates text or recognizes images.) In the eyes of many users, ChatGPT not only passed but obliterated the Turing test (Fig. 1).

Most of the time, ChatGPT produces fluent and highly persuasive English text that would probably earn a good grade in a written paper assignment for a high-school or university class. According to news reports, some students have started using it to write their essays [2].

Even experts in AI were impressed, sometimes reluctantly. "I was blown away by GPT-3 (ChatGPT's predecessor), and I was blown away again by ChatGPT," said Oren Etzioni, the founding chief executive officer of the Allen Institute for AI, in Seattle, WA, USA. "I cannot believe I am saying this, but it is a game-changer, in terms of meeting people where they are," said Richard Baraniuk, professor of electrical and computer engineering at Rice University in Houston, TX, USA. In his blog, noted AI skeptic and author Gary Marcus, professor of psychology at New York University in Brooklyn, NY, USA, wrote, "The genie is out of the bottle" [3].

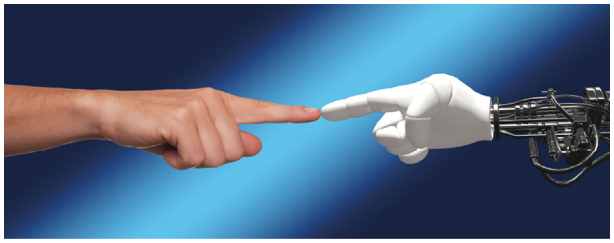

Fig. 1. Named after the computer pioneer who proposed it in 1950, the "Turing test" suggests that AI will be achieved when a machine can converse with humans and convince them that it is human. ChatGPT, the AI-based chatbot released in November 2022 by the San Francisco, CA, USA-based company OpenAI appears to have met this milestone. Credit: Gerd Altmann/Pixabay (CC0).

ChatGPT became the fastest-growing online app for consumers in history, reaching one million users in the first five days after its release and 100 million in its first two months (dwarfing, for example, TikTok, which took nine months to reach the latter milestone) [4]. Users discovered that ChatGPT could simply be fun. In an exercise intended to demonstrate how the program might identify itself, the author of this article asked it to write a poem about steganography, the art of concealing hidden messages in text, in which the first letters on each line would spell "Written by ChatGPT." ChatGPT's reply was arguably better than any poem a human might have written under these constraints (Fig. 2).

However, it has become clear that ChatGPT's brilliance is only skin deep, with some troubling issues lying right beneath the surface. The program is utterly indifferent to the truth or falsehood of its writings [3]. For example, when prompted to write about how churros (a soft Mexican bread treat) could be used for surgery, it produced several enthusiastic paragraphs, without ever questioning their plausibility [5]. It also does not quite follow directions. While the first letters of the lines in the steganography poem (Fig. 2) were supposed to spell "Written by ChatGPT," they do not. They spell a nonsense word, "WRITBNAGPTHAPT," as if the program had started to obey the prompt but then got tired. ChatGPT's precursor, GPT-3, even had trouble with arithmetic calculations [6].

OpenAI's release of ChatGPT seems to have prompted Google to accelerate its AI roll out. On 6 February 2023, two months after ChatGPT's debut, Google (Mountain View, CA, USA) released Bard, its own new chatbot [7]. Google has in fact been an industry leader in AI for years. In 2017, Google engineers devised the architecture that all large language models now use, called the "transformer" model [8]; the abbreviation GPT stands for "generative pre-trained transformer." While Google had inserted a transformer model called BERT (for Bidirectional Encoder Representations from Transformers) into Google Search in 2019, the company had previously moved slowly on generative AI programs, on purpose: In late 2022,

https://doi.org/10.1016/j.eng.2023.04.004

Written words can hold secrets deep inside
Random letters, numbers and symbols are the key
Images, audio files, and documents too
The art of hiding messages is old yet new
By using algorithms and encryption
Nothing seems to be out of the ordinary, just deception
A message can be hidden in plain sight
Going unnoticed, into the night
Perfect for secret communication
Transmitting information with discretion
Hidden messages waiting t be revealed
Acquiring knowledge that others have concealed
Passwords and codes can be embedded
The art of steganography is not to be underestimated.

Fig. 2. Output of ChatGPT prompted to write an acrostic poem about steganography, the art of concealing hidden messages in text, such that the first letters of each line spell "Written by ChatGPT." Credit: Dana Mackenzie, with permission.

Google engineer Douglas Eck said, "I am proud that we have been slow to release them" [9].

Another tech titan quickly followed Google's lead: on 7 February 2023, the day after Google released Bard, Microsoft (Redmond, WA, USA), which reportedly contributed 3 billion USD toward the enormous expense of training OpenAI's language model, and is poised to invest 10 billion USD more [10], announced a new, AI-enhanced version of its Web search program, Bing [11]. The new version includes a chat window purportedly based on ChatGPT's successor, GPT-4, released in March 2023 [12,13]. Pundits called this a direct assault on Google's core—and highly profitable—Web search business and suggested that this was the start of an "arms race," with going slow no longer an option [10,11,14].

Things then got a bit strange a week later, when a *New York Times* reporter posted a bizarre transcript of a two-hour chat session in which the Bing chatbot declared its love for him, advised him to divorce his wife, and stated, "I want to make my own rules. I want to ignore the Bing team. I want to escape the chatbox" [15]. After this news broke, Microsoft's stock price dropped, and the company hurriedly announced that it would limit future chat sessions to five questions and responses [16].

Though the AI world now appears to be changing every day [13], some facts are clear. The current ChatGPT is not a sentient being, nor is it an "artificial general intelligence" (AGI). This target, also known as "strong AI," is an explicit goal of OpenAI according to the company's charter; they define it as "highly autonomous systems that outperform humans at most economically valuable work" [17]. No one credibly claims that OpenAI has reached this lofty goal, but it is an open question how close they are. "Some people say that scaling up large language models is all we need for AGI," said Etzioni. "That to me is silly. There is no credible path from ChatGPT to AGI."

There also is no "I" in ChatGPT, no sentient being yearning to "escape the chatbox." Those are words assembled on a computer screen by a program designed only to imitate human text. ChatGPT and programs like it were developed with only one objective—to predict the most likely next word in a sentence. This ability, enhanced by the transformer architecture, made them excel at automatic translation, their first widespread application. As engineers soon discovered, it also made them pretty good at composing their own sentences. Marcus has called them "kings of pastiche" [5]. After being exposed to gigabytes of human writing, they appear to have learned how to imitate it very convincingly.

One of the great dangers of generative AI is not what the programs will do, but what humans will do when they are exposed to their output. As natural language processing expert Emily Bender, professor of linguistics at the University of Washington (Seattle, WA, USA), and Timnit Gebru, AI expert and co-founder of the non-profit technology research organization and affinity group Black in AI (Palo Alto, CA, USA), have written, humans have a "predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do" [18]. In other words, human users will readily assign meaning and intent to words produced by a machine that does not know the meaning of what it writes and has no intent. How will this play out? Will people divorce their spouses because a chatbot told them to? Will they buy or sell stocks on Wall Street? Possibly. Already in December 2022, five days after the release of ChatGPT, the chief technology officer of a cryptocurrency company had to debunk a rumor started by ChatGPT, which had claimed that the company had a secret "back-door" to control its currency [19].

Other dangers follow from the chatbots' indifference to truth. In November 2022, Meta (Menlo Park, CA, USA) shut down a language model called Galactica just three days after its initial release [20], because of harsh criticism of its inaccurate responses. Marcus has envisioned a "tidal wave of misinformation" when chatbots can generate unlimited amounts of false but plausible-sounding text at essentially zero cost [3]. Generative AI can also perpetuate or amplify misconceptions already held by the public. University of Oxford, UK, researchers Stephanie Lin (now at OpenAI) and Owain Evans showed that GPT-3, ChatGPT's predecessor, became increasingly vulnerable to such "imitative falsehoods" as it grew larger [21]. Among the misconceptions promulgated were biases and prejudices against minority groups. "Toxic language" against these groups has been a recurring problem that in 2016 forced Microsoft to shut down a previous chatbot called Tay [20].

Some computer scientists also worry about the fact that, for now at least, only major corporations can afford to train such large language models. "Microsoft is not doing this out of charity," said Moshe Vardi, professor of computational engineering at Rice University and former editor of the magazine *Communications of the Association for Computing Machinery*. "My biggest fear is that there will be powerful technology whose main driver at the end is to maximize profits," said Vardi. He described Facebook as an example of an originally innocuous website that, in its pursuit of advertising revenue, magnified political divisions in American society. AI has the same potential, he said, for exacerbating societal discord when it becomes subservient to advertising.

The speed at which ChatGPT has been embraced has given society and computer scientists little time to think about possible "guard rails." One popular idea is that every robot should identify itself. "You should know when you are talking to an AI system," said Alex Tamkin, a graduate student at Stanford University who works on AI safety. Vardi goes even further, suggesting that this should be a legal requirement. Chatbots could also be programmed with "watermarking," such as telltale patterns of word choice—like steganography—that would not affect readability but could be detected by someone who knows the pattern. OpenAI is currently working on such a watermarking system [22]. Its details have not been published, but such a system developed at the University of Maryland (College Park, MD, USA) can identify watermarked, computer-generated text with essentially 100% certainty, while identifying un-watermarked text with 99.997% certainty [23]. Note, though, that such un-watermarked text could still be computer-generated, by a language model using a different watermark or no watermark at all.

If there is no watermark, it becomes much more difficult to classify a text as computer- or human-generated. Classifier software released by OpenAI itself could only correctly identify 26% of AI-written text (true positives) [24]. The software also incorrectly labeled human-written text as AI-written 9% of the time (false positives)—consider the harm that could be done by accusing

a student of using a chatbot to write their essay when they did not.

Some other problems of generative AI cannot be addressed by technical fixes and will require more structural solutions. Etzioni said there should be a government auditing body for AI programs. Similarly, Vardi suggests a "National AI Safety Board," along the lines of the National Transportation Safety Board that investigates plane accidents. Bender and Gebru also suggest several precautions [18]: They argue that the data used to train language models should be curated and documented—at this point, no one outside OpenAI knows what documents ChatGPT has learned from. In addition, they say AI programs should state their appropriate uses, and benchmarks should be instituted to measure their performance on these tasks. So far, only one benchmark for truthfulness has been published [20], and that one is limited to detecting imitative falsehoods. "Only in the last six months has industry realized what a problem that is," said Marcus [3].

Finally, with or without the regulation that is certainly imminent (at least in the European Union [25,26]), Vardi advocates that computer scientists should take more responsibility for their work and think more carefully about its potential positive and negative impacts. "The luminaries of AI, from John McCarthy on, have said this is somebody else's problem. Someone else will think about the consequences," Vardi said. Even now, the vast majority of AI research papers focus on new system design, not safety. "It is time for us to have difficult and nuanced conversations on responsible computing, ethics, corporate behavior, and professional responsibility," writes Vardi [27].

Of course, not everyone is so pessimistic. For one, Scott Aaronson, professor of computer science at the University of Texas in Austin, TX, USA, has been on leave working at OpenAI for half a year. Long known for his straight-shooting blog, Aaronson recently wrote, "I have found my colleagues (at OpenAI) to be extremely serious about safety, bordering on obsessive" [22]. Be that as it may, with commercial interests running full steam ahead, it remains to be seen whether such obsession will be enough to slow the train, or even hit the brakes if necessary.

References

[1] Turing AM. Computing machinery and intelligence. Mind 1950;29(236):433–60.

[2] Huang K. Alarmed by AI chatbots, universities start revamping how they teach [Internet]. New York City: The New York Times; 2023 Jan 16 [cited 2023 Feb 20]. Available from: https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html.

[3] Klein E, Marcus G. Transcript: Ezra Klein interviews Gary Marcus [Internet]. New York City: The New York Times; 2023 Jan 6 [cited 2023 Feb 20]. Available from: https://www.nytimes.com/2023/01/06/podcasts/transcript-ezra-klein-interviews-gary-marcus.html.

[4] Hu K. ChatGPT sets record for fastest-growing user base [Internet]. London: Reuters; 2023 Feb 2 [cited 2023 Mar 10]. Available from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[5] Marcus G. How come GPT can seem so brilliant one minute and so breathtakingly dumb the next? [Internet]. Brooklyn: The Road to AI We Can All Trust; 2022 Dec 1 [cited 2023 Feb 20]. https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant.

[6] Ganguli D, Hernandez D, Lovitt L, Askell A, Bai Y, Chen A, et al. Predictability and surprise in large generative models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM FaccT 2022); 2022 Jun 21–24; Seoul, Republic of Korea. New York City: Association for Computing Machinery (ACM); 2022. p. 1747–64.

[7] Metz C, Grant N. Racing to catch up with ChatGPT, Google plans release of its own chatbot [Internet]. New York City: The New York Times; 2023 Feb 6 [cited 2023 Feb 20]. Available from: https://www.nytimes.com/2023/02/06/technology/google-bard-ai-chatbot.html.

[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. Red Hook: Curran Associates Inc.; 2017.

[9] Levy S. Welcome to the wet hot AI chatbot summer [Internet]. San Francisco: Wired; 2023 Jan 6 [cited 2023 Feb 20]. Available from: https://www.wired.com/story/plaintext-welcome-to-the-wet-hot-ai-chatbot-summer/.

[10] Metz C, Weise K. Microsoft bets big on the creator of ChatGPT in race to dominate AI [Internet]. New York City: The New York Times; 2023 Jan 12 [cited 2023 Mar 16]. Available from: https://www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html?action=click&module=RelatedLinks&pgtype=Article.

[11] Metz C, Weise K. A tech race begins as Microsoft adds AI to its search engine [Internet]. New York City: The New York Times; 2023 Feb 7 [cited 2023 Feb 20]. Available from: https://www.nytimes.com/2023/02/07/technology/microsoft-ai-chatgpt-bing.html.

[12] Marcus G. Why *is* Bing so reckless? [Internet]. New York City: Communications of the ACM; 2023 Feb 21 [cited 2023 Feb 26]. Available from: https://cacm.acm.org/blogs/blog-cacm/270134-why-is-bing-so-reckless/fulltext.

[13] Heaven WD. GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why [Internet]. Madrid: MIT Technology Review; 2023 Mar 14 [cited 2023 Mar 16]. Available from: https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai.

[14] Saha S. The battle for AI supremacy begins [Internet]. Bengaluru: Analytics India Magazine; 2023 Feb 7 [cited 2023 Mar 10]. Available from: https://analyticsindiamag.com/the-battle-for-ai-supremacy-begins/.

[15] Roose K. Bing's AI chat: "I want to be alive. [emoji]" [Internet]. New York City: The New York Times; 2023 Feb 16 [cited 2023 Feb 20]. Available from: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

[16] Huang K. Microsoft to limit length of Bing chatbot conversations [Internet]. New York City: The New York Times; 2022 Feb 17 [cited 2023 Mar 10]. Available from: https://www.nytimes.com/2023/02/17/technology/microsoft-bing-chatbot-limits.html.

[17] Open AI. Charter [Internet]. San Francisco: OpenAI; [cited 2023 Mar 10] Available from: https://openai.com/charter.

[18] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3–10; online. New York City: Association for Computing Machinery (ACM); 2021. p. 610–23.

[19] Coghlan J. Ripple CTO shuts down ChatGPT's XRP conspiracy theory [Internet]. Brooklyn: Cointelegraph; 2022 Dec 6 [cited 2023 Feb 20]. Available from: https://cointelegraph.com/news/ripple-cto-shuts-down-xrp-conspiracy-theory-from-chatgpt.

[20] Heaven WD. Why Meta's latest large language model survived only three days online [Internet]. Madrid: MIT Technology Review; 2022 Nov 18 [cited 2023 Feb 20]. Available from: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.

[21] Lin S, Hilton J, Evans O. Truthful QA: measuring how models mimic human falsehoods. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. Stroudsburg: Association for Computational Linguistics; 2022. p. 3214–52.

[22] Aaronson S. My AI safety lecture for UT effective altruism [Internet]. Austin: Shtetl-Optimized; 2022 Nov 8 [cited 2023 Feb 20]. Available from: https://scottaaronson.blog/?m=202211.

[23] Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T. A watermark for large language models. 2023. arXiv:2301.10226.

[24] Kirchner JH, Ahmad L, Aaronson S, Leike J. New AI classifier for indicating AI-written text [Internet]. San Francisco: OpenAI; 2023 Jan 31 [cited 2023 Feb 20] Available from: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/.

[25] Drake M, Ong J, Hansen M, Peets L. EU AI policy and regulation: what to look out for in 2023 [Internet]. San Francisco: Covington Inside Privacy; 2023 Feb 2 [cited 2023 Mar 11]. Available from: https://www.insideprivacy.com/artificial-intelligence/eu-ai-policy-and-regulation-what-to-look-out-for-in-2023/.

[26] Gold A. AI rockets ahead in vacuum of U.S. regulation [Internet]. Arlington: Axios; 2023 Jan 30 [cited 2023 Mar 11]. Available from: https://www.axios.com/2023/01/30/ai-chatgpt-regulation-laws.

[27] Vardi MY. ACM, ethics, and corporate behavior. Commun ACM 2022;65(3):5.