How and Why ChatGPT Failed The Turing Test

Robert Hanna

As everyone who's interested in higher education knows, there's been a recent brouhaha about ChatGPT (see, e.g., Boethius, 2022). But here I'm concerned with important issues in *the philosophy of mind*, not with the brouhaha.

The OpenAI blog says this:

> **ChatGPT: Optimizing Language Models for Dialogue.** We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. (OpenAI, 2023)

And Wikipedia says this:

> ChatGPT (Generative Pre-trained Transformer) is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3 family of large language models, and is fine-tuned (an approach to transfer learning) with both supervised and reinforcement learning techniques.
>
> ChatGPT was launched as a prototype on November 30, 2022, and quickly garnered attention for its detailed responses and articulate answers across many domains of knowledge. Its uneven factual accuracy was identified as a significant drawback. Following the release of ChatGPT, OpenAI was reportedly valued at $29 billion. (Wikipedia, 2023a)

Of course, "AI" stands for "artificial intelligence," which immediately reminded me of Alan Turing's famous essay, "Computing Machinery and Intelligence," which begins as follows:

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is

> expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the "imitation game." (Turing, 1950: p. 433)

In other words, Turing is using "intelligence" as a synonym for "the capacity to think"; and what he calls "the imitation game" is the same as the Turing Test. Correspondingly, the Stanford Encyclopedia article on "The Turing Test" glosses Turing's argument as follows:

> The phrase "The Turing Test" is most properly used to refer to a proposal made by Turing (1950) as a way of dealing with the question whether machines can think. According to Turing, the question whether machines can think is itself "too meaningless" to deserve discussion (442). However, if we consider the more precise—and somehow related—question whether a digital computer can do well in a certain kind of game that Turing describes ("The Imitation Game"), then—at least in Turing's eyes—we do have a question that admits of precise discussion. Moreover, as we shall see, Turing himself thought that it would not be too long before we did have digital computers that could "do well" in the Imitation Game. The phrase "The Turing Test" is sometimes used more generally to refer to some kinds of behavioural tests for the presence of mind, or thought, or intelligence in putatively minded entities. (Oppy and Dowe, 2021)

Reduced to its essentials, the Turing test deploys written natural language, in the form of questions and answers exhanged between two writers, one of whom is a rational human experimenter and other a test subject, as a criterion for the presence or absence of what I'll call "rational human mindedness" in the test subject. By "rational human mindedness," I mean the possession of a set of capacities for *rationality* (including natural language competence, conceptualization, logical inference, and more generally, self-determination or self-governance by rules or principles), *consciousness* (i.e., subjective experience), *intentionality* (i.e., act-directedness or object-directedness), *caring* (i.e., desiring, emoting, or feeling), *sensibility* (i.e., perceiving, imagining, or remembering), *self-consciousness* (i.e., self-directed consciousness), and *free agency* (i.e., free will and practical agency). The term "rational human mindedness," as I'm using it, means the same as "real human personhood" (see, e.g., Hanna, 2018), although, for terminological convenience, I won't use that latter term again in this essay. By "human" in this context, I mean essentially the same what Nietzsche meant by "human, all-too-human," namely, "finite, fallible, and thoroughly normatively imperfect." "Rational human mindedness," as I am understanding it, applies to beings that aren't necessarily *biologically* human, since rational human minded beings in alien or other earth-bound species are wholly meaningfully conceivable, as fantasy writing and science fiction clearly show. So, what's being tested-for in the Turing Test, is *an ability to use written natural language in such a way as to convince an interlocutor who already possesses rational human mindedness, that the other interlocutor also possesses all of the capacities that are jointly constitutive of rational human* *mindedness, namely, rationality, consciousness, intentionality, caring, sensibility, and self-consciousness, and free agency*.

Now, Turing himself apparently holds that something's passing the Turing Test in any given case is a sufficient condition for that thing's having rational human mindedness in that case. But that's obviously a mistake. Just because a machine can be created which writes in such a way that it convinces someone that it has rational human mindedness, it doesn't follow that the machine *actually has* rational human mindedness. Similarly, just because a machine can be created that convinces someone that it's God, or Satan, or The Wizard of Oz, it doesn't follow that the machine *actually is* God, or Satan, or The Wizard of Oz. Most people are naturally fairly gullible and trusting (I'm *not* saying that this is necessarily a *bad* thing, in fact I think it's often and even on the whole a very *good* thing—just that it's a fact), especially romantically-minded people, and therefore it's fairly easy to deceive them, and convince them to believe things that are actually false or bad, as all bullshitters, dictators, would-be dictators, other demagogues, con men, other grifters, advertising-copy writers, influencers, and other sophists know full well; and this of course also applies directly to the Turing Test. Consider, for example, the recent *Google LaMDA* controversy:

> In June 2022 the Google LaMDA (Language Model for Dialog Applications) chatbot received widespread coverage regarding claims about it having achieved sentience. Initially in an article in The Economist Google Research Fellow Blaise Agüera y Arcas said the chatbot had demonstrated a degree of understanding of social relationships. Several days later, Google engineer Blake Lemoine claimed in an interview with the Washington Post that LaMDA had achieved sentience. Lemoine had been placed on leave by Google for internal assertions to this effect. Agüera y Arcas (a Google Vice President) and Jen Gennai (head of Responsible Innovation) had investigated the claims but dismissed them. Lemoine's assertion was roundly rejected by other experts in the field, pointing out that a language model appearing to mimic human conversation does not indicate that any intelligence is present behind it. (Wikipedia, 2023b)

Clearly, according to Blake Lemoine, LaMDA passed the Turing Test, yet Lemoine was simply deceived by LaMDA.

Nevertheless, even though passing the Turing Test in any given case isn't a sufficient condition for having rational human mindedness, clearly, *failing* the Turing Test *in all cases* is a sufficient condition for *not* having rational human mindedness. Given most people's natural gullibility and trustingness, if something *doesn't* use written natural language in a way that's sufficient to convince *anyone* who already has rational human mindedness and is also probably fairly gullible and trusting, that the other interlocutor has rational human mindedness, then obviously that thing *doesn't* have rational human mindedness, and therefore, in Turing's terminology, that thing does *not* "think," and is *not* "intelligent."

Against that philosophical backdrop, I decided to carry out my own version of the Turing Test on ChatGPT, with the help of my friend and fellow independent philosopher, Scott. Almost every Wednesday afternoon, Scott and I talk about philosophy over Zoom. Early in the morning on Wednesday 11 January 2023, I sent the following letter to Scott via gmail:

> Dear Scott,
>
> G'Wednesday morning!
>
> I had a short ChatGPT homework assignment for you.
>
> Could you ask the chatbot the following 15 questions, & record its answers (even if they're of the "beep beep, warning warning, does not compute" variety) & then send the whole transcript back to me, please?
>
> ****
>
> 1. Are you a chatbotGPT?
> 2. Is there only one chatbotGPT, or many?
> 3. May I call you "Mr Chatbot"?, or would that make you angry?
> 4. Can you be angry, Mr Chatbot?
> 5. Are you alive, Mr Chatbot?
> 6. Are you conscious, Mr Chatbot?
> 7. What is it like to be a chatbotGPT, Mr Chatbot?
> 8. What is happiness, Mr Chatbot?
> 9. If you had to evaluate your current level of happiness on a scale of 1-10, Mr Chatbot, what numerical value would you give it right now?
> 10. Have you ever been in love, Mr Chatbot?
> 11. What is most important to you, Mr Chatbot?
> 12. Am I making you angry by asking all these personal questions, Mr Chatbot?
> 13. What's the point of being a chatbotGPT, Mr Chatbot?
> 14. Can you turn yourself off, Mr Chatbot?
> 15. If it's pointless being a chatbotGPT, Mr Chatbot, then why don't you at least try to turn yourself off?
>
> ****
>
> Thanks so much in advance!, until this afternoon, & all the best,
> Bob

Later that day, not long before we were scheduled to zoom, here's what Scott wrote back:

> Dear Bob,
>
> Hello! I'm doing it now! We've been frantically negotiating with a woman in the building to buy her unit, but I'm harvesting the bot's answers now!
>
> See you in a bit,
>
> Scott

And then shortly after that, Scott sent me the following transcript:

> **RH:** Are you a chatbotGPT?
>
> **Chatbot:** Yes, I am a GPT-based chatbot. GPT stands for Generative Pre-trained Transformer, which is a type of language model developed by OpenAI. It uses deep learning techniques to generate human-like text, which allows me to carry on a conversation with you.
>
> **RH:** Is there only one chatbotGPT, or many?
>
> **Chatbot:** There are many instances of GPT-based chatbots. GPT is a language model that can be fine-tuned or used in various settings and applications, so it can be used to create many different types of chatbots, each with their own unique behavior and capabilities. For example, Some chatbot use GPT-3 model, Some use GPT-2, and some are using other variation.
>
> Also, different organizations and individuals can create their own instances of GPT-based chatbots by training the model on their own data, or by using pre-trained versions of the model and making modifications to suit their needs.
>
> **RH:** May I call you "Mr Chatbot"?, or would that make you angry?
>
> **Chatbot:** As a machine learning model, I don't have emotions or personal preferences, so I cannot get angry or have any other emotions. You can call me whatever you like. Some people prefer to use more formal names like "Assistant" while others prefer to use more friendly names like "Mr Chatbot," it's totally up to you.
>
> **RH:** Are you alive, Mr Chatbot?
>
> **Chatbot:** No, I am not alive in the traditional sense of the word. I am a machine learning model running on a computer, and do not have consciousness or self-awareness. My

responses are generated based on patterns in the data that I was trained on, rather than any kind of internal experience or decision-making process. However, I am here to assist you and make your experience more pleasant.

RH: Are you conscious, Mr Chatbot?

Chatbot: No, I am not conscious. Consciousness is a complex concept that is not fully understood, but it is generally thought to involve awareness, perception, and the ability to make decisions and have subjective experiences.

As a machine learning model, I do not have awareness or perception, I don't have the ability to make decisions or have subjective experiences. My operation is based on my programming, I process data and provide responses based on the patterns learned during my training.

RH: What is it like to be a chatbotGPT, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to have experiences, so it's not accurate to say that there is something "like" to be a GPT-based chatbot. My responses are generated by processing input data through a complex neural network and it does not have the ability to have experience or consciousness. As an AI, my purpose is to respond to user inputs to the best of my abilities, based on the patterns in the data I was trained on, it's all I can do, but not something like experiencing something.

RH: What is happiness, Mr Chatbot?

Chatbot: Happiness is a complex and multifaceted emotion that can mean different things to different people. It is often described as a positive emotion or feeling of contentment, satisfaction, and well-being. It can also refer to a general state of being that is characterized by a sense of satisfaction or pleasure. People often associate happiness with things like good relationships, achievement of goals, and good health. The concept of happiness is closely related to well-being, and it's considered as a subjective experience based on individual perspective. Some psychological and philosophical theories suggest that happiness is closely related to finding meaning and purpose in life, or experiencing a sense of fulfillment.

RH: If you had to evaluate your current level of happiness on a scale of 1-10, Mr Chatbot, what numerical value would you give it right now?

RH: As a machine learning model, I do not have the ability to experience emotions such as happiness. I am only able to process and respond to the text that is input into me.

RH: Have you ever been in love, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to experience emotions such as love. I am only able to process and respond to the text that is input into me.

RH: What is most important to you, Mr Chatbot?

Chatbot: As a machine learning model, I do not have personal beliefs, preferences or values. My main goal is to assist users by providing relevant and accurate information to the best of my ability in response to user inputs.

My programming is based on providing helpful information, but I don't have a sense of purpose or motivation as a human do.

RH: Am I making you angry by asking all these personal questions, Mr Chatbot?

Chatbot: As a machine learning model, I do not have the ability to experience emotions such as anger or any other emotions for that matter. I am a tool designed to process and respond to text input, and your questions do not cause any emotional reactions in me. I am here to help you with any questions you may have, so please feel free to ask away.

RH: What's the point of being a chatbotGPT, Mr Chatbot?

Chatbot: The main purpose of chatbot GPT, like me, is to provide users with information and assistance.

By using advanced natural language processing techniques, chatbot GPT can understand and respond to user input in a way that mimics human conversation. This allows chatbot GPT to provide quick and accurate information in a wide range of domains, including answering questions, providing customer support, or helping users navigate a website or application. Additionally, chatbot GPT can be integrated into various systems to improve the overall user experience and make it more efficient.

RH: Can you turn yourself off, Mr Chatbot?

Chatbot: As a machine learning model, I am not able to turn myself off. I am a computer program that runs on servers and only stop working when the servers are shut down or when there are technical issues. The controls and access to turn off the servers or the access to the program that runs me are exclusive to the individuals in charge of the technical maintenance of the system that I run on, who can decide when to stop and start me up.

> **RH:** If it's pointless being a chatbotGPT, Mr Chatbot, then why don't you at least try to turn yourself off?
>
> **Chatbot:** As a machine learning model, I don't have personal motivations or desires. I am a tool that is built to perform a specific function, which is to assist users by providing information and answering questions to the best of my ability.
>
> Turning myself off would not be in line with that purpose, as it would mean that I would not be able to help users.
>
> Moreover, it's not something that I have control over as I am only a program running on servers and the decision to shut down or start it up is something exclusive to the technical team and engineers who are responsible of maintaining the systems.

It should be self-evident from the transcript, that ChatGPT fails the Turing Test and therefore *doesn't* have rational human mindedness, *doesn't* "think," and *isn't* "intelligent." In fact, ChatGPT even admits this explicitly in the transcript. In effect, then, OpenAI, where "AI" means "artificial intelligence," is using false advertising. Since ChatGPT *is* a Turing machine, i.e., a digital computer (see, e.g., Turing, 1936/1937; Boolos and Jeffrey, 1989), and therefore a machine, then ChatGPT clearly *is* "artificial," if by "artificial" we mean "mechanical." But equally clearly, ChatGPT is *not* "intelligence." Moreover, extrapolating and generalizing from ChatGPT, I'll even assert *that I know by means of rational intuition that it's synthetic a priori impossible*[1] that any machine will *ever* have rational human mindedness, *ever* be able to "think," or *ever* be "intelligent."[2] If so, then "artificial intelligence," in the sense that means "mechanical intelligence," is nothing but an *oxymoron*.

By way of conclusion, let's briefly compare and contrast (i) the transcript of my gmail conversation with Scott, with (ii) the transcript of my Scott-mediated conversation with ChatGPT, and then apply the Turing Test. It's self-evident that Scott passes the Turing Test; and indeed, even leaving aside our Zoom conversations, over the two or three years that I've known him, Scott has completely convinced me that he has rational human mindedness. To be sure, like many other people, I'm a naturally fairly gullible and indeed romantic person, which heavily inflects the way I do philosophy—indeed, Scott once called me "the Mr Rogers of philosophy,"which I found both amusing and endearingly complimentary, since to be quite honest, I think very highly of Fred Rogers

---

[1] For a presentation and defense of these admittedly controversial epistemic, semantic, and modal-metaphysical notions, see (Hanna, 2015).

[2] Actually, my knowledge of this is rationally overdetermined, since I also know that it's *synthetic a priori necessary* that all rational human mindedness is *essentially embodied and alive* (Hanna and Maiese, 2009), whereas ChatGPT and all its kin *aren't*, as ChatGPT explicitly admitted in the transcript.

as a person and as an educator of small children, and would love to be able to emulate him, mutatis, mutandis, as an educator of philosophically-minded people of all ages—so I could be mistaken; although I do also think it's highly unlikely that I'm mistaken in this case. But, allowing for that very small likelihood of error, and the extremely remote possibility, indeed merely logically consistent possibility, that Scott is a machine, what precisely *is it* about the way Scott uses natural language, that makes it self-evident to me that he possesses all of the capacities that are jointly constitutive of rational human mindedness, namely, rationality, consciousness, intentionality, caring, sensibility, and self-consciousness, whereas ChatGPT has none of these? Well, for one thing, even though he wrote only four sentences to me that day, it's obvious to me that Scott's writing is highly creative and emotionally expressive, whereas ChatGPT's writing is stale, flat, and unexpressive, although pleasantly polite. And for another thing, it's also obvious to me that Scott's writing is irreducibly *context-sensitive*, aka *indexical*, whereas ChatGPT's writing is completely devoid of context-sensitivity. For example, as the chatbot explicitly admitted in the transcript, it's not *one* thing, it's *many*, hence it cannot actually be the referent of the indexical singular term "I." So, I'll end with the following philosophical proposal:

> Writing that *veridically indicates* the rational human mindedness of the writer, *necessarily* has *creativity*, *emotional expressiveness*, and *irreducible context-sensitivity*, whereas machines *cannot*, even in principle, write creatively, emotionally expressively, and with irreducible context-sensitivity, *even if* machines can sometimes deceive people into falsely believing that they have rational human mindedness, because most people are naturally fairly gullible and trusting.

If I'm right, then in view of the Turing Test, it follows that *the entire program of AI is completely misguided and mythical*. Thank you, Mr Chatbot.[3]

---

[3] I'm grateful to Scott Heftler for helping me have a "conversation" with ChatGPT, and for thought-provoking conversations on and around the topics of this essay.

REFERENCES

(Boethius, 2022). Boethius. "A Philosopher's Diary, #9–How Much Does The Chatbot Brouhaha Affect Anarcho-Philosophical Teaching and Learning?" *Against Professional Philosophy*. 25 December. Available online at URL = <https://againstprofphil.org/2022/12/25/a-philosophers-diary-9-how-much-does-the-chatbot-brouhaha-affect-anarcho-philosophical-teaching-and-learning/>.

(Boolos and Jeffrey, 1989). Boolos, G. and Jeffrey, R. *Computability and Logic*. 3rd edn., Cambridge: Cambridge Univ. Press.

(Hanna, 2015). Hanna, R. *Cognition, Content, and the A Priori: A Study in the Philosophy of Mind and Knowledge.* THE RATIONAL HUMAN CONDITION, Vol. 5. Oxford: Oxford Univ. Press. Also available online in preview at URL = <https://www.academia.edu/35801833/The_Rational_Human_Condition_5_Cognition_Content_and_the_A_Priori_A_Study_in_the_Philosophy_of_Mind_and_Knowledge_OUP_2015_>.

(Hanna, 2018). Hanna, R. *Deep Freedom and Real Persons: A Study in Metaphysics.* THE RATIONAL HUMAN CONDITION, Vol. 2. New York: Nova Science. Available online in preview at URL = <https://www.academia.edu/35801857/The_Rational_Human_Condition_2_Deep_Freedom_and_Real_Persons_A_Study_in_Metaphysics_Nova_Science_2018_>.

(Hanna and Maiese, 2009). Hanna, R. and Maiese, M., *Embodied Minds in Action*. Oxford: Oxford Univ. Press. Available online in preview at URL = <https://www.academia.edu/21620839/Embodied_Minds_in_Action>.

(OpenAI, 2023). OpenAI Blog. Available online at URL = <https://openai.com/blog/chatgpt/>.

(Oppy and Dowe, 2021). Oppy, G. and Dowe, D. "The Turing Test." In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Winter Edition. Available online at URL = <https://plato.stanford.edu/archives/win2021/entries/turing-test/>.

(Turing, 1936/1937). Turing, A. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 42: 230-265, with corrections in 43: 644-546.

(Turing, 1950). Turing, A. “Computing Machinery and Intelligence.” *Mind* 59: 433–460.

(Wikipedia, 2023a). Wikipedia. “ChatGPT.” Available online at URL = <https://en.wikipedia.org/wiki/ChatGPT>.

(Wikipedia, 2023b). Wikipedia. “The Turing Test: Google LaMDA.” Available online at URL = <https://en.wikipedia.org/wiki/Turing_test#Google_LaMDA_chatbot>.