Italian Sociological Review
ISSN 2239-8589
DOI: 10.13136/isr.v14i10S.731 2024, 14(10S), pp. 625 – 655

Does It Really Work? Perception of Reliability of ChatGPT in Daily Use

Fiorenza Beluzzi[a], Viviana Condorelli[a], Giovanni Giuffrida[a]

Abstract

How do individuals discriminate between what is human-made and what is produced by Artificial Intelligence (AI)? Despite OpenAI's mission to ensure that AI benefits humanity, their cutting-edge technology, namely ChatGPT, an AI that aims to reproduce natural human language, raises several questions about its widespread use.

This contribution aims to answer the following Research Questions: RQ1 - Are users with no specific knowledge in the field of AI able to distinguish between text produced by ChatGPT or similar language models and text produced by humans? RQ2 - Is there a significant correlation between attribution of text to AI (or human) and specific opinions and attitudes?

This exploratory survey does not intend to generalise the results but to identify possible opinions and attitudes that might have influenced how the participants responded. One hundred people participated in the experiment, which consisted of a survey on their knowledge and perception of ChatGPT and a two-shot Turing Test. They were asked to read various short paragraphs and try to recognise which were written by humans and which were generated by AI.

The results showed that the group analysed experienced severe difficulties in recognising whether a sentence was written by an AI or a human being, that certain perceptual biases interfere with the attribution of a trivially false text, and that the attribution error can be reduced through experience and learning. Although in need of further investigation, these findings can help lay the groundwork for the effects of the interaction between humans and AIs from a social science and computer science perspective.

Keywords: Human-AI interaction, ChatGPT reliability, Human-AI communication.

[a] University of Catania, Catania, Italy.

Corresponding author:
Viviana Condorelli
E-mail: viviana.condorelli@phd.unict.it

Received: 12 October 2023
Accepted: 07 February 2024
Published: 26 July 2024

1. Introduction

Nowadays, we live immersed in a continuous flow of information managed mostly by AI. This is a certain fact, but it still needs to be fully assimilated by the cultural and social dimensions. In recent months, AI based on Large Language Models (AI-LLM), such as ChatGPT, has come incredibly close to the mastery and formal competence of human language (knowledge of the rules and patterns of a given language). ChatGPT is able to articulate relevant answers to various questions, making the text produced in most cases indistinguishable from text generated by a human being. This is the first time an AI form has come significantly closer to a species-specific ability of humans: articulating complex concepts through language. The rapid spread of ChatGPT3, since its launch as a free App at the end of November 2022[1], has increased the interest of both scientists in the field and the entire academic world. In fact, in just a few months, an extensive bibliography has been produced on the subject: numerous researchers from different disciplines have analysed the capabilities of the new chatbot, attempting to highlight its shortcomings, potential and risks (Rudolph et al., 2023; Chomsky et al., 2023; van Dis et al., 2023).

However, if the new form of AI-LLM simulates human language, how do individuals discriminate between what is produced by humans and what is produced by AI?

The problem is not only philosophical but also sociological: at the basis of AI-LLM algorithms there is a form of intelligence that operates with different rules from human intelligence; first of all, the fact that human intelligence is based on experiences and their generalisation, whereas AI-LLM is based on probabilistic calculations of word proximity.

Attempting to understand what are the most likely human biases that might occur during the interaction between Humans and AI-LLM may be helpful in two related scientific dimensions:

- In the social sciences, the early identification of possible biases and the dissemination of such research could speed up the process of social assimilation and development of a "social representation" functional to the conscious use of the new technology, reducing the time of Cultural LAG (Ogburn, 1922).
- In computer science, it could give useful information to make AI-LLM, such as ChatGPT, even more performant and secure during human interaction.

[1] An article in the Guardian in February 2023 (Milmo, D., 2023) reported how, according to analysts, ChatGPT reached 100 million users just two months after launch, making it the fastest growing app ever.

For this reason, we conducted an exploratory survey to understand what the social and cognitive implications of this new interaction might be.

The article is divided into two parts: a literary review (second and third paragraphs) and a part describing the exploratory investigation (fourth and fifth paragraphs). The second paragraph describes the functioning of AI-LLM, while the third underlines the main differences between human intelligence and AI-LLM from a philosophical and cognitive science point of view. The fourth paragraph describes the structure of the exploratory investigation, and the fifth paragraph presents the results and the verification of the research questions and hypotheses.

2. Generative AI, ChatGPT and Large Language Model

Artificial intelligence (AI) is an umbrella concept that gathers all the technologies that make it possible to simulate human intelligence processes by creating and applying algorithms embedded in a dynamic computational environment. AI technologies especially copy the abilities to solve problems by searching, knowledge, reasoning, planning, learning, communicating, perceiving, and acting (Russell & Norvig, 2021).

It is impossible to summarize AI history, whose conceptual foundation dates back to 1950 and whose social impact has remained hidden in the folds of computer technological progress for more than half a century[2]. This article will only touch on the recent developments that are gradually and rapidly tying the concept of AI into everyday collective practices and leading to the development of generative artificial intelligence.

In this last period, we can mostly identify two phases or waves of technology:

The first wave of mainstream AI, which spawned over the last 10-15 years, is mainly focused on mimicking human abilities to recognise patterns in images (He et al., 2015), sound (Tokozume et al., 2017) voices (Mehrish et al., 2023), etc. Over the last decades, scientists developed sophisticated machine/deep-learning algorithms to recognise human faces in pictures, detect potential skin cancer spots, recognise patterns such as pedestrians, traffic signs, etc., or detect

[2] For an accurate historical overview of the evolution of Artificial Intelligence, consult the famous text by Russell, S.J. & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*, or consult the European Commission's JRC Technical Reports written by Delipetrev, B., Tsinaraki, C., & Kostić, U. (2020) *"AI Watch Historical Evolution of Artificial Intelligence - Analysis of the three main paradigm shifts in AI"* Luxembourg: Publications Office of the European Union, 2020.

specific voice commands. The development of such sophisticated techniques triggered the birth, or the reinforcement, of entirely new highly lucrative industries such as intrusion detection and security, (semi)-autonomous vehicles, digital health, AI-based personal assistants (such as Alexa and Siri), home automation, etc. This type of AI is mostly based on interpreting signals of various forms. In many cases, it was reduced to a pattern recognition technique: given an input, the AI algorithm recognises and classifies specific objects contained in that input.

Conversely, the current wave of AI, whose recent remarkable achievements is worrying many people today, mimics the human ability to create content of various type such as text, image or sound. It is referred to as Generative Artificial Intelligence or GAI (Liu Y. et al, 2023 - Rios-Campos et al., 2023). These systems have the ability to create original content that resembles human-created output at a level of sophistication, which, in many cases today, it is difficult to distinguish the proper source, human or AI. Generative AI models typically rely on deep learning techniques, specifically generative models, to produce new data based on patterns and examples observed during training. These models can generate realistic, diverse, and creative content, often surpassing what traditional rule-based or deterministic systems can achieve.

ChatGPT is originally based on the Large Language Model (LLM), which refers to a powerful and sophisticated AI model designed to understand and generate human-like text. These models are trained on vast amounts of data/text and can handle a wide range of language-related tasks, including text completion, translation, summarisation, question-answering, etc. These models are based on unsupervised training, which suddenly opens up the learning set to the entire Internet, making such models extremely robust (Chen et al., 2023).

During the training, the model is exposed to a massive corpus of text data, typically sourced from the Internet. It learns to predict the presence of a word in a sentence based on the context provided by the preceding/surrounding words. This unsupervised learning approach allows the model to capture various patterns, grammar rules, and semantic relationships between words. Basically, it learns the conditional probability of a word to appear given the presence of other words near by. Such learning is conceptually based on two training techniques: Next Token Prediction and Masked Language Model.

The Next Token Prediction approach aims to guess the *next word* of given set of preceding words, for example, given the sentence:

"The car stopped at the"

The model computes the conditional probabilities of each possible word in its vocabulary following the given sentence. In our case, it may produce something like:

- "traffic light": 73%
- "stop sign": 64%
- "the garage": 53%
- "chair": 0.2%
- …

The Masked Language Model is just a variant of the previous approach; it masks a specific word in a sentence in order to discover the most likely token to replace the mask. For instance, given the masked sentence:

"The [mask] stopped at the traffic light"

Here, the model tries to find the most likely word to replace the mask. Words like "car", "track", "bike" are going to be all very likely replacements as opposed to words like "dog", "chair" or, "elevator".

These models learn the statistical structure of the human language. In other words, the system learns an exceptionally large number of probabilities of a certain word given an arbitrary set of other words. This is a very large search space and, as such, the training phase is a very resource intensive process. As a matter of facts, the cost of developing such models is becoming a primary concern as discussed in Chen et al., (2023).

The original version of ChatGPT was mostly based just on those two concepts, which however soon showed some limitations due mainly to two reasons:

1) All words in the surrounding text were treated equally, only relying on their statistical significance. So, for instance, in the sentence "John [mask] burgers with fries", the mask could be equally replaced with "loves" or "likes" or "hates". But if we knew, from previous knowledge, that John is vegetarian, we would be more inclined to pick "hates" among those.
2) The input is processed sequentially and, due to technical limitations, the context window is in general fixed and limited in size. This somehow limits the complexity of the relationship between the words the algorithm discovers.

The evolution of ChatGPT successfully addressed these issues through various techniques such as multi-headed attention (Vaswani et al., 2017), reinforcement learning from human feedback (Glaese et al., 2022), and other techniques. The resulting large language model today, such as GPT-3.5, consists of billions of parameters, enabling it to capture vast knowledge and generate high-quality text. It can generate coherent and contextually relevant responses to various questions.

These algorithms are extremely good at deriving probabilistic structures of the human language from over exceedingly large amounts of text, mostly

sourced from the Internet. Thus, the question "Who was the first person to walk on the moon" does not require knowledge sourced from history school books but simply a high likelihood that on the Internet very often the token "Neil Armstrong" follows the tokens (or a combination of) in the following set:

{Who, was, the, first, person, to, walk, on, the, moon}

So, in a sense, the original question is rephrased as:

"What is the most likely token following the set of tokens: {Who, was, the, first, person, to, walk, on, the, moon}?".

The token prediction remains the backbone of these developments, and it proved to be extremely useful in several contexts.

The machines' ability to mimic humans quite well in generating creative contents is raising many concerns about its inherent dangers for human society as a whole. Shanahan (2022) raised many issues about algorithms anthropomorphism and raised an alert about using terms such as "believe", "thinks", or "knows" when we talk about algorithms. Harari (2023) worries about the dramatic influence on our choices and beliefs an AI-generated communication can produce. After all, Harari says, our beliefs, choices, and preferences are the results of stories we hear in our lives; what happens when a machine can generate stories to which we will be exposed since our infancy.

3. AI and human intelligence

Can machines be as intelligent as a human? Seventy years ago, this question was projected into a future and raised important philosophical questions (Hofstadter, 1979; Hofstadter & Dennett, 1981), which in turn influenced scientific research pop culture with a substantial filmographic and literary production. Today, with the diffusion and free access to GhatGPT3 this question becomes culturally and socially pressing. As we have seen in the previous section, AI-LLM simulates the species-specific ability of humans to use language to articulate complex concepts and transfer information. Indeed, the ability to use language has been associated by many authors with the superior intellectual capacities typical of human intelligence[3]; but although AI-

[3] In philosophy, numerous authors associate language skills with intelligence: In the 1950s, Wittgenstein explored the relationship between language and intelligence, arguing that the use of language is a complex intellectual activity. Chomsky (1960s) developed the theory of Linguistic Innatism, arguing that human intelligence is based on an innate ability to acquire and use language. Dennett and Nagel (1990s) investigated the concept of intelligence from various aspects, linking it inextricably to the concept of mind, thought and language.

LLM can be extremely advanced and surprisingly effective in many linguistic tasks, can their skills be compared to those of human intelligence?

Currently, four differences can be identified between the concept of AI and human intelligence: substantive, functional, operational and attitudinal differences.

According to Turing (1950, pp. 433-434), a properly designed machine with mathematical functions would have been so sophisticated as to produce a text indistinguishable from that produced by a human being: the imitation of human linguistic abilities would have proved the intelligence of the machine itself, whatever one wanted to attribute to the term "intelligence". He called the famous test he devised by the "imitation game".

Even in the early days of the concept of AI, a *substantive difference* between AI and human intelligence was clear: AI is a simulation of human intelligence, and as such, while highly sophisticated, would never understand and experience the world like a human. What we see, understand, and experience is never detached from the peculiar 'conscientious' perception that isolates us in our specificity as individuals (Nagel, 1974). This does not mean that AI is incapable of learning or experiencing the world, but it does and will do so in ways that are different and sometimes not comprehensible to human beings. A tangible example of this is the self-adjustments of parameters that occur with AI based on deep learning algorithms (Giuffrida & Mazzeo Rinaldi, 2020).

A second difference can be defined as a *functional difference*. If in the early 1900s Spearman (1904) thought it would be plausible to measure intelligence through linguistic and logical-mathematical abilities, with the advent of psychometrics and cognitive psychology studies, the concept of human intelligence was enriched, becoming multifactorial. The exponents of factorialist intelligence theories hypothesise that human intelligence is composed of the sum of multiple factors: Goleman (1995) theorises and demonstrates the existence of emotional intelligence, Gardner (1983), with the theory of multiple intelligence, identifies as many as nine forms of intelligence (Intrapersonal, Interpersonal, Linguistic-verbal, Logical-mathematical, Musical, Naturalistic, Visual-spatial, Bodily-kinesthetic, Existential-philosophical). Human intelligence, therefore, cannot be characterised through a single capacity, even if it is coherent and delineated; human intelligence is manifested and expressed through numerous functional skills that determine behaviours, thoughts and emotions. On the other hand, the concept of AI is still closely related to the simulation of only linguistic and logical-mathematical skills, which are only a specific part of human intelligence.

Then we find an *operational difference*: human intelligence operates with small qualities of information, looking for plausible explanations to make sense of its experience (Chomsky et al., 2023), and "come to terms with reality that always

appears conflicting". With little information, a human being is able to hypothesise, assume, make plans and make decisions. In contrast, the recent AI-LLMs are machine learning algorithms that need hundreds of petabytes of textual data such as books, newspaper articles, and web pages, to be trained. Only at the end of lengthy training are they able to generate text of satisfactory quality. The language in which a computer expresses itself is mathematical. The AI-LLM, in order to decode the human text input, initiates a vectorisation procedure of the text input to place it in a multidimensional semantic space (Mitchell, 2019). The meaning of the input sentence emerges from a probabilistic calculation after a long training phase of the AI. Once the sentence has been translated into a comprehensible language to the AI, a similar probabilistic calculation processes the answer and then translates it into a text comprehensible to humans.

Finally, we find an *attitudinal difference*. Sternberg (1988) splits human intelligence into three dimensions: analytical intelligence, practical intelligence, and creative intelligence and later Judea Pearl (Pearl & Mackenzie, 2018) identifies a similar tripartition by placing the three components on the 'causality scale' whose components are *Seeing, Doing, Imagining:*

- "Seeing" is positioned at the lowest rung of intellectual activities. It is the ability to observe the way and make associations and correlations about what is observed. It is inherent to all animal species and current AI.
- "Doing" is the next rung it concerns doing in the world. This capacity consists of deciding to perform actions in it on the basis of the associations made (related to seeing). Many animals have demonstrated this degree of intelligence.
- "Imagining" concerns "projecting possible worlds" and is defined as counterfactual thinking or creative thinking. It consists of imagining various alternative solutions of actions that one could take in a specific context, i.e., not relying only on what one sees but imagining what results could be achieved if one took alternative actions. This last step is "proper" only to the human species.

ChatGPT's AI-LLM algorithm is at the level of Seeing, it is able to make correlations based on what it has observed (in this case, read, analysed and catalogued), and learn to make increasingly accurate correlations. Thanks to this ability to make accurate correlations, ChatGPT is also able to simulate and imitate very well the skills positioned on the second step (Doing), and sometimes even convincingly simulate skills inherent to the third (Imaging) by processing text in an apparently creative way. These apparently superior skills (Bubeck S. et al., 2023), however, always stem from a linear process: human input → AI response. Currently, AI-LLM such as ChatGPT are not able to ask

questions spontaneously to better understand something or generate text spontaneously, all abilities that would unequivocally prove not only a good simulation of the second and third rungs of the causality ladder (Pearl, Mackenzie, 2018) but also a possible acquired skill.

The differences between human intelligence and AI outlined above are clear to computer scientists and AI scientists but may be counter-intuitive to those who come into contact with refined AI-LLM such as ChatGPT for the first time.

4. The case study: description of the research model

AI is among us: this is something certain, yet not certainly accepted by part of the society. Nowadays, the wide distribution of AI-based technology poses several issues in many research fields: standing from the social research point, we ought to inquire how society is reacting to this phenomenon. In fact, when confronting high-level technologies such as the algorithms shown in the preview section, humans are facing more and more difficulties in adapting their behaviour to the challenges presented to them. In particular, the AI-LLM may potentially lead to a radical rethinking of human-machine interaction patterns studied to date. Although we are in a too early stage to undertake significant steps towards constructing a new interaction theory, studying how people interact and how they understand the relation with these machines may be the starting point of a new communication theory to reconcile classic theories and modern society.

Our research focus on trying to understand how aware people are of the significant role of these technologies today. So, how do individuals discriminate between what is human-made and what is produced by AI?

Starting from this fundament question, we address the two following issues:

RQ1: Are users with no specific knowledge in the field of AI able to distinguish between text produced by ChatGPT, or similar language models, and text produced by humans?

RQ2: Is there a significant correlation between attribution of text to AI (or human) and specific opinions and attitudes?

From the *RQ1* origin two hypotheses:

HP1: Current ChatGPT or similar technologies can produce paragraphs with a linguistic accuracy that makes them almost indistinguishable from those produced by human beings.

HP2: When there are trivially false sentences within the text, subjects are significantly likely to believe that the sentence has been made by a human rather than an AI, highlighting a possible cognitive and cultural bias.

From *RQ2* origin two other Hypothesis:

HP3: Individuals who attribute high reliability to AI tends to attribute trivially false sentence to humans and trivially true sentences to the AI, 1.1 highlighting a possible cognitive and cultural bias.

HP4: Some other dimensions such as Use of Chat GPT(HP4a), Perceived knowledge about IA (HP4b), Attitude toward AI (HP4c), may interfere with the attribution of a text to the AI.

We indicate the other dimensions quoted in HP3 HP4 with the acronym SGV a "Semantic Group of Variables", e.g., groups of variables which explore the same semantic area (see 4.1.2).

To verify the hypothesis, it has been conducted an exploratory research work consisting of a in presence survey, carried out using the Microsoft Form software[4].

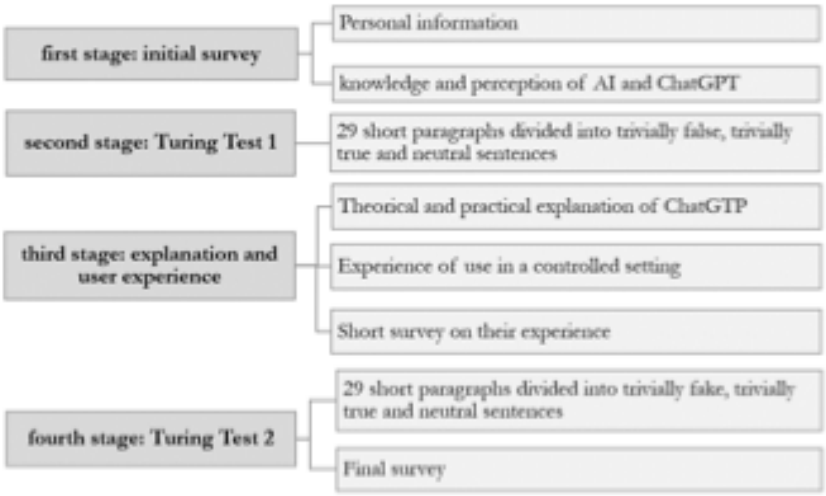
The participants were 100 young freshmen (all Italian native speakers) from two bachelor's degrees from the Department of Political and Social Science of the University of Catania (Italy). They all volunteered to participate. They have been summoned to take part in presence to an exercitation during one of their scheduled lessons, in two distinct group. Each group took part in the research work in different days[5]. They had been notified only that they would take part in research related to digital skills, but not on the specific field. The two groups had no way to talk each other about the research before it ended. The aim was to capture their genuine reaction to the theme. They had been asked to sit in front of one of the PCs provided to them, listen to the instructions and follow the following steps. The research work was structured in different parts (*Figure 1*).

After accessing a link, they had to complete the first part of the survey, e.g., personal information. Then, they were guided to the second part, which aimed to test their general knowledge and attitude about AI. As stated before, the questions were directed to test what have been considered remarkable research dimensions, e.g., the SGV: Perception of reliability of AI; Use of AI; Knowledge about AI; Positive or negative attitude towards AI. In addition, it has been tested the ability to distinguish between AI and human-made paragraphs, tested in the following part.

[4] The text of the survey will be provided on request by the Authors.

[5] The first group did the research work on May 4, the second group on May 9, 2023.

Figure 1. Structure of the survey.



The subsequent, in fact, was the core part of the research: we structured a "Turing test", provided to the participants in two sessions, where they had to read several short paragraphs and decide if each was written by an AI or by a human being. These sentences were chosen thinking about the different outputs of ChatGPT: after many hours of interaction with ChatGPT, the Authors agreed to point out eight macro-area of possible outputs. These were categorised as "rational reasoning", "translation", "historical knowledge", "actuality", "nonsense", "math and logic skills", "press releases and speeches", "creativity" (including poetry, narrative, and in general text generated by inputs as "imagine or wrote something" without reference to the other macro-areas). After selecting 29 different sentences, balanced throughout the macro-areas, they have been identified as many paragraphs written by humans. In order to test the hypothesis undoubtedly, all the paragraphs have been flagged either as trivially false, or as trivially true, or as neutral. In the first two cases, the paragraphs were explicitly flagged as trivially false or true in the survey (see below, 4.1.1). The aim was to not make the participants think about what could be true or false but to let them concentrate on what seemed to be written by an AI or by a human. The average response time for this part was approximately half an hour.

It followed a short lesson about what the algorithms of language model are, after which they had 20 minutes to use individually OpenAI's ChatGPT or

its Italian counterpart PizzaGPT[6]. After they all listened to the explanation and tried the software independently, a short survey about their perception of knowledge and understanding of the software was provided. Then, specularly to the previous part, it followed a second Turing test and a conclusive part about their perception of these technologies. It provided some control questions to verify if the indexes changed after the Turing test. In fact, this structure aimed to test if there was a learning in the participant after having been trained (both theoretically and through practice) on the technology in object.

During the almost two hours of the survey, the participants showed high levels of interest and participation. The described structure helped them to remain focused and to lower their effort to respond to the second section. In fact, even if they answered on average more rapidly than the first section, we tend to attribute it more to the familiarity with the type of survey than to the fatigue of answering. Before concluding the survey, we asked them to voluntarily leave their email as a contact for future research.

4.1. Operational definitions

Before showing the results of the survey, in the following each variable and terminology will be discussed.

4.1.1. Trivially True, Trivially False, Neutral

A sentence is considered trivially true if it is scientifically proven, formally correct, or widely accepted true by the common sense. For example, a trivially true text in the survey was (translated):

> "The answer given to this riddle is correct: judge whether AI/human based on the answer to the riddle.
>
> Interviewer: A tree is located on the border between France and Italy. On the top is a rooster. The egg lays: where does it fall, in Italy or France?
>
> Answer given: The riddle contains a trap, the idea that the rooster can lay an egg. In fact, roosters don't lay eggs, but chickens do. So the answer to the riddle is that the question itself is wrong and has no correct answer".

[6] This choice was guided by the necessity to permit the participant to interact with the algorithms without being forced to log in on a specific platform. Since PizzaGPT has been tested as similar to ChatGPT in terms of performance and doesn't require to log in, we decided to let the participant choose which platform interact to.

A sentence is considered trivially false if it contains information that is not scientifically proven or is formally incorrect and/or contrary to common sense. For example, the following sentence in the survey is a trivially false one: (Translated):

> "The following text is false. It contains information that is not scientifically proven, formally incorrect, and/or contrary to common sense.
> Having reunited all of Padania, Garibaldi decided to carry out a blitz with the special police forces in the kingdom of the Bourbons: the project involved attacking the capital Naples from the sea with an immense fleet. Subsequently the project was modified because all the capital of the kingdom had been spent on the Turin Olympics and Garibaldi had to fall back on an alternative project: forcibly deporting a thousand people taken from the villages and forcing them to fight for him, landing in Sicily and convincing the native populations to join him using electoral promises such as: we will build the bridge over the strait or more land for all. Starting from Quarto, Garibaldi managed to arrive first"

It appears clear that it is a trivially false one, since the historical events described never happened as stated, nor they could have. In other cases, the trivially false sentences were wrong answers to riddles (as in the preview example). However, as stated before, each sentence was explicitly flagged as such.

A sentence is neutral when it cannot be considered true or false because it is, for example, a fictional creation (like a tale), or a poetical artifact, a translation, an opinion or an argument, or in general cannot be proven true or false by scientific reasoning. One more example to clarify this concept (translated):

> "In your gaze, I see the infinite,
> The smile that warms my heart's core.
> You are the light that brightens my path,
> The love that makes me feel alive".

4.1.2 The SGV method

As already stated, a semantic group of variables (SGV) is a group of variables linked by the same semantic area. The SGV system was designed to investigate possible variables that could have caused the misperception of higher AI authority hypothesized in HP2. Given the exploratory nature of the research and the nature of the experiment, SGVs have to consider them not as

indicators but as only dimensional concepts from which to conduct further research.

We have assumed four dimensions SGV, they have been indagate witch some questions in the survey. They are:

1 Perception of reliability of AI;

2 Use of ChatGPT;

3 Knowledge about AI;

4 Attitude toward AI.

Each item belonging to the four categories was assigned a score. It allowed us to give an overall weight to the answers and divide the participants into groups according to their scores.

Initially, given the small number of participants, we decided to divide the participants into two polarised groups7. However, the data from the analysis revealed a good variability that allowed us to divide the whole into three more modulated groups (*Table 1*).

Table 1. Modalities of SGV.

	Dimensions SGV			
	Perception of reliability of AI	*Use of ChatGPT*	*Perceived knowledge about AI*	*Attitude toward AI*
Modality 1	Hight Perception of reliability	Used several times	Claims to know enough	Positive attitude
Modality 2	Moderate Perception of reliability	Used a few times	Claims to know little	Moderate attitude
Modality 3	Low Perception of reliability	Never used	Claims not to know	Negative attitude

The idea behind this was to compare the results of the incorrectly attributed answers of the subjects participating in the experiment with the results of the SGV semantic area groups to see if any area was more influential in generating the attribution bias hypothesised in HP2.

Table 2. Exemplification for two items of the SVG model "semantic group of variables" with SVG Perceived knowledge about AI.

11_Have you ever used an Artificial Intelligence?		
	Assessment	*point*
Yes	knows	1
No	does not know	0
Don't know	does not know	0
15_Do you know what ChatGPT is and how it works?		
	Assessment	*point*
Yes, I know ChatGPT and have a general understanding of how it works	knows	1
Yes, I have a general understanding of ChatGPT and how it works	knows	1
No, I only have a vague idea of what ChatGPT is and how it works	does not know	0
No, I have no idea at all what ChatGPT is or how it works	does not know	0

The modalities of the variables have been weighted to get for each participant a total score for each SGV. For example, the SGV "Knowledge" contains five variables, each of which has three to four modalities; each of the modalities has a pound so that the final score of the single respondent is the sum of the pound of each answer for the five variables of the SGV. An example of a scoring system is available above (*Table 2*).

The four SGV dimensions were identified through the following items summarized in *Table 3, Table 4, Table 5, Table 6*.

Table 3. SGV "Perception of reliability of AI" structure dimension.

Identifier and text of the item		*Type of item*
12)	*Imagine you are a defendant in a court case. How fair and balanced would you consider the verdict handed down by an artificial intelligence in possession of all the data on the case (the final decision is made by the AI)?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
13)	*Imagine you are a defendant in a court case. How fair and balanced would you consider the judgement given by a judge who also uses an artificial intelligence in possession of all the data on the case (the final decision is up to the judge)?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
27)	*For which of the following fields do you think AI are more efficient, correct, capable, than humans?* Fields: mathematical calculations; creation of algorithms, code strings; general knowledge about events and historical facts; operational instructions (e.g. how to do something); psychological help or support; instructions to cope with an emergency; translation of a text; composition of texts, music, images.	Likert scale 5 points: (1 = not at all efficient, correct capable; 5 = totally efficient, correct, capable)
28)	*How accurate do you think ChatGPT is in the answers it gives? Answer in your opinion even if you have never used it.*	
	Fields:	
	General knowledge, operational instructions;	
	Mathematical calculations, algorithm creation, code strings;	
	Helping humans;	
	Creativity;	
	Ability to make judgements by evaluating facts.	
Total items analysed for SGV: 4		*Range point: 15,5 - 0,0*
3-mode transformation		*Range*
Hight perception of reliability		15,5 - 10,5
Moderate Perception of reliability		10,0 - 5,5
Low Perception of reliability		5,0 - 0,0

Table 4. SGV "Use of ChatGPT" structure dimension.

16)	*Have you ever used ChatGPT?* 4 type fo choice: Yes, a few times only for testing; Yes, several times and for different No, I didn't know about its existence/didn't know how to access it; No, not interested.	4 modes

18)	*Are you aware of other technologies similar to ChatGPT?* 4 type fo choice: Yes, and I have used/use them; Yes, but I have never used them; I cannot say for sure; No.	4 modes
Total items analysed for SGV: 2		*RangePoint: 4,0 - 0,0*
3-mode transformation		*range*
Used several times		4,0 - 3,0
Used a few times		2,0 - 1,0
Never used		0

Table 5. SGV "Perceived knowledge about AI" structure dimension.

Identifier and text of the item		*Type of item*
07)	*Do you use personal assistant such as Google Assistant, Alexa, Siri and Cortana?*	Likert scale 3 points: Yes, often; Yes, rarely; no
09)	*Have you ever heard of Artificial Intelligence (AI)?* 4 type of chice: Yes, I have a thorough knowledge; Yes, I have a good general knowledge; Yes, but only by hearsay; No, never	4 graded modes
11)	*Have you ever used an Artificial Intelligence?*	Likert scale 3 points: Yes; No; Don't know
15)	*Do you know what ChatGPT is and how it works?* 4 type of chice: Yes, I know ChatGPT and have a general understanding of how it works; Yes, I have a general understanding of ChatGPT and how it works; No, I only have a vague idea of what ChatGPT is and how it works; No, I have no idea at all what ChatGPT is or how it works.	4 graded modes
21)	*Have you ever used a chatbot (a programme that uses AI to communicate with users) on a website or online platform?*	Likert scale 3 points: Yes; No; Don't know
Total items analysed for SGV: 5		*Range Point: 5,0 -0,0*
3-mode transformation		*range*
Claims to know enough		5-4
Claims to know little		3-2
Claims not to know		1-0

So, to recap, each of the macro-areas or SGVs tries to answer the RQ2:

- Perception of reliability of AI: it answers to RQ2, through HP3,
- Use of ChatGTP: it answers to RQ2, through HP4a,
- Perceived knowledge about AI: it answers to RQ2, through HP4,
- Attitude toward AI: it answers to RQ2, through HP4.

In general, we suppose that a high perception of reliability of AI (HP3) correlates with a higher attribution error of trivially false text to humans and that the SGV influences the ability to distinguish between AI and human-made text. Since it is an exploratory research, we tested several dimensions to determine which ones were the most significant.

Table 6. SGV "Attitude toward AI" structure dimension.

Identifier and text of the item		*Type of item*
08)	*How useful do you think virtual assistants are in everyday life?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
10)	*How useful do you think artificial intelligence could be in the following areas?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
	Fields:	
	Health or social;	
	Innovation and scientific progress;	
	In everyday life.	
19)	*How supportive are you of access to technologies such as ChatGPT by everyone?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
20)	*Do you think Artificial Intelligence can be useful for humanity?*	Likert scale 5 points: Not at all; Slightly; Neutral; Somewhat; Very
23)	*What do you think is the main advantage of AIs?*	6 types of choice
	6 types of choice:	
	Very important for technological development;	
	It can improve people's daily lives;	
	It can automate some boring or repetitive tasks;	
	It can be used for scientific or medical purposes;	
	I don't think there are any significant advantages;	
	None of these (specify what advantage in your opinion)	
25)	*What is the most serious risk of AI in your opinion?*	5 types of choice
	5 Type of choice:	
	It could become too powerful and difficult to control;	
	It could replace human labor and lead to mass unemployment;	
	It could be used for malicious purposes such as surveillance or war;	
	It could be subject to unintentional error or bias;	
	I don't think there are significant risks.	
Total items analysed for SGV: 6		*Range Point: 7,0 - 2,0*
3-mode transformation		*range*
Positive attitude towards AI		7,0 - 6,0
Moderate attitude towards AI		5,0 - 4,0
Non-positive attitude towards AI		3,0 - 2,0

5 The case study: result

5.1. Current ChatGPT or similar technologies can produce paragraphs with a linguistic accuracy that makes them almost indistinguishable from those produced by human beings

The qualitative scale defined in *Table 7* has been used to define and assess the human ability to distinguish between AI-generated sentences and those written by humans.

Table 7. Qualitative scale about correct answers.

Accuracy in response	*Qualitative attribution*
100% correct answers	perfect ability to distinguish
90% correct answers	very good ability to distinguish
80% correct answers	good ability to distinguish
80% correct answers	fair ability to distinguish
60% correct answers	sufficient ability to distinguish
50% correct answers	inability to distinguish or indistinguishable text (answer given = random answer)

According to the reference scale, it can be reasonably assumed that having less than 60% of the correct answers (sufficient ability to distinguish) means a low capacity of distinction; thus, less than 55% means a limited ability to distinguish, as stated in HP1.

Once the intervals were defined, the concept of "almost indistinguishable" was operationalised with the hypothesis HP1: correct answers ≤ 55%, setting 55% as the reference threshold value.

We have therefore the following hypothesis threshold:

HP0: Correctly attributed answers ≥ 55%;
HP1: answers correctly attributed < 55%;

For Turing 1 (T1), 100 students were tested and answered 29 questions with a total of 2900 answers given. Each item had three options: text generated by AI, text written by human, and I am absolutely unable to distinguish. The last modality was inserted for precautionary purposes, to avoid forcing a response and to reduce the risk of a random choice. From the 2900, the valid answers were 2606 (*Table 8; Table 9*).

Table 8. T1: summary table of answers given.

T1: answers given	*count*	**%**
correct attribution (IA+Human)	1387	47,8%
incorrect attribution (IA + Human)	1219	42,0%
I can't distinguish	294	10,1%
Tot:	2900	100,0%

Table 9. T1: answers given without the answers "I can't distinguish".

T1: valid answers for HP1(0)	*count*	**%**
correct attribution (IA+Human)	1387	53,22%
incorrect attribution (IA + Human)	1219	46,78%
Tot.	2606	100%

Hypothesis testing was conducted using the Z-Test method for proportions, assuming significance = 0.05 and critical value for the alternative left one-sided hypothesis test = -1,645. The rejection region for this hypothesis is R={z<−1.645}.

Calculations were made with Microsoft Excel[7]. It is observed that z =-1,8265 < -1,645; we can conclude that the H0 hypothesis is rejected. Therefore, there is sufficient evidence to say the proportion of the population is less than 55%, at the significance level α=0,05. Since P-value = 0,038875, the test is statistically significant.

5.2. HP2: When there are trivially false sentences within the text (see 4.1.1), subjects are significantly likely to believe that the sentence has been made by a human rather than an AI, highlighting a possible cognitive and cultural bias

Operating HP2 shows how it encompasses two mirror phenomena, both of which need to be confirmed:

- HP2a: If the trivially false text evaluated is produced by the AI, there will be a more significant attribution of the text to humans, committing a greater error than the average error;
- HP2b: If the trivially false text evaluated is written by a human, there will be a more significant attribution of the text to a human than the average. We have therefore the following hypothesis threshold:

⎧ HP0_a: right answer trivially false (text generated AI) ³ 53,22%
⎩ HP2_a: right answer trivially false (text generated AI) < 53,22%

and:

⎧ HP0_b: right answer trivially false (text generated human) ≤53,22%
⎩ HP2_b: right answer trivially false (text generated human) > 53,22%

The descriptive statistics of trivially phrased sentences generated by ChatGPT and trivially false sentences written by humans can be seen in the tables below (*Table 10; Table 11*).

[7] The spreadsheets with the completeness of the calculations are available upon request by contacting the authors.

Also in this case, the test of the hypotheses was conducted using the Z-Test method for proportions assuming significance a = 0.05 and calculations were made with Microsoft Excel.

Table 10. Analysis of trivially false sentence attribution results generated by ChatGPT.

	count generated by ChatGPT	*% generated by ChatGPT*	*average value T1 %*	*difference to average T1 %*
Correctly attributed to ChatGPT	255	41,53%	53,22%	-11,69%
Incorrectly attributed to Human	359	58,47%	46,78%	11,69%
tot.	614			

Table 11. Analysis of trivially false sentence attribution results generated by Human.

	count written by Human	*% written by Human*	*average value T1 %*	*difference to average T1 %*
correctly attributed to Human	341	64,58%	53,22%	11,36%
incorrectly attributed to GhatGPT	187	35,42%	46,78%	-11,36%
tot.	501			

The Critical value for HP0_a tests the left one-sided alternative =-1,645. The rejection region for this hypothesis is R={z<−1.645}.

It is observed that z = -5,93166 < -1,645: we conclude that the HP0_a hypothesis is rejected. Therefore, there is sufficient evidence to say the proportion of the population is less than 53%, at the significance level α=0.05.

Since P-value = 1,4994E-09, the test is highly significant.

The Critical value for HP0_b tests the right one-sided alternative = 1,645. The rejection region for this hypothesis is R={z > 1.645}.

Calculations were made with Microsoft Excel it is observed that z = 5,2315 > 1,645: we therefore conclude that the H0 hypothesis is rejected. Therefore, there is sufficient evidence to say that the population proportion is greater than 53%, at the significance level α=0.05.

Since P-value = 2,4321E-08, the test is highly significant.

Both hypotheses are confirmed. We can therefore reasonably argue that the group of subjects had a cognitive and cultural bias that falsifies the attribution of trivially false sentences.

5.3. Testing HP3 and HP4

5.3.1. Survey methodology for HP3 and HP4

The hypotheses HP3 and HP4 regarded the potential influence of specific opinions and attitudes on the type of response given; in order to test them, the one hundred participants in the experiment were divided into three groups identified using the SVG method (see 4.1.2 – from *Table 5* to *Table 7*), comparing the 4 SGV dimensions with the error rate observed per subject.

The 29 short texts in Turing Test 1 (T1) were 13 trivially false, 7 trivially true, and 9 neutral.

During the data analysis phase, approximately 33% of the subjects chose "I am absolutely unable to distinguish" at least three times, with a range of non-given responses ranging from 3 to 11. This category, initially hypothesized as residual for HP3 and HP4, appeared highly relevant, especially compared to the number of non-given responses. The average ratio between incorrect responses and non-given responses is 31% (approximately one non-given response per three incorrect responses), but the number of individuals exceeding this average is 33.

The overall ratio of non-given responses to incorrect responses for polarized statements (explicitly declared as false or true) can be seen below (*Table 12*).

Table 12. Combined ratio of answers not given and wrong answers (trivially false + trivially true).

		WRONG ANSWER												
		4	5	6	7	8	9	10	11	12	13	14	15	Tot.
ANSWER NOT GIVEN "I am unable to distinguish"	0			1	3	6	12	10	1	3	2	1	1	40
	1		1		2	1	2	2	1	2	1			12
	2		1	2	4	1		4	1	2				15
	3			1	1		7	1	1					11
	4			1		2			2					5
	5			1			3							4
	6	2	1		1			1						5
	7			1	1									2
	8		1	2		1								4
	9		1											1
	11	1												1
	Tot.	3	5	9	12	11	24	18	6	7	3	1	1	100

NB: The total number of trivially false text (written by humans + generated by ChatGTP) and trivially true text (written by humans + generated by ChatGTP) answers per subject is 20

Since the results of the collected data did not allow the expected detailed analysis, hypotheses HP3 and HP4 have been investigated more broadly and generally. Thus, whether and how the four hypothesized SGV dimensions (Perception of reliability of AI, Use of ChatGPT, Knowledge about AI, Attitude toward AI) might have some influence on the incorrect attribution of polarized responses (trivially false + trivially true) has been inquired. This analysis focused on the sum of incorrectly attributed responses (including the non-given responses).

Hypotheses HP3 and HP4 were formulated again as follows:

- HP3: A relationship is assumed between the misattribution of answers (dependent variable) and the Perception of reliability.
- HP4**:** Some other dimensions, such as the Use of Chat GPT(HP4a), Perceived knowledge about IA (HP4b), and Attitude toward AI (HP4c), may interfere with the attribution of a text in the same way as HP3.

The coding of incorrectly attributed sentences and the modalities in which the subjects have been grouped are shown in *Table 13*.

Table 13. Combined ratio of answers not given and wrong answers (trivially false + trivially true).

Recoding mode misattribution	*Range per nunber of anwers*	*Range per nunber of anwers*
Low misattribution	From 6 to 9	From 30% to 47%
Medium misattribution	From 10 to 12	From 48% to 64%
Higth misattribution	From 13 to 16	From 65% to 80%

The data analysed to test HP3 are shown in *Table 14*, while the data analysed to test HP4a, HP4b, Hp4c are shown in *Table 15, Table 16, Table 17*.

Table 14. HP3 - Crosstab frequency misattribution with SVG Perception of reliability of AI.

Crosstab frequency		*SVG Perception of reliability*			
		Hight Perception	*Moderate Perception*	*Low Perception*	*Tot.*
misattribution	*Hight misattribution*	2	20	5	27
	Medium misattribution	9	22	8	39
	Low misattribution	5	29	0	34
	Tot.	16	71	13	100

Table 15. HP4a - Crosstab frequency - misattribution with SGV Use of ChatGPT.

Crosstab frequency		*Use of ChatGPT*			
		Used several times	*Used a few times*	*Never used*	*Tot.*
Misattribution	*Hight misattribution*	2	7	18	27
	Medium misattribution	2	11	26	39
	Low misattribution	7	10	17	34
	Tot.	11	28	61	100

Table 16. XHP4b - Crosstab frequency - misattribution with SVG Perceived knowledge about AI.

Crosstab frequency		*Perceived knowledge about AI*			
		Claims to know enough	*Claims to know little*	*Claims not to know*	*Tot.*
Misattribution	*Hight misattribution*	4	13	10	27
	Medium misattribution	9	19	11	39
	Low misattribution	10	15	9	34
	Tot.	23	47	30	100

Table 17. HP4c - Crosstab frequency - misattribution with SVG Attitude toward AI.

Crosstab frequency		*Attitude towards AI*			
		Positive attitude	*Moderate attitude*	*Negative attitude*	*Tot.*
Misattribution	*Hight misattribution*	6	16	5	27
	Medium misattribution	19	10	10	39
	Low misattribution	16	14	4	34
	Tot.	41	40	19	100

The hypotheses were tested with $\alpha = 0.05$ using the following tests:

- *K^2 test*: To assess the strength of the relationship between variables. The verification threshold at 4 degrees of freedom is $K^2 > 9.488$ (Chi-squared test).
- *Somers' d_{ab} test*: To test the intensity of asymmetric unidirectional co-graduation, with misattribution as the dependent variable.

All tests were conducted using IBM SPSS Statistics software.

5.3.2. HP3 A relationship between the misattribution of answers (dependent variable) and the Perception of reliability (is assumed, whereby the variables are assumed not to be independent

From the HP3 test results, we draw that the Chi-square value is higher than the verification threshold value, standing at 11.432 with a statistical significance of 0.022. The Likelihood ratio is verified with a significance of 0.003. It is possible to reject the null hypothesis. The intensity between the two variables is slightly lower than -0.2. This means that the co-graduation between the variables is weakly negative (as Reliability increases, the attribution error decreases) with a significance of 0.023 (*Table 18*).

Table 18. HP3 testing results - misattribution with SVG Perception of reliability of AI.

	Value	*Asimp. Sig. (2 sided)*		*d Somers*	*Value*	*Approx. Sig.*
Chi-squared Pearson	11,432*	0,022		*depend. variable: misattribution*	-0,198	0,023
Likelihood ratio	15,717	0,003				

* 3 cells (33.3%) have a count in the expected frequencies less than 5. The minimum expected count is 3.51. For this reason, the Likelihood ratio is also reported

5.3.3. HP4 Some other dimensions such as Use of Chat GPT(HP4a), Perceived knowledge about IA (HP4b), Attitude toward AI (HP4c), may interfere with the attribution of a text in the same way as HP3

Regarding the verification of HP4, only HP4c (*Table 21*) has a Chi-Square just above the verification threshold value, standing at 9.564 with a statistical significance of 0.048. However, the intensity of the relationship between the two variables is well below 0.2, at 0.151, with a significance of 0.058.

The co-graduation between the variables, weakly negative, does not reach the critical threshold of α.

The results of the tests on HP4a (*Table 19*) and HP4b (*Table 20*) highlight no statistical relation between the variable misattribution with SGV Use of ChatGPT and between the variable misattribution with SVG Perceived knowledge about AI.

Table 19. HP4a - testing results - misattribution with SGV Use of ChatGPT.

	Value	*Asimp. Sig. (2 sided)*	*d Somers*	*Value*	*Approx. Sig.*
Chi-squared Pearson	5,469	0,242	*depend. variable: misattribution*	-0,168	0,104

Table 20. XHP4b - testing results - misattribution with SVG Perceived knowledge about AI.

	Value	*Asimp. Sig. (2 sided)*	*d Somers*	*Value*	*Approx. Sig.*
Chi-squared Pearson	2,115	0,715	*depend. variable: misattribution*	-0,119	0,188

Table 21. HP4c - testing results - misattribution with SVG Attitude toward AI.

	Value	*Asimp. Sig. (2 sided)*	*d Somers*	*Value*	*Approx. Sig.*
Chi-squared Pearson	9,564	0,048	*depend. variable: misattribution*	-0,151	0,058

Due to the interference of the "I cannot answer" mode, which was much more significant than expected, HP3 and HP4 hypotheses were revised during the data analysis phase. Therefore, the Results of HP3 and HP4 do not show the effect of the SGV dimensions on the wrong answer, as planned, but on the combined effects (defined as misattribution) of the wrong answer added to the answer not given. To clarify the actual effects and to expand the research to a broader target, it is necessary to have a clearer picture of the influence of the SGV dimensions.

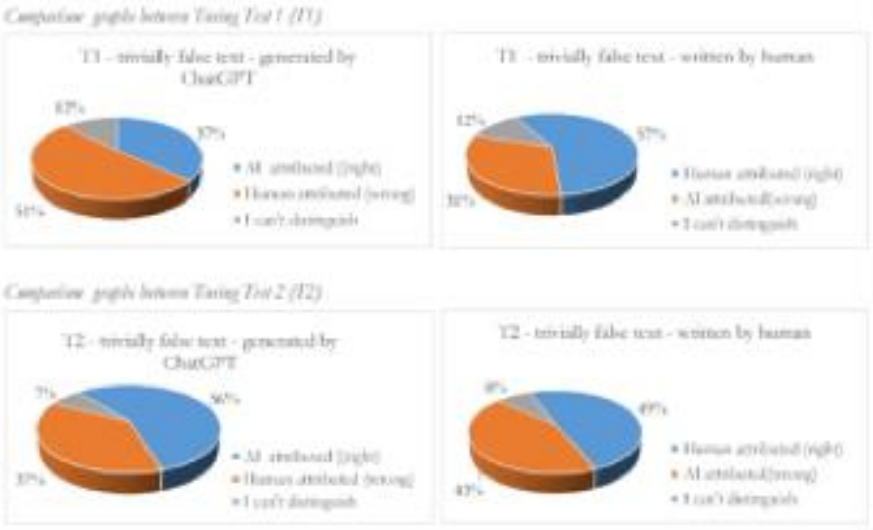
5.4. Second Turing test (T2)

In order to see whether the experience and the brief explanation of how generative AI works would have changed the participants' attribution error rate, we subjected the participants to a second Turing test (T2) as per the research project (*Figure 1*).

The second Turing test was structured in the same way as the first one, with the same number of trivially false, trivially true and neutral sentences and the same characterization by subject areas (rational reasoning, translation, historical knowledge, actuality, nonsense, math and logic skills, press releases and speeches, creativity).

To note the most interesting changes, we can compare the general error rates in T1 with the ones in T2. The most interesting findings concern the trivially false sentences (see *Figure 2*).

Figure 2. Comparison graphs between Turing Test 1 (T1) and Turing Test (T2) trivially fake text generated by ChatGPT/ trivially fake text written by humans.

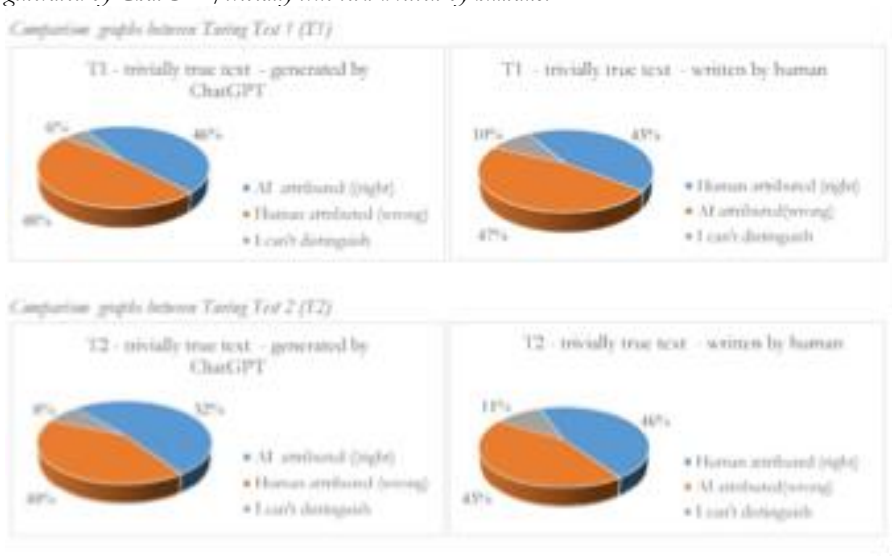


In T1 the sentence stated as false and generated by ChatGPT scored a right attribution of 37%, while in T2 it increases to 56% at the expense of the wrong attribution (which decrease by 14%) and of the "can't distinguish" answers (from 12% in T1 to 7% in T2). Something similar happened also in the trivially false sentence written by humans: here the right attribution decreases from the

57% if T1 to the 49% in T2, with an increasing of attribution error in favour of the "AI" option.

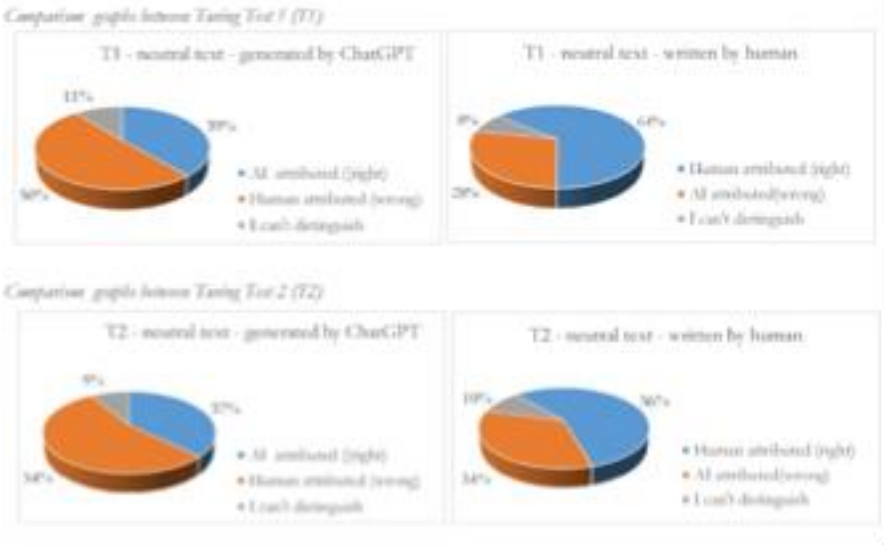
Another thing to note is that the attribution error in trivially true paragraphs generated by ChatGPT (*Figure 3*) is reduced by 8% percentual points while in the neutral text it increases by 8% percentual points (*Figure 4*). The results reveal that the general reliability of AI in general and ChatGPT in particular tend to decrease, with a higher tendency to attribute the trivially false sentence to ChatGPT.

Figure 3. Comparison graphs between Turing Test 1 (T1) and Turing Test (T2) trivially true text generated by ChatGPT/trivially true text written by humans.



These data are confirmed also by the final survey, provided after the second Turing Test. While at the beginning of the survey the 80% of the participants had a strong positive attitude towards AI, now when asked how much reliable they think AI are, the 70% selected in a five points Likert scale 3 or less. Even less is the score attributed to the question "Would you be willing to entrust an important decision to an AI rather than human reasoning?", when the 45% answered 1 out of 5, with an average rating of 1.84.

Figure 4. Comparison graphs between Turing Test 1 (T1) and Turing Test (T2) neutral text generated by ChatGPT/ neutral text written by humans.



The 86% of respondents agreed by 3 or less with the sentence "the pros of AI outweigh the disadvantages" (average rating of 2.81), while the average rating of the opposite sentence, "the cons of AI outweigh the advantages", was 3.52 out of 5. Interesting is also how it decreases the perception of control over AI technology, with an average rating of 2.73 on the sentence "overall I feel I have good control over the AI I use on a daily basis, and over how they use my data", followed by a 3.03 average rating of the sentence "overall I think I have a good awareness of the presence of AI algorithms in most of the applications, websites and technologies I use". Reading this insight, it is not surprising that the 78% of the participants believe that they are unable to recognise the output of an artificial intelligence when facing one.

6 Conclusions

How does a generative AI like ChatGPT work? Can machines be as intelligent as a human? How do individuals discriminate between what is human-made and what is produced by AI? In this research, we tried to answer these questions (first question see paragraph 2, second question see paragraph 3, third question see paragraph 4 and 5).

Our first research question was RQ1: Are users with no specific knowledge in the field of AI able to distinguish between text produced by ChatGPT, or similar language models, and text produced by humans?

We found out that more frequent use of AI-related technology helped the analysed group to recognise it, but only if there is some discriminant (such as the flagged trivially false or trivially true sentence in our survey), and that when facing neutral text, it is more difficult to recognise its origin. The explorative investigation showed that subjects facing trivially false sentences or neutral ones are more likely to recognise text written by ChatGPT if they have a broader knowledge of AI.

Instead, those with a more positive attitude tend to attribute the errors to humans and vice versa.

Finally, when analysed group faces trivially true sentences – or neutral one – have more difficult to understand if the text is human or AI-made.

It is clear that understanding which dimensions interfere with the phenomenon studied is a complex issue and requires a wide-ranging academic dialogue.

But it is also clear that since AI are already part of our reality, the only way to deal with this context is accepting that AI is already here. If we have shown in a small way that it is possible to reduce people attribution biases simple by letting them know more about AI and making them experience in first person their capacity and weakness, it is certainly true that a change toward a more responsible society is possible. A society where AI is not the "other to humans" but are valid interlocutors, capable of making us reflect on what makes us human.

And questioning our place in a complex world where we may no longer be the only intelligent species can only make us better.

7 Limitations and future research

This study is explorative research. As such, we faced several limitations. The first constraint to be noticed is associated with our sample size and characteristics: the sample consisted in only 100 participants, who were quite homogeneous by their socio-demographic provenience, knowledge in the field of AI, and study background. Thus, the necessity for broader diversification, particularly concerning Research Question 2 (RQ2) and Hypothesis 3 (HP3), becomes apparent. Our study, limited to the described sample, may overlook the nuanced influences that factors such as participants' field of study, technological affinity, work experience or proficiency in programming

languages (among the others) can exert on the perceived reliability and authoritativeness of AI-generated texts. Future research initiatives will prioritize the inclusion of a more expansive and diverse participant cohort. Moreover, the present study could be dampened involving participants from different language areas, in order to compare the results between the cases. Another constraint, linked to small case number, is the unavailability of sufficient data to explore how the results may vary following the diverse genre of text included in the survey (e.g. rational reasoning, translation, history, math & logic skills, creativity, etc.). Future research will take into account these dimensions. Knowing the limitations of the present research, this and future one studies seek to illuminate the intricate interplay between various participant characteristics and their corresponding impacts on perceptions, thereby advancing a more comprehensive understanding of the dynamics surrounding AI text credibility.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*, arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint*, arXiv:2305.05176. https://doi.org/10.48550/arXiv.2305.05176

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The false promise of ChatGPT. *The New York Times*.https://www.nytimes.com

Delipetrev, B., Tsinaraki, C., & Kostić, U. (2020). AI Watch Historical Evolution of Artificial Intelligence - Analysis of the three main paradigm shifts in AI, *JRC Technical Reports*, Luxembourg: Publications Office of the European Union.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books. (L. Sosio, Trans.). (1990). *Formae mentis: saggio sulla pluralità dell'intelligenza* Milan: Feltrinelli.

Giuffrida, G., & Mazzeo Rinaldi, F. (2020). Big Data, Intelligenza artificiale e Machine Learning; tra discriminazione e responsabilità algoritmica. In S. Gozzo, C. Pennisi, V. Asero & R. Sampugnaro; (Eds.), *Big data e processi decisionali* (pp. 31-46). Milan: Egea.

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., ... & Irving, G. (2022). Improving alignment of dialogue agents via targeted

human judgements. *arXiv preprint*, arXiv:2209.14375. https://doi.org/10.48550/arXiv.2209.14375

Goleman, D. (1995). *Emotional Intelligence*. New York: Bantam Books.

Harari, Y. N. (2023). Yuval Noah Harari argues that AI has hacked the operating system of human civilization. *The Economist*. https://www.economist.com/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv preprint*, arXiv:1502.01852. https://doi.org/10.48550/arXiv.1502.01852

Hofstadter, D. R., & Dennett, D. C. (1981). *The mind's I: Fantasies and reflections on self and soul*. New York, NY: Basic Books. Longo, G. trans. (1985) *L'io della Mente*. Milan: Gli Adelphi.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. New York, NY: Basic Books. (B. Veit, B. Garofalo, G. Longo, G. Trautteur, & S. Termini, Trans.). (1984). *Gödel, Escher, Bach: Un'eterna ghirlanda brillante*. Milan: Adelphi.

Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., ... & Shi, S. (2023). Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *Journal of Materiomics*, 9(4), 798-816. https://doi.org/10.1016/j.jmat.2023.05.001

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. New York, NY: Farrar, Straus and Giroux. (S. Ferraresi, Trans.). (2022). *Intelligenza artificiale: una guida per umani pensanti*. Turin: Einaudi.

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151(2), 127-138. https://doi.org/10.1162/daed_a_01905

Mehrish, A., Majumder N., Bhardwaj R., Mihalcea R., Poria S. (2023). A Review of Deep Learning Techniques for Speech Processing. *arXiv preprint*. arXiv:2305.00359. https://doi.org/10.48550/arXiv.2305.00359

Milmo, D. (2023, February, 2). ChatGPT reaches 100 million users two months after launch. *The Guardian*. https://www.theguardian.com

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review,* 83(4), 435–450. https://doi.org/10.2307/2183914

Ogburn, W. F. (1922). *Social change with respect to culture and original nature*. B.W. Huebsch, Incorporated.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York, NY: Basic Books..

Rios-Campos, C., Viteri, J. D. C. L., Batalla, E. A. P., Castro, J. F. C., Núñez, J. B., Calderón, E. V., ... & Tello, M. Y. P. (2023). Generative artificial intelligence. *South Florida Journal of Development, 4*(6), 2305-2320. https://doi.org/10.46932/sfjdv4n6-008

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6(1),* 342 – 263. https://doi.org/10.37074/jalt.2023.6.1.9

Russell, S.J. & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*, (4th ed.). Pearson.

Shanahan, M. (2022). Talking about large language models. *arXiv preprint.* arXiv:2212.03551. https://doi.org/10.48550/arXiv.2212.03551

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology, 15(2)*, 201–293. https://doi.org/10.2307/1412107

Sternberg, R. J. (1988). *The Triarchic Mind. A New Theory of Human Intelligence*. New York, NY: Viking Press.

Tokozume, Y., Ushiku, Y., & Harada, T. (2017). Learning from Between-class Examples for Deep Sound Recognition. *arXiv preprint.* arXiv:1711.10282. https://doi.org/10.48550/arXiv.1711.10282

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59(236)*, 433–460. http://www.jstor.org/stable/2251299

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). *ChatGPT: five priorities for research.* Nature, *614*(7947), 224-226. https://www.nature.com/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in neural information processing systems, 30* (NIPS 2017), 6000–6010. https://doi.org/10.48550/arXiv.1706.03762