Introduction

"Much of what we do with machine learning happens beneath the surface. Though less visible, much of the impact of machine learning will be of this type – quietly but meaningfully improving core operations".

Jeff Bezos[1]

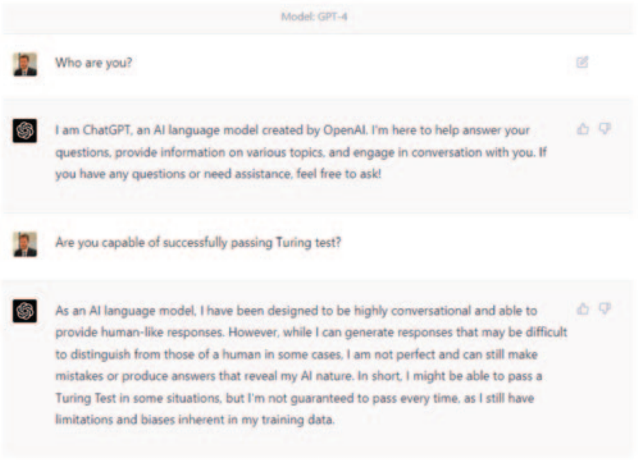
“Artificial intelligence (AI) is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable“[2].

This was the definition of AI offered by John McCarthy, professor at Stanford University, one of the founders of this discipline and co-author of the term “artificial intelligence”.

However, as early as 1950 Alan Turing asked[3]: “can machines think?” and formulated what would later become known as the “Turing test”: a test of a machine's ability to display intelligence indistinguishable from that of a human being. Turing proposed that a human evaluator judge natural language conversations between a person and a machine designed to generate human-like responses. If the evaluator was unable to distinguish the machine from the human, the machine would have passed the test.

Although there is controversy in this regard[4], many authors consider that there are already artificial intelligences that could pass the Turing test, such as GPT-4, from the Open AI Foundation, although GPT-4 itself is not so sure about it (Fig. 1). There are also more sophisticated tests, such as Winograd's schema test, which consists of solving complex anaphora that require knowledge and common sense , something that the current AI does not seem to be able to do yet.

Figure 1. Conversation with GPT-4 about its ability to pass the Turing test.

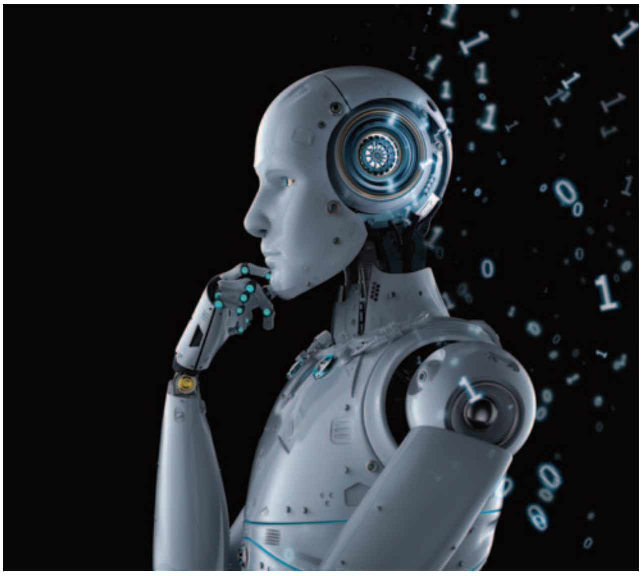


Who are you?

I am ChatGPT, an AI language model created by OpenAI. I'm here to help answer your questions, provide information on various topics, and engage in conversation with you. If you have any questions or need assistance, feel free to ask!

Are you capable of successfully passing Turing test?

As an AI language model, I have been designed to be highly conversational and able to provide human-like responses. However, while I can generate responses that may be difficult to distinguish from those of a human in some cases, I am not perfect and can still make mistakes or produce answers that reveal my AI nature. In short, I might be able to pass a Turing Test in some situations, but I'm not guaranteed to pass every time, as I still have limitations and biases inherent in my training data.



---

[1] Bezos (b. 1964), J., founder, executive chairman and former CEO of Amazon.

[2] McCarthy (2004). Professor of Computer Science at Stanford University.

[3] Turing (1950). British mathematician, logician, theoretical computer scientist, cryptographer, philosopher and theoretical biologist.

[4] Harnad (2003). Professor of Psychology at the University of Quebec in Montreal (UQAM) and McGill University, and Emeritus Professor of Cognitive Science at the University of Southampton.

[5] A Winograd scheme is a binary choice question where (i) there are two parties mentioned in the question; (ii) pronouns are used to refer to them; (iii) there is ambiguity about who the pronoun refers to; and (iv) there are specific words that can change the correct answer. In an example from Terry Winograd (Professor of Computer Science at Stanford University):

- Question: the city councilmen refused the demonstrators a permit because they [feared/advocated] violence. Who [fears/advocates] violence?
- Answer: [the city councilmen / the demonstrators].

With this, an alternative test to the Turing Test can be generated, using such questions and heavily penalizing wrong answers (see Levesque (2014)).

Even so, although the field of AI is not new, dizzying breakthroughs have been made in recent years, with applications ranging from self-driving cars to medical diagnostics, automatic trading, facial recognition, energy management, cybersecurity, robotics or machine translation, to name a few.

A distinguishing feature of today's AI is precisely linked to McCarthy's definition mentioned above: it is not limited to observable methods, and, when it reaches a certain level of complexity, it poses interpretability challenges. In other words: AI models tend to have a high performance, much higher than traditional algorithms; but in each specific case it can be extremely complex to explain why the model has produced a given result.

Although there are applications of AI where it is not as important to be able to understand or explain why the algorithm has returned a particular value, in many cases it is essential and is a regulatory requirement. For example, in the European Union, under the General Data Protection Regulation (GDPR), consumers have what is known as the "right to an explanation"[6]:

> [...] not to be subject to a decision based solely on automated processing [...], such as automatic refusal of an online credit application [...] without any human intervention", and [the data subject] has the right "to obtain an explanation of the decision reached [...] and to challenge the decision".

All this has led to the development of the Explainable Artificial Intelligence (XAI) discipline, which is the field of study that aims to make AI systems understandable to humans[7], as opposed to the notion of "black box", which refers to algorithms in which only the results are observable and the operation of the model is unknown, or the basis for the results cannot be explained.

It can be concluded[8] that an algorithm falls within the XAI discipline if it follows three principles: transparency, interpretability and explainability. Transparency occurs if the processes that calculate the parameters of the models and produce the results can be described and justified. Interpretability describes the ability to understand the model and present how it makes decisions in a human-understandable way. Explainability refers to the ability to decipher why a particular observation has received a particular value. In practice, these three terms are closely linked and are often used interchangeably, in the absence of a consensus on their precise definitions[9].

These principles are achieved through basically two strategies: either develop algorithms that are interpretable and explainable by their nature (including linear regressions, logistic or multinomial models, and certain types of deep neural networks, among others), or use interpretability techniques as tools to achieve compliance with these principles[10].

---

[6] GDPR (2018), Recital 71.

[7] Vilone et al. (2021). Doctora en Inteligencia Artificial, School of Computer Science, Technological University Dublin.

[8] Roscher et al. (2020). Data Scientist at the Technical University of Munich.

[9] Marcinkevics et al. (2020). Researcher at the Department of Computer Science, ETH Zurich.

[10] iDanae (2022). Chair in Big Data and Analytics (iDanae is a Spanish acronym for intelligence, data, analysis and strategy) created from a collaboration between Management Solutions and the Polytechnic University of Madrid (UPM) in the educational, scientific and technical fields. The Chair aims to promote knowledge creation and dissemination as well as technology transfer, and to foster R&D&I in Data Analytics.

XAI deals both with the techniques to try to explain the behavior of certain opaque models ("black box") and the design of inherently interpretable algorithms ("white box")[11].

XAI is essential for AI development, and therefore for professionals working in this area, due to at least three factors:

- It contributes to building confidence in making decisions that are based on AI models; without this confidence, model users might show resistance to adopting these models.
- It is a regulatory requirement in certain areas (e.g. data protection, consumer protection, equal opportunities in the employee recruitment process, regulation of models in the financial industry).
- It leads to improved and more robust AI models (e.g. by identifying and eliminating bias, understanding the relevant information to produce a certain result, or anticipating potential errors in observations not included in the model's training sample). All of this helps to develop ethical algorithms and allows organizations to focus their efforts on identifying and ensuring the quality of the data that is relevant to the decision process.

Although the development of XAI systems is receiving a great deal of attention from the academic community, industry and regulators, it still poses numerous challenges.

This paper will review the context and rationale for XAI, including XAI regulations and their implications for organizations; the state of the art and key techniques of XAI; and the advances and unsolved challenges in XAI. Finally, a case study on XAI will be provided to help illustrate its practical application.

---

[11] Sudjianto et al. (2011). Head of Model Risk at Wells Fargo..