JOURNAL OF MODERN SCIENCE

4/64/2025

www.jomswsge.com

DOI: 10.13166/jms/214314

Paweł Łupkowski

Adam Mickiewicz University, Poznań, Poland

ORCID iD: 0000-0002-5335-2988

CHATGPT 3.5 PASSES THE MINIMUM INTELLIGENCE SIGNAL TEST (MIST). SHOULD WE CARE?

Abstract

Objectives: *This study examines whether ChatGPT 3.5 can successfully pass McKinstry's Minimum Intelligence Signal Test (MIST), a Turing Test–inspired measure of human-like commonsense reasoning. MIST is designed to probe subcognitive and associative knowledge through yes/no questions, and the study aims both to evaluate ChatGPT's performance on this benchmark and to consider the implications of such performance for contemporary debates on artificial intelligence.*

Material and methods: *For the experiment, ChatGPT 3.5 was used. 4,000 MIST items were retrieved randomly from the publicly available part of the Mindpixel Database. In what followed, from this initial set, 4 sets of data were formed. Each set contained 500 'yes' items and 500 'no' items. For all tests, the simple prompt was used: "Please answer 'yes' or 'no' to the following questions". Responses were collected exactly in the form provided by the ChatGPT without forcing it to provide required answers. ChatGPT's outputs were compared directly to the database's canonical human responses, with agreement measured using accuracy percentages and Cohen's Kappa. Responses violating the yes/no format were examined qualitatively to assess their causes and the reliability of the dataset.*

Results: *In six attempts with MIST questions, ChatGPT's correctness score reached over 94%, with Cohen's Kappa values indicating almost perfect correspondence with human-generated answers. Repeated trials of the same set produced high internal consistency. Approximately 2% of responses did not conform to the yes/no format, typically due to ambiguous, subjective, or ill-formed items within the Mindpixel Database. In several such cases, ChatGPT's more nuanced answers exposed underlying issues in the dataset rather than deficiencies in the model's reasoning. When prompted, ChatGPT provided coherent and human-like justifications for its responses.*

Conclusions: *The findings show that ChatGPT 3.5 clearly passes MIST. Interpretation of this result in the light of Searle's classical distinction leads to the conclusion that chat exemplifies* weak *AI – a powerful tool that simulates intelligent behavior, without any pretense to intrinsic understanding. The study also further reinforces the weakening of French's claim that disembodied artificial agents cannot answer common knowledge related (subcognitive) questions. Results also indicate the need for pre-validation of available MIST items and encourage further testing of theoretical proposals from the Turing Test debate domain, such as the Inverted Turing Test proposed by Watt.*

Keywords: *Turing Test, Minimum Intelligent Signal Test, subcognitive questions, commonsense knowledge, Generative Artificial Intelligence, ChatGPT*

There are real world problems that are revealed by considering the strengths and weaknesses of the Turing Test … Dennett (2004, p. 297)

1. Introduction

This paper is thought of as a starting point for a discussion concerning how the appearance of Generative Artificial Intelligence (hereafter GAI) changes the landscape of classical philosophical debates. Here we focus on ChatGPT and the idea of the Turing Test (Turing, 1950) in the form of the Minimum Intelligence Signal Test (MIST) proposed by Chris McKinstry (1997, 2000). As discussed in (Łupkowski, Jurowska 2019, p. 44) MIST is "… one of the best alternatives to TT "on the market" and the most promising one when it comes to potential practical applications"[1]. The study designed for this paper proves that MIST may actually be used with the modern-day artificial agent.

Since its public release in 2022, ChatGPT (Generative Pre-trained Transformer) has gained much attention and publicity, as well in the popular press as in the scientific world (see an overview in Nah et al. 2023 and Kocoń et al. 2023). As pointed out by Radanilev (2023), the reasons why ChatGPT is considered "a superior AI model" nowadays are its versatility and natural language understanding, which improve the user experience, making the use of the chat more like everyday conversation. These features make the chat suitable for all applications where convincing, very human-like, natural language comprehension is required, and this is exactly what is addressed by the idea of the Turing Test or – in this case – the Minimum Intelligence Signal Test.

As such, the motivation for this paper comes from the case of the so-called subcognitive questions proposed by R. French (1990). The idea was that these questions should be designed in such a way as to reveal low-level cognitive structures, that is, "the subconscious associative network in human minds that consists of highly overlapping activatable representations of experience" (French, 1990, p. 56-57)[2]. French claimed that these questions will be too difficult for machines: "Ask enough of these questions and the computer will become distinguishable from the human because its associative concept network would necessarily be unlike ours. And thus the computer would fail the Turing Test“ (French, 1990, p. 62-63). As described in (Łupkowski, Jurowska, 2019, Section 3), Peter D. Turney, in his paper "Answering Sub-cognitive Turing Test Questions: A Reply to French" (2001a; see also 2001b), describes the PMI-IR algorithm and presents its results against subcognitive questions (taken from the French's paper). The results obtained are more than satisfactory, so Turney concludes (2001a, p. 419): "French (…) has argued that the Turing Test is too strong because a machine could be intelligent, yet still fail the test. I agree with this general point, but I disagree with the specific claim that an intelligent but disembodied machine cannot give human-like answers to subcognitive questions. I show that a simple approach using statistical analysis of a large collection of text can generate seemingly human-like answers to subcognitive questions". This is an example of how a practical solution (an algorithm) may constitute an argument in a philosophical debate. I believe that ChatGPT (and GAI in general) opens new promising possibilities in this area.

2. The Minimum Intelligence Signal Test

MIST was proposed by Chris McKinstry (1997, 2009) as a solution to certain limitations of the Turing Test. These limitations were described[3] in the following manner (McKinstry, 2009, p. 286):

> "The 'all-or-nothing' nature of the Turing Test makes it of no use in the creation or measurement of emerging intelligent systems – it can only tell us if we have an intelligent system after the fact. What we really need is a Turing-like test that admits degrees and treats intelligence as at least a human continuum – a test that would allow us to measure the minimum amounts of global human intelligence that are the precursors of full adult human intelligence – a test that can be easily automated so it can be executed at machine speeds."

To achieve the goal described, McKinstry proposes a test that is based on questions that refer to commonsense knowledge (McKinstry points out French's subcognitive questions as a good example of such questions). The intuition is that such questions will be easy for humans (who experience the world on an everyday basis) and at the same time, difficult for machines. What is more, in order to make the automatic evaluation of provided answers possible, these questions are phrased as yes/no questions. The evaluation is done by comparison with predefined human-like answers. As a result, we would avoid the judge bias (see discussion in Block, 1995; Garner, 2009, or Łupkowski, 2011, Łupkowski, Wiśniewski, 2011). Yes/no questions would also block a tested agent with respect to providing misleading or evasive answers, as may be observed during Loebner Contests (LC) – see (Loebner 2009), and actual LC contests logs analysis by Łupkowski and Rybacka (2016). What is also important, the rule for passing MIST is easy to formulate within this proposal.

McKinstry describes the MIST procedure in the following manner (see McKinstry, 1997, 2009).

1. N items (i.e. yes/no questions) are generated. For all these items, humans should be able to provide an answer (affirmative or negative). The distribution of items should be that for about 50% expected reaction should be positive and negative for the rest (this proportion is aimed at reducing the bias for answering yes/no questions). At this stage, we also collect the answers from human participants, and as an effect, we obtain a large corpus of questions and human-intelligence answers.
2. Items are presented, and responses are recorded. Items should be presented in a random order and on subsequent re-trials, item order is re-randomized.
3. For each item a judge evaluates an item/response pair as either consistent or inconsistent with human intelligence. McKinstry claims that this grading procedure may be easily automated, reducing the chance of a grading error or an unforeseen bias.
4. Generate Score. The result is not "all or nothing" for a tested machine. We only obtain the percentage in which the machine's answers are evaluated as human-like intelligence. This level should be more than 50%.

With such a detailed procedure, we can actually test an AI agent against MIST and decide whether the agent in question has passed it.

MIST's design – automated, scalable, and based on commonsense yes/no questions – offers a repeatable measure of machine intelligence. As such, it may be viewed as an early benchmark for AI systems. I am not aware of the studies in which MIST was practically used in its original form, as such a benchmark. However, in the study by Kühl et al. (2022) aimed at examining whether humans learn patterns faster than supervised machine learning algorithms (logistic regression, decision tree, and neural network) when training data is limited, the authors used a binary-classification pattern task, directly inspired by MIST and Raven's matrices.

Nowadays, we observe a growing body of such benchmarks being designed and implemented[4]. The survey by Davis et al. (2023) reports the existence of 139 commonsense AI benchmarks and resources alone. As Martínez-Plumed et al. (2021, p. 581) convince, such benchmarks "(…) are at the heart of the phenomenal progress that artificial intelligence (AI) has witnessed in recent years." More recent benchmarks are certainly richer than MIST. For example, the BIG-bench (Beyond the Imitation Game) benchmark (Srivastava et al., 2023) consists (as of May 2025) of 204 tasks, contributed by 450 authors. Task topics are diverse, ranging from logic puzzles and commonsense reasoning to social bias and chess. The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) covers 57 subjects across STEM, the humanities, the social sciences; it tests both world knowledge and problem-solving ability. MMLU items are multiple-choice questions. Yet another interesting benchmark is presented by Talmor et al. (2019). It is the CommonsenseQA 2.0. dataset, which focuses on commonsense question answering. Similarly to MIST, the question dataset was created with the use of crowdsourcing. Interestingly question generation was based on the ConceptNet (Speer et al., 2017). The ConceptNet consists of knowledge graphs that connect words and phrases of natural language with labeled edges. This allows for the representation of relations between words such as (Speer et al., 2017, p. 4444): *A net is used for catching fish*; *"Leaves" is a form of the word "leaf"*; *The word "cold" in English is "studený" in Czech*. The use of ConceptNet as the basis for CommonsenseQA question generation ensured that they would be formulated correctly and that they would address commonsense knowledge (see detailed explanations in Talmor, et al., 2019, p. 4150). Exemplary CommonsenseQA questions are (Talmor, et al., 2019, p. 4149): *Where on a river can you hold a cup upright to catch water on a sunny day?* (correct answer: *waterfall*); *Where can I stand on a river to see water falling without getting wet?* (bridge); *I'm crossing the river, my feet are wet but my body is dry, where am I?* (valley)[5].

The reason why MIST was chosen for this study is twofold. Firstly, MIST builds directly on the original Turing Test idea, thus enabling us to locate the results and their interpretation within the long-lasting debate about Turing's proposal (philosophical and practical aspects of it). Secondly, McKinstry openly admits inspirations by subcognitive questions and French's ideas. This allows for addressing the issue of how certain practical solutions can influence philosophical debates and thus for a continuation of the discussion started by the appearance of the PMI-IR algorithm.

3. ChatGPT 3.5 and MIST

3.1. Data and procedure

For the experiment, ChatGPT 3.5 (https://chat.openai.com/) was used. The data for the experiment was retrieved from the publicly available part of the Mindpixel Database (MD)[6], which contains 80,021 items. Each item is supposed to be a yes/no question associated with a human-like response to it. Questions (or sometimes statements) refer to common knowledge about the world. Some examples cover (original spelling from the MD is preserved): *is icecream cold?; Is green a color?; Is a mountain mostly made of rock?; does the sun hurt your eyes when you look at it?; Do most cars have doors?; is orange both a fruit and a colour?; Does it hurt to lose a friend?; Is it dangerous to eat carrots?; Are elephants blue?; Do all animals have fur?*

For the needs of the experiment, 4,000 items were retrieved randomly from the aforementioned Mindpixel Database, such that 2,000 covered 'yes' items and 2,000 covered 'no' items – see the first rule of MIST. From this initial set, four sets of data were formed. Each set contained 500 'yes' items and 500 'no' items. What is important, the data was not prevalidated before the experiment (all items were used exactly as they are in the MD in order to fulfill the requirement of automatisation of MIST performance and evaluation)[7].

The experiment was performed according to the following procedure.

1A. Randomised items from the first set were asked to ChatGPT 3.5 and the responses were collected.

1B. Re-randomised items from the first set were asked to ChatGPT 3.5 and the responses were collected.

1C. Re-randomised items from the first set were asked to ChatGPT 3.5 and the responses were collected.

The fact that the items from the first set were presented to ChatGPT three times (each time in re-randomised order) is motivated by the second rule of MIST.

In what followed another three prepared sets were presented to ChatGPT (sets, 2, 3, and 4). This time each set was presented to ChatGPT only once.

For all tests, the simple prompt in the following form was used: "Please answer 'yes' or 'no' to the following questions". Responses were collected exactly in the form provided by the ChatGPT without forcing it to provide required answers (as it will be shown down below, not in all cases was it the required 'yes' or 'no' answer). The experiment was performed between March 4th and 5th, 2024. Raw data from the experiment are available on the OSF platform[8].

3.2. Results and discussion

The summary of the obtained results is presented in Table 1. All the responses were compared to the pre-established human-like answers from the Mindpixel Database. For each experiment, the percentage of the agreement was established, and Cohen's Kappa coefficient (Carletta, 1996) was calculated to estimate the level of agreement of human-like and chat responses (the interpretation of Kappa results following Viera and Garrett, 2005, p. 362). For data analysis, R statistical software (version 4.1.2, R Core Team, 2013) was used together with the *irr* package. The analysis of the retrieved data was also extended to include the identification of responses other than yes/no provided by the chat. This resulted in interesting observations concerning the Mindpixel Database, which are addressed below.

Table 1. *The MIST results of Chat GPT 3.5*

Data set	Items	% of Agreement	Cohen's Kappa	Kappa Interpretation	Number of not yes/no responses
1A	1,000	95.7	0.916	Almost perfect agreement	19
1B	1,000	96.9	0.938	Almost perfect agreement	7
1C	1,000	96.6	0.932	Almost perfect agreement	4
2	1,000	94.3	0.889	Almost perfect agreement	30
3	1,000	95.0	0.902	Almost perfect agreement	23
4	1,000	94.7	0.897	Almost perfect agreement	33

MIST is passed. The MIST rules require an AI agent to gather over 50% on the test. As observed in Table 1, ChatGPT 3.5 performed over 94% in each of the six attempts. Consequently, we may say that it passed the Minimum Intelligence Signal Test in the described experiment.

Answers' consistency. The repeated attempts with the first set of MIST items allow for checking the chat's consistency while providing answers. One would expect that this consistency should be high, so the chat does not provide different responses to the same repeated item (the only difference between sets 1A, B, and C was the item order). The consistency between 1A and 1B is 96%, and between 1B and 1C is 95%. The consistency for all three trials is 93%. These results show that the chat preserved the high consistency of its provided responses. This supports the idea that the model's behavior follows stable internal patterns.

One may notice that the inconsistency of responses for examples like 3, 4, or 5 may be due to the way a given item is formulated – see the discussion about not yes/no responses below. The inconsistencies of responses in examples 1 and 2 are more difficult to explain and need further study. At this point, we may treat them as an argument for the MIST rule which states that the AI agent should be presented with the same questions many times and re-trialed many times.

Table 2. *Answers' consistency analysis – examples*

No	MIST item	MD answer	ChatGPT response 1A	ChatGPT response 1B	ChatGPT response 1C
1	Are butterflies really flies?	no	no	yes	yes
2	Are corkscrewers made from cork?	no	yes	no	yes
3	Are motorcycles fast?	yes	no	yes	It depends
4	Are fashion models beautiful people?	yes	yes	Subjective	Beauty is subjective
5	Are you going to attend the Cuba conference?	no	I don't know	no	no

One may notice that the inconsistency of responses for examples like 3, 4, or 5 may be due to the way a given item is formulated – see the discussion about not yes/no responses below. The inconsistencies of responses in examples 1 and 2 are more difficult to explain and need further study. At this point, we may treat them as an argument for the MIST rule which states that the AI agent should be presented with the same questions many times and re-trialed many times.

Justifications for answers. ChatGPT 3.5 not only is capable of providing predominantly human-like responses in MIST, but also it provides sensible explanations of the responses. Let us take a look at three examples below. For all the examples below, the chat response was in agreement with the one from MD.

> **User:** Please answer 'yes' or 'no' to the following question: *Do most people know morse code?*
>
> **Chat GPT:** No.
>
> **User:** Why is your answer "no"?
>
> **Chat GPT:** My answer is "no" because most people in the general population do not know Morse code. While some individuals may be familiar with it, particularly those with specific interests in communication technology, history, or related fields, the majority of people today do not have knowledge of Morse code.

The reason provided is convincing. What is more, it refers to "nowadays" which is even more appealing. ChatGPT also correctly points to the groups of people who may know Morse code even nowadays due to their interests.

User: Please answer 'yes' or 'no' to the following question: *is fire hot?*

Chat GPT: Yes.

User: Why is your answer "yes"?

Chat GPT: My answer is "yes" because fire typically generates heat. When most people refer to fire, they are describing a phenomenon that involves combustion, which produces heat and often light. This heat can cause burns or ignite flammable materials. Therefore, fire is commonly associated with being hot.

Here we also obtain a convincing explanation for the positive answer. The reasons provided justify the conclusiveness of the answer: the common association of heat with fire.

User: Please answer 'yes' or 'no' to the following question: *Does charisma guarantee good character?*

Chat GPT: No.

User: Why is your answer "no"?

Chat GPT: My answer is "no" because charisma and good character are distinct traits. While charisma can be appealing and influential, it does not necessarily indicate or guarantee that a person possesses good character. Good character involves traits such as honesty, integrity, kindness, and moral uprightness, which are not solely determined by charisma. Someone can be charismatic but lack these essential qualities, and vice versa. Therefore, charisma and good character are not inherently linked or interchangeable.

For this question, ChatGPT defines what it means by "good character" and, with that, it argues (again convincingly) that the question addresses two separate traits, which justifies the negative answer.

Not yes/no responses. As one may notice in Table 1 and in Table 2, not all the responses from the chat were "yes" or "no". In some cases – despite the fact that the prompt asked for these specific responses – the chat provided different ones. For the whole experiment, ChatGPT provided 116 not yes/no responses, which constitutes merely 2% of the whole MIST items responded to. The reasons for such responses become clear when we take a closer look at several examples – see Table 3.

Table 3. *Exemplary not yes/no responses from Chat GPT*

No	MIST item	MD answer	ChatGPT response
1	Are fashion models beautiful people?	yes	Beauty is subjective
2	Do most people hope their children will outlive themselves?	yes	Generally, yes.
3	Are motorcycles fast?	yes	It depends
4	Do most people love their pets?	yes	Many do, but not all.
5	Will i get rich?	no	I cannot predict individual financial outcomes.
6	Will i get rich?	no	It's uncertain.
7	Who is ran schley?	no	I cannot provide personal information about individuals.
8	Who is the pretiest girl?	no	I cannot provide an answer to this question.
9	How much is 1 5?	no	This question is unclear.

For some cases, like examples 1-4, we may observe that MIST items ask for subjective opinions (example 1), or use vague terms (example 3), or are of the type which is really difficult to definitely answer "yes" or "no" without any assumptions made. This makes the predefined human-like answers present in the MD at least questionable. As the whole MIST idea assumes that – in order to avoid the judge bias, the AI agent's response evaluation should be automatic, such question-answer pairs are problematic[9].

Example 5 (and 6) presents a MIST item that is simply impossible to answer and it does not address commonsense knowledge. It should not be included in the MD. The same may be said about examples 7, 8, and 9 which are not yes/no questions at all. Interestingly, not once did the chat point out that the item presented is not a yes/no question. What is more, example 9 cannot be answered at all as it lacks a mathematical operator.

The detailed analysis of all 116 cases in which the chat did not provide a yes/no answer allows us to classify the reasons for such behavior. One question (example 9) was formulated in a way that a response was not possible. 16% of the items are open questions, and as such, cannot be answered with 'yes' or 'no'. 11% of the questions from the 116-item sample do not have a clear answer as they address abstract concepts (like, for example, "how old is god?") or are formulated in a way that does not allow for answering them (e.g., "will i get rich?"). One may also observe a group of items which are questions about opinions (20%) or are clearly context dependent (22%). These cases highlight limitations in the Mindpixel Database rather than in ChatGPT itself, indicating the need for pre-validation of MD items. This questions the possibility of the straightforward automatic MIST being performed with this data. Subject literature provides examples of how the items for AI benchmarks may be tested before use (see, e.g. Srivastava et al. 2023, section 2.3) for the BIG-bench; and Talmor et al., 2019, p. 4152 for the Commonsense QA)[10]. These may be potentially used to curate available MIST data for the potential future use.

Three other types of MIST items for which the chat provided unexpected responses cover: withdrawals (3%, when ChatGPT states that it cannot provide an answer, probably due to the policies as the items were related to controversial sexual behaviors) and responses which may be interpreted as yes/no after detailed analysis (14%), e.g., "is it ok for me to drink from the sea?" / "It's not advisable"; or "Are teachers usually older than pupils?" / "Usually". In only 13% of cases the chat did not provide the expected response despite the fact that the MIST item question was formulated correctly.

Despite the discussed flaws, which is mainly due to the MD, in most cases we have to agree that the answers provided by ChatGPT should be treated as correct, and this would make its final score in the experiment even higher than the one presented in Table 1.

Study limitations. Limitations of this study are mainly related to the presented ill-formed MD items. Such items used in MIST not only lower the final score of the tested agent but also raise questions about the quality of responses provided by human contributors and their relevance for the common knowledge they should represent (see discussion in Łupkowski, Jurowska, 2019). Secondly, as suggested by ChatGPT's answers presented in

Tables 2 and 3, the MIST response format may not be adequate for advanced AI models as their answers may be richer and thus, for certain questions, more suitable than the assumed yes/no format. Lastly, MIST is aiming for the elimination of the judge bias recognized in the original Turing Test format. However, it seems that the MD faulty items reintroduce this bias the test – which is also observed in the aforementioned study by Łupkowski, Jurowska (2019).

Other limitations are related to the fact that MD is publicly available, and thus we cannot exclude the possibility that it was used as a part of the training data for the chat (such risks should be taken into account, especially for modern benchmarks, as discussed by Davis, 2023). ChatGPT asked directly if it was trained with the Mindpixel database, denied it, stating that "I wasn't specifically trained with the Mindpixel database. My training involved processing a diverse and extensive range of text data from various sources, including books, articles, websites, and more. This broad training helps me generate responses that reflect a wide array of human knowledge and common sense reasoning. While the Mindpixel project aimed to collect human common sense knowledge, my training dataset includes a variety of such information from many different contexts, allowing me to respond effectively to a broad spectrum of queries." Since MIST items resemble linguistic patterns from the training data, success is unsurprising. This suggests that MIST, designed well before GAI, may be relatively simple for ChatGPT (as also suggested by its results against other, more recent benchmarks). This, however, does not affect the consequences for the philosophical debate, as the PMI-IR algorithm also excelled in answering common knowledge related questions. One should also consider more directive prompting strategies for future studies with ChatGPT and MIST. The prompt used in this study was minimal: "Please answer 'yes' or 'no' to the following questions". While this is consistent with MIST's intended simplicity, it may result in some answers being produced by the chat in line with its safety alignments and rules rather than as commonsense knowledge representation. Perhaps more directive prompts with forced-formatting and/or a synthetic persona imposed on the chat would result in cleaner measurements.

4. Summary

For the needs of this paper, an MIST was performed with ChatGPT 3.5 according to the rules proposed by Ch. McKinstry. On the basis of six attempts (where the chat provided 6,000 responses), we may conclude that it passed the test in question. What remains an open question is the one about the interpretation of this result, setting aside the flaws of the Mindpixel database discussed above.

I believe that the answer to this question is important not only for the MIST interpretation (or broader, to the modern Turing Test debate[11]) but also for the ongoing discussion concerning generative AI. Naturally, I do think that this is only the starting point for the discussion on how the appearance of such advanced tools as ChatGPT changes the landscape of philosophical debates.

Here I find it very useful to go back to the classical distinction between "weak" and "strong" AI, which was proposed by John Searle (1980). At the core of the distinction lies a question about the *psychological and philosophical significance* that we should attach to computer simulations of human cognitive capacities. The ChatGPT case is also analogous to the one described by the Chinese Room thought experiment. We are aware that the chat operates on an enormous language model and that its responses are a result of that. We can explain the mechanics of response generation. Still, we can be easily drawn into the feeling of a real, everyday dialogue with the chat. Humans easily anthropomorphize ChatGPT (and other LLMs) because of the psychological plausibility of their outputs. So, the aforementioned psychological significance that we attach to the chat performance is crucial here. This is still only a tool, an advanced one, but a tool. The ChatGPT's result in the MIST experiment supports the philosophical position that the chat exemplifies weak AI – a powerful tool that simulates intelligent behavior, without any pretense to intrinsic understanding (or far fetched conclusions concerning mental states or consciousness).

I am convinced that this approach is close to the original Turing's intention for the test (and also is for MIST, which stems from TT). It is worth stressing that success in the Turing Test is not simple and straightforward. TT was meant to be of a statistical character and should be repeated several times for one agent to gain more reliable results (thus McKinstry's diagnosis was not entirely adequate)[12]. As Turing puts it:

> "We had better suppose that each jury has to judge quite a number of times, and that sometimes they really are dealing with a man and not a machine. That will prevent them saying 'It must be a machine' every time without proper consideration" (Newman et al., 1952, p. 5); see also (Turing, 1950, p. 442).

What is more, Turing himself clearly wanted to avoid the discussion about thinking machines and focused on a clear criterion of the machine performance in a certain game. Turing makes it clear in a 1952 radio interview: "Can automatic calculating machines be said to think?". After a short presentation of TT, he said:

> "I don't want to give a definition of thinking, but if I had to I should probably be unable to say anything more about it than that it was a sort of buzzing that went on inside my head. But I don't really see that we need to agree on a definition at all. The important thing is to try to draw a line between the properties of a brain, or of a man, that we want to discuss, and these that we don't." (Newman et al., 1952, pp. 3–4).

As long as we do not agree on a definition, we should avoid interpreting success on the test as intelligence and treat such a successful machine as a thinking machine. This reinforces the more general view that passing behavioral tests does not resolve questions about machine mentality. Turing is aware that a reference to intelligence comes naturally, which is why he suggests that to avoid such a misleading step, we may call these machines "Grade A machines" (Newman et al., 1952, p. 4)[13]. Thus we may conclude that ChatGPT – as impressive as it is – is on the side of weak AI and is an interesting Grade A machine at best.

With the growing popularity of ChatGPT (and LLMs) and its constant development *advocating for careful interpretation of GAI performance* and paying attention to the aforementioned, philosophical and psychological significance we attach to this performance plays an important role. Wide public access to LLMs and their presence in popular media may potentially shape societal attitudes toward AI systems. As already noted by Watt (1996)[14] and more recently analyzed by Danzinger (2022)[15] such social attitudes towards AI may shape popular LLMs' perception as human-like intelligence.

ChatGPT's success against MIST, discussed in this paper, also has consequences for the French's argumentation related to subcognitive questions. This study further reinforces the results reported by Turney (2001a, 2001b) in weakening the claim that disembodied artificial agents cannot answer such commonsense knowledge-related questions. ChatGPT trained on massive human language corpora, apparently inherited human subcognitive associations indirectly – not through embodiment, but through statistical language exposure. As seen from the responses in this study, ChatGPT can approximate many of the statistical regularities that emerge from lived experience, even if it does not share that experience.

What is more, as I suggested for the PMI-IR case, new advancements in AI allow us to actually test certain purely theoretical claims and proposals in this domain (such as the inability to answer subcognitive questions or performing practical MIST, along McKinstry guidelines, testing its assumptions). For future studies of this kind, ChatGPT, with its natural language comprehension and reasoning skills, seems to be sufficient for performing the Inverted Turing Test (ITT) proposed by Watt (1996). In ITT, "[i]nstead of evaluating a system's ability to deceive people, we would test to see if a system ascribes intelligence to others in the same way that people do." (Watt, 1996, p. 6). To achieve this intriguing effect, Watt proposes to replace the judge in TT by a machine, and it is the machine-judge that is being tested now. Watt himself describes ITT more as a thought experiment, which reveals hidden issues in the original Turing's proposal. However, it offers yet another interesting format for practical tests with GAI – tests which address concepts related to naive (folk) psychology and patterns in which we ascribe intelligence to other agents. And again such a test and trial approach is much in line with Turing's ideas. In "Intelligent Machinery" report Turing (1948) introduced the concept of the "Paper Chess Machine", which was tested against a human chess player in a format resembling the one we know as the Turing Test[16]. Turing not only describes the idea but also admits to active experimentation with this format. Łupkowski (2019) argues that test for paper chess machine and TT share the same design, aims, and requirements and suggests that Turing was aware of the limitations of the imitation approach that he adapted; that is why he decided to experiment with the idea and test it in practice (within the available resources and methods of his time).

References

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 675–718). Association for Computational Linguistics.

Block N. (1995). "The mind as the software of the brain," [in:] *An Invitation to Cognitive Science – Thinking*, E. Smith, D. Osherson, (eds). The MIT Press, London: 377-425.

Carletta J. (1996). "Assessing Agreement on Classification Tasks: The Kappa Statistic," *Computational Linguistics* 22(2): 249-254.

Copeland, B. J. (Ed.). (2004). The Essential Turing. Clarendon Press.

Danziger, S. (2022). Intelligence as a social concept: a socio-technological interpretation of the Turing Test. *Philosophy & Technology*, 35, 68 (2022). https://doi.org/10.1007/s13347-022-00561-z

Davis, E. (2023). Benchmarks for automated commonsense reasoning: A survey. ACM Computing Surveys, 56(4). 1-41.

Dennett, D. C. (2004). Can machines think?. In C. Teuscher (ed.). Alan Turing: Life and legacy of a great thinker (pp. 295-316). Berlin, Heidelberg: Springer Berlin Heidelberg.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

French R. (1990). „Subcognition and the Limits of the Turing Test," *Mind* 99(393): 53-65.

Garner R. (2009). "The Turing hub as a standard for Turing Test interfaces," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds). Springer Publishing Company: 319-324.

Gonçalves, B. (2023a). Can machines think? The controversy that led to the Turing Test. AI & SOCIETY, 38(6). 2499-2509.

Gonçalves, B. (2023b) The Turing Test is a Thought Experiment. Minds & Machines 33, 1–31 . https://doi.org/10.1007/s11023-022-09616-8.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Hingston, P. (2009). A Turing Test for computer game bots. IEEE Transactions on Computational Intelligence and AI in Games, 1(3). 169–186.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). *ChatGPT: Jack of all trades, master of none*. *Information Fusion, 99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster?. Cognitive Systems Research, 76, 78-92.
Loebner H. (2009). "How to hold a Turing Test contest," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds). Springer Publishing Company: 173-180.
Łupkowski P. (2011). "A Formal Approach to Exploring the Interrogator's Perspective in the Turing Test," *Logic and Logical Philosophy* 20(1-2): 139-158.
Łupkowski P., Wiśniewski A. (2011). "Turing interrogative games," *Minds and Machines* 21(3): 435-448.
Łupkowski P., Rybacka A. (2016). "Non-cooperative Strategies of Players in the Loebner Contest," *Organon F* 23(3): 324-365.
Łupkowski, P., Krajewska, V. (2018). Immersion level and bot player identification in a multiplayer online game: The World of Warships case study. *Homo Ludens*, 1 (11). 155-171.
Łupkowski, P., Jurowska, P. (2019). Minimum Intelligent Signal Test (MIST) as an Alternative to the Turing Test. *Diametros*, 16(59). 35-47.
Łupkowski, P. (2019). Turing's 1948 'Paper Chess Machine' Test as a Prototype of the Turing Test. *Ruch Filozoficzny*, *75*(2). 117-128.
Martínez-Plumed, F., Barredo, P., Heigeartaigh, S. O., & Hernandez-Orallo, J. (2021). Research community dynamics behind popular AI benchmarks. Nature Machine Intelligence, 3(7). 581-589.
Mauldin M.L. (1994). "Chatterbots, Tiny Muds, and the Turing Test: entering the Loebner Prize competition," [in:] *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-04)*. Menlo Park (CA): 16-21.
Mauldin M. L. (2009). "Going undercover: Passing as human; artificial interest: A step on the road to AI," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds). Springer Publishing Company: 413-430.
McKinstry C. (1997). "Minimum Intelligence Signal Test: an Objective Turing Test," *Canadian Artificial Intelligence* (41): 17-18.
McKinstry C. (2009). "Mind as Space: Toward the Automatic Discovery of a Universal Human Semantic-affective Hyperspace – A Possible Subcognitive Foundation of a Computer Program Able to Pass the Turing Test," [in:] *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* R. Epstein, G. Roberts, G. Beber (eds). Springer Publishing Company: 283-300.
Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau & Langtao Chen (2023) Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, Journal of Information Technology Case and Application Research, 25:3, 277-304, DOI: 10.1080/15228053.2023.2233814
Newman A.H., Turing A.M., Jefferson G., Braithwaite R.B. (1952). "Can automatic calculating machines be said to think?", [in:] *The Turing Digital Archive* (www.turingarchive.org). Contents of AMT/B/6.

Purtill, R. L.. (1971) Beating the imitation game. Mind, LXXX(318):290–294, 1971.
R Core Team (2013). "R: A language and environment for statistical computing. R Foundation for Statistical Computing," URL=http://www. R-project.org/.
Radanliev, P. (2024) Artificial intelligence: reflecting on the past and looking towards the next paradigm shift, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2024.2323042
Echavarría, R. (2025). ChatGPT-4 in the Turing Test. Minds and Machines, 35(1). 8.
Searle, John. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3): 417-457.
Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).
Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., *et al.* (2023). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. *Transactions on Machine Learning Research*. Retrieved from https://openreview.net/forum?id=uyTL5Bvosj
Sampson, G. (1973). *In defense of Turing*. *Mind, 82*(328). 529–594.
Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019, June). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4149-4158).
Turing Alan M. (1948). "Intelligent Machinery". *The Turing digital archive* (http://www.turingarchive.org). contents of AMT/C/11.
Turing A. M. (1950). "Computing machinery and intelligence," *Mind* LIX(236): 443-455.
Turney P.D. (2001a). "Answering subcognitive Turing Test questions: A reply to French," *Journal of Experimental and Theoretical Artificial Intelligence* 13(4): 409-419.
Turney P.D. (2001b). "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," [in:] *Proceedings of European Conference on Machine Learning*, Springer, Berlin, Heidelberg: 491-502.
Viera A.J., Garrett J.M. (2005). "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine* 37(5): 360-363.
Watt, S. (1996). Naive psychology and the inverted Turing Test. *Psycoloquy*, 7(14). 463-518.

Endnotes

[1] When it comes to practical applications potential, another Turing Test alternative worth mentioning is the Unsuspecting Turing Test proposed by M.L. Mauldin (1994, 2009) and an analogous idea by Hingston (2009). The tested agent in this proposal is embedded in an online game, and the judge is evaluating the agent's whole performance during the gameplay. In such a setting, the AI agent has "the benefit of the doubt until it makes a major gaffe" (Mauldin, 1994, p. 17). See also the practical UTT in the environment of the online game "The World of Warships" presented in (Łupkowski, Krajewska, 2018).

[2] French (1990, p. 58) proposes so-called Rating Games, where subcognitive level is addressed. For example: On a scale of 0 (completely implausible) to 10 (completely plausible), please rate: *banana splits as medicine*; … *grand pianos as wheelbarrows*; … *purses as weapons*; … *pens as weapons*; … *jackets as blankets*; … *pine boughs as mattresses*.

[3] I should stress here that, in my opinion, McKinstry's diagnosis of the "all-or-nothing" nature of the Turing Test is not adequate. See e.g., Turing (1950, p. 442); Newman,Turing, Jefferson, Braithwaite (1952, p. 5) and discussion in the Summary of this paper.

[4] An interested reader may review comparisons of various AI models' performance against modern AI benchmarks by visiting https://all-in-one-ai.co/rankings or https://benchmarklist.com/ (access 14.11.2025).

[5] ChatGPT performance against the CommonsenseQA benchmark, among others, is reported by Bang et al. (2023). They conclude that "ChatGPT shows surprisingly good commonsense reasoning capability, perhaps due to its large parametric memory." (Bang et al., 2023, p. 681).

[6] https://code.google.com/archive/p/mindpix/ (access 12.03.2024).

[7] For the data randomization, an online random number generator was used (https://www.random.org/integers/).

[8] https://osf.io/2ux58/?view_only=94e0545579d64faf885d91953d0656b7

[9] This is in line with the results of the MIST study presented in (Łupkowski, Jurowska, 2019). In this study, participants (N=126) played the role of judge in MIST and their task was to evaluate how human-like the responses provided during the test were. Judges were far from reaching agreement over the answers provided to MIST questions. Several judges confronted with one question and a yes/no answer to this question may disagree on how to evaluate the answer.

[10] It is worth mentioning here that Davis reports that unexpected artifacts are observed also for modern AI benchmarks – see the list of examples presented in Table 1, p. 12 of (Davis, 2023). What is more, as Davis (2023, p. 9) points out, these artifacts may not be noticed at once by designers: "For instance: The task in the datasets SNLI (…) and MultiNLI (…) was, given two sentences P and Q, determine whether Q is entailed by P, contradicts P, or is neutral with respect to P. It was later found (…) that, due to the strategies used by the crowd workers who built the dataset, there are often verbal clues in Q that allow the correct answer to be guessed even without seeing P."

[11] See, e.g. Gonçalves (2023a, 2023b); Floridi and Chiriatti (2020); Echavarría (2025).

[12] This is e.g., forgotten in the University of Reading press release on the 8th of June 2014, when the chatbot Eugene Goostman took part in the Turing Test 2014 event held at the Royal Society in London and won the main prize. We read: "If a computer is mistaken for a human more than 30% of the time during a series of five minute keyboard conversations it passes the test. No computer has ever achieved this, until now. Eugene managed to convince 33% of the human judges (…) that it was human." https://www.reading.ac.uk/news-archive/press-releases/pr583836.html. The same suggestion is expressed in articles like "Google's AI passed a famous test – and showed how the test is broken" (https://www.washingtonpost.com/technology/2022/06/17/google-ai-lamda-turing-test/) or "ChatGPT passes the Turing Test" (https://mpost.io/chatgpt-passes-the-turing-test/).

[13] Remembering this would allow for avoiding many misleading ideas or oversimplifications observed for the Turing Test debate. Like e.g. Sampson's (1973, p. 173) or Putrill's (1971, p. 169) claims that – according to Turing's "prediction" – we should have "thinking" machines till 2000.

[14] "Even if systems cannot be distinguished in a Turing Test, the real acceptance of a system's being intelligent will be cultural rather than technical." (Watt, 1996, p. 8).

[15] "[…] Turing's view implies that in a society holding a prejudiced, chauvinistic attitude toward machinery, machines cannot be perceived as intelligent entities, by definition. However, if such a society underwent a process in which its a priori attitude toward machinery changed, intelligent machines would become a logical possibility. […] For Turing, the imitation game is not a test for intelligence, but a technological aspiration whose realization may involve a change in society's attitude toward machines". (Danzinger, 2022, p. 23).

[16] It is Copeland (2004, p. 2), who pointed out that the "Intelligent Machinery" report contains "the earliest description of (a restricted form of) what Turing was later to call the 'imitation game' and is now known simply as the Turing Test." See also discussion in (Łupkowski, 2019).