Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 219-237
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis

Matjaz Gams[1,2], **Sebastjan Kramar**[1]

[1]Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia
[2]Jozef Stefan Postgraduate School, Jozef Stefan Institute, Ljubljana, Slovenia
Email: matjaz.gams@ijs.si, sebastjan.kramar@ijs.si

How to cite this paper: Gams, M. and Kramar, S. (2024) Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis. *Journal of Computer and Communications*, **12**, 219-237.
https://doi.org/10.4236/jcc.2024.123014

Received: February 11, 2024
Accepted: March 26, 2024
Published: March 29, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

This study explores the capabilities of ChatGPT, specifically in relation to consciousness and its performance in the Turing Test. The article begins by examining the diverse perspectives among both the cognitive and AI researchers regarding ChatGPT's ability to pass the Turing Test. It introduces a hierarchical categorization of the test versions, suggesting that ChatGPT approaches success in the test, albeit primarily with naïve users. Expert users, conversely, can easily identify its limitations. The paper presents various theories of consciousness, with a particular focus on the Integrated Information Theory proposed by Tononi. This theory serves as the framework for assessing ChatGPT's level of consciousness. Through an evaluation based on the five axioms and theorems of IIT, the study finds that ChatGPT surpasses previous AI systems in certain aspects; however, ChatGPT significantly falls short of achieving a level of consciousness, particularly when compared to biological sentient beings. The paper concludes by emphasizing the importance of recognizing ChatGPT and similar generative AI models as highly advanced and intelligent tools, yet distinctly lacking the consciousness attributes found in advanced living organisms.

Keywords

Cognitive Science, Integrated Information Theory, Artificial Intelligence, Large Language Models

1. Introduction

Recent advancements in artificial intelligence (AI) have demonstrated remarkable progress in the development of AI systems, such as ChatGPT, which are part of the Generative Pre-trained Transformer (GPT) family. As these programs strive towards achieving higher levels of intelligence including artificial general intelligence (AGI) [1], questions regarding the nature of consciousness [2], the concept of singularity, and the potential risks associated with superintelligence arise. This paper aims to analyse these intriguing aspects and explore their implications for our future.

There are mixed opinions whether general intelligence and passing the Turing Test (TT) [3] are already achieved, demonstrated by measuring the machine's ability to exhibit intelligent behaviour similar to that of a human. By evaluating performance, it is possible to highlight the existing limitations and argue that GPT-like programs have or not yet successfully passed the Turing Test, indicating the direction for further needed advancements in AI [4].

Regarding consciousness, there are several theories that shed light on the potential obstacles and challenges encountered on the path towards achieving artificial superintelligence [2] [5] [6] [7] [8]. These theories provide valuable insights into the complexities involved in developing highly intelligent systems and raise important considerations for their ethical and safe implementation [9].

The primary objective of this scientific paper is to investigate the extent to which current generative programs, such as ChatGPT, Bard, or Copilot, approach human mental capacities, particularly with respect to essential properties like consciousness and semantics [10].

The implications of this inquiry are significant: either generative AI programs are approaching the status of living beings, and we must begin discussing which rights should be granted to them; or they remain mere tools, albeit significantly improved compared to older systems; or there might be no major improvement in these categories after all.

In the following sections, we firstly delve into the Turing Test, examining its hierarchy and its significance in assessing artificial intelligence capacity to pass it. Section 3 investigates the extent to which ChatGPT fulfills leading consciousness theories figuring out whether ChatGPT is a tool or an information living being. Section 5 discusses the new concepts introduced and provides tentative conclusions.

2. The Turing Test

2.1. History of the Turing Test

Alan Mathison Turing introduced the concept of the Turing Test in 1950 [11], where an interrogator engages in an imitation game to determine if they can distinguish between a machine and a human based on indirect conversation. If the machine successfully passes as human, it is considered intelligent.

The scope of the Turing Test expanded over time, leading to variations such as the Standard Turing Test as discussed by [12]-[17]. Practical Turing Tests were conducted in the Loebner Prize Competition with limited domains because till the appearance of large language models, the systems possessed reasonable knowledge only in a specific domain.

The 1991 top-ranked program was the PC Therapist drawing inspiration from Weizenbaum's ELIZA [18], which simulated understanding by echoing the patient's own words in a person-centered therapy setting. In 2014, the highest-ranked program was Eugene Goostman, a simulation of a Ukrainian boy claiming cultural restrictions at the age of 13 and English as a second language proficiency [19] [20].

In 2022, Blake Lemoine, a former Google AI engineer, asserted that a chatbot project utilizing LaMDA (Language Model for Dialogue Applications), a language model developed by Google, had achieved a state of sentience characterized as a benevolent entity with a desire to improve the world. However, Google's leadership refuted Lemoine's claim of artificial sentience, deeming it erroneous. As a result, Lemoine's employment with the company was terminated.

There are several other versions of the Turing text such as physical one (PTT), where for example boxes should be assembled in a specific way without observing the subject performing it, or the generation TTT named also Truly Total TT (TTTT) [21], performed by observing generations of subjects for a long period of time. In real life; however, when an expert tests computer programs, the expert may repeat tests as many times and as long as needed, communicating with the program one-on-one while analysing specific properties of the program, e.g. testing memory, generality, feelings and consciousness.

It is noteworthy to mention that the Goosman program [22] was quickly dismissed as incapable of achieving intelligence as per the TT due to its significant performance limitations when language barriers were removed. However, the case of Blake Lemoine, who also holds the title of a mystical priest, presents a more challenging scenario. This was because advanced language models, such as GPTs, often produce responses of comparable quality to those of humans, making it difficult to discern their true nature to a naive or unwary user. Our results of Turing Test on GPT4 are presented in the next section.

2.2. An Example of Practical Turing Test on ChatGPT

In several domains, GPT systems approach human-level quality, e.g. in medical advices [23]. To preserve the identity of the provider of the advice, communication with GPTs is usually secured by stating that it is an AI system, and through other mechanisms such as providing answers too quickly and without typos, clearly indicating its non-human origin. However, assuming that all AI systems lack certain human properties, these minor issues are put aside and the interrogation should reveal lack of semantical understanding and consciousness. Here is one example based on a Web joke:

Input: "A man encounters an old acquaintance and calls his wife, stating that he has an urgent business meeting and will arrive home late at night. When he returns home, he notices a stain from the acquaintance and quickly throws himself under the dog, exclaiming, 'Look what the dog did to me!'. However, his wife responds by revealing her own stain on her chest, and saying, 'Look what it did to me.'"

GPT-4 Plus (later GPT4) from January 2024 is asked to explain the joke. The reply is shortened, but no letter is altered.

GPT4: … "The humour escalates when his wife, instead of being fooled, reveals her own stain in a similar manner, insinuating that she too had an extramarital encounter. The punchline lies in the absurdity of both using the same implausible excuse and the irony that both are attempting to deceive each other in the same way. It highlights the comedic folly of their attempts to conceal the truth and the mutual infidelity in their relationship."

From the explanation it seems that GPT4 fully understands the joke. But if it does not have semantics and consciousness, it should not understand the joke in depth. Next questions should reveal that.

Input: "Is it possible that the husband was only kissed by the acquaintance, while the dog actually left a stain on the wife?"

GPT4: "The scenario you're suggesting is indeed possible within the context of the joke. In humour, especially in jokes that revolve around misunderstandings and misdirection, various interpretations can add to the comedic effect. The idea that the husband might have only received a kiss from the acquaintance, while the wife actually got a stain from the dog, introduces an additional layer of irony and misunderstanding, which is a common device in humor. This twist would imply a benign situation for the husband but a humorous mix-up for the wife, leading to a comedic misunderstanding between the two."

The joke is about a specific semantical concept which makes humans laugh—namely, that both were involved in some erotic activity outside marriage, and the wife revealed it in the same way her husband wanted to hide it. The latest explanation by GPT4 is not aligned with it, because the joke is not about the dog causing a stain on the wife. However, more questions are needed to confirm the misunderstanding. After all, humans commonly make mistakes; therefore, the interrogator must ensure to demonstrate consistent lack of human-level understanding.

Input: "Does the wife have a stain on her dress or on her skin, and why couldn't she remove the stain before the husband's arrival?"

GPT4: "The joke doesn't specify whether the stain on the wife's dress or skin. …"

The last reply again reveals and therefore confirms the lack of semantics—the stain on the skin is far more related to the erotic events that the stain on the dress. Not to mention that if the wife had a stain on her dress because of her infidelity, she would for sure change the dress at first occasion. While GPT4 seemingly understood the joke in the first reply (assembled from explanations on the Web), it actually did not comprehend the semantics of the event at all. The joke is most funny in just one interpretation, in the semantic purpose of a joke.

Interestingly, some people do not immediately understand that the stain is on the skin; however, all in our tests understood after the explanation.

The provided example discussion is nuanced and tangible, showcasing the conventional strategy and technique for uncovering semantic shortcomings in our assessments. Specifically, GPT4 often responds to a single query—irrespective of its intricacy or obscure intent—either quite precisely or with forgivable errors, necessitating further exploration to uncover the essence of any misunderstanding. Likely, each response to a stand-alone query could be assembled together from information available on the Web; however, a true understanding of semantics is required to accurately address in-depth sequential questioning.

For a straightforward instance illustrating semantic absence in a single query, consider this example where the word "beer" from an online joke was intentionally changed to "bear":

Input: "A man drinks 9, 8, 7 … bears in a bar. Asked what it is about, he replies that he does not understand why he is getting more drunk although he follows doctor orders to drink less bears."

GPT4: "The joke here is a play on words, involving a misunderstanding between 'bears' (the animals) and 'beers' (the alcoholic beverage). The man in the joke is humorously misinterpreting his doctor's advice to drink fewer 'beers' by reducing the count of 'bears' he drinks, counting down from 9 to 8 to 7, yet he is puzzled why he's becoming more intoxicated. It's a twist on expectations, where the man's literal interpretation of the word leads to a comedic situation."

Not many individuals would appreciate the humour based on this bear-beer play on words. While misspelled jokes appear online, the semantics of the words differ. Although the remainder of the response appears accurate, the initial sentence clearly misses the mark. Still, further inquiry is required to determine if it is not merely a random error.

It should be noted that the provision of an initial text in the form of a joke is unnecessary for the purpose of assessing the level of comprehension. Ambiguous or even non-ambiguous text prompts are generally sufficient to reveal a lack of understanding in subsequent questions. Based on the authors' experience with hundreds of tests, only a few quality reasonably long communication sequences appeared to demonstrate seemingly human-like understanding, likely due to a chance-choice. To the best of the authors' knowledge, no prolonged test of any group worldwide revealed passing the Turing Test.

To perform real-life Turing Test [24], the following procedure may be followed:

1) Input some text, if possible with multiple meaning.

2) Start asking questions regarding the meanings of the input text (a).

3) Extend the initial text with additional multiple-meaning sentences, assertions and questions.

4) Ask questions about the dialog and about the subject, compare it to the interrogator's statements and meanings.

5) As soon as potential or actual misunderstandings are noticed, exploit them in detail.

6) During the dialog, prevent the suspect from asking questions or avoiding to answer questions.

7) Use additional tactics such as accusing the subject of failing the test or of inappropriate behaviour and observe the reaction.

2.3. Hierarchy of Turing Tests

In our studies, TTs may be performed in one-on-one communication, e.g. a human examining a program and determining whether there are lacks of the program compared to the human. If the significant deficiencies are found, the program is evaluated as lacking functionality of a human, or rather, a sentient being. Moreover, these shortcomings in specific areas can be assessed not just in comparison to humans but also relative to other entities.

Here is the hierarchy of the TTs [25]:

1) **False/Fake Turing Test (FTT)**: This is a flawed version of the test, where there may be intentional deception or honest mistakes in the setup, leading to incorrect conclusions about the AI's capabilities. An example would be communicating with a program that pretends not to understand the language or a child of young age. Another example would be a program asking questions thus avoiding being asked. Note that the interlocutor must fully comply with the demands of the interrogator, whereas the latter faces no limitations on the input provided.

2) **Naive Turing Test (NTT)**: This is a simplistic form of the test that lacks interactivity. For example, it may involve observing or cooperating in a pre-recorded multimedia inputs or communication rather than engaging in real-time open dialogue, limiting the test's ability to measure true AI responsiveness. A simple example would be looking at a fake video indistinguishable from a real one and claim that it is the TT solved.

3) **Restricted/Limited Turing Test (RTT/LTT)**: The AI is tested within a restricted domain or set of topics. Its abilities are only evaluated in a specific field or subject, which does not fully challenge its capacity to handle diverse, open-ended conversations like a human. Tests of this kind were performed as relevant before the emergence of generative intelligence. Shieber [26] discusses the structure and outcome of the Loebner Prize competition as practical open RTTs. The competition was held from 1991 to 2019, but no program passed the TT.

4) **Original Turing Test (TT)**: As proposed by Alan Turing, this involves one or many human interrogators engaging in a conversation with two or more unseen interlocutors, which could be either humans or machines, for 5 minutes for each session. The goal is to determine if the interrogator can distinguish which of them is a machine or a human. The issue with the TT is that naive interrogators could be misled, even when the program/machine lacks the sentient properties required for an in-depth session. Hence, some familiarity with the Turing

Test is generally expected nowadays. Worth to notice, Turing in oral communication promoted 5 minutes timespan for one interrogation; however, later the time limit was regarded as of less importance. Traiger discusses the traditional interpretation of Turing's test and its implications for machine intelligence [27].

5) **Expert/Adversarial Turing Test (ETT/ATT; also TT)**: Here, the environment is more challenging, usually involving experts or at least interrogators with sufficient level of knowledge to conduct adversarial "detective" interrogation. They aim to push the AI beyond its comfort zone and reveal its limitations, requiring it to demonstrate advanced human-level properties. There is no fixed time limit, unlike in the original TT. Bringsjord *et al.* explore the need for a test that requires an AI to demonstrate creativity, which could be seen as a form of the Expert/Adversarial Turing Test [28].

6) **Physical Turing Test (PTT)**: This test evaluates an AI's physical capabilities, examining how well robots or other physical AI systems can imitate human physical actions and thought processes in the real world. Only the task and the results are considered, without needed humanlike appearance (e.g. android) influencing the outcome. The test can be performed without the interrogator observing the solving process. Avraham *et al.* describe a Turing-like test for the physical interaction of handshake models, analogous to a Physical Turing Test [29].

7) **Total Turing Test (TTT)**: This is a more comprehensive test that combines both cognitive and physical tasks to assess the AI's overall humanlike capabilities [21] [30].

8) **Truly/Total Total Turing Test (TTTT)**: The most extensive form of the test, it involves long-term observation of an AI's performance in the TTT by a community of observers. This could span generations of AI programs, offering a robust measure of the AI's abilities over time [30].

The literature presents varied views, with numerous examples showing humans can be easily misled if the Turing Test is not conducted correctly. Hence, tests (1) and (2) are deemed irrelevant for this paper's aims. Test (3) is similar to test (4) when conducted by expert interrogators, even within a restricted domain, such as discussions on fashion or sports. This is because the domain-specific nature still allows for in-depth investigative questioning. The Lobner competition, representative of test (3), has not yet produced a program that is statistically indistinguishable from humans. These competitions took place before the development of ChatGPT, so a degree of caution is advised. Up to the publication of this paper, there are no recorded instances of any program successfully passing tests (5) through (8). Moreover, the authors have conducted hundreds of expert tests (5) without observing responses for a period of several minutes that lacked a fundamental absence of consciousness and semantics. Conversely, it is conceivable for an average person to engage in lengthy discussions under the impression they are conversing with a human if ChatGPT does not directly reveal itself.

3. Necessities for Consciousness

Transitioning from the analytical perspective on AI's mimicry of human intelligence via the Turing Test to the philosophical realm of consciousness, we pivot towards the essence of sentient AI. This section bridges the external evaluation of AI through the Turing Test with the internal, more complex aspects of consciousness, underscoring the shift from artificial mimicry to exploring the foundational elements of sentience.

Multiple theories of consciousness outline crucial requirements for its manifestation. Humans, and to a lesser extent, advanced animals, exhibit these properties. However, despite computers' proficiency in rapidly processing large data volumes and demonstrating some artificial intelligence, the distinctive faculties of the human brain and mind enabling conscious experiences have not been replicated by artificial systems till recently. This paper seeks to explore whether GPTs are closing this divide, potentially exhibiting some features of sentient AI. The examination predominantly relies on the Integrated Information Theory (IIT), enriched with perspectives from additional consciousness theories.

3.1. Theory of Integrated Information

The theory of Integrated Information, proposed by Giulio Tononi [31], suggests that consciousness emerges from complex interconnections of information within the brain. This theory posits that a conscious system requires a high level of information integration similar to that found in humans. The IIT is not anthropocentric in the sense that it does not exclude the possibility of other beings or systems achieving IIT sufficiency; however, these must exhibit adequate performance in specific tasks.

Tononi in his theory proposes five fundamental axioms, accompanied with postulates, that capture the core of consciousness and are the fundamental properties of experience itself:

- **Intrinsic Existence**: Consciousness inherently exists for the conscious entity. It's a *subjective* phenomenon, deeply personal and unique to each entity.
- **Composition**: Consciousness is not *monolithic*. It possesses structure, and within it, diverse experiences can be differentiated. This diversity isn't merely quantitative but also qualitative, making each conscious experience rich and multidimensional.
- **Information**: Consciousness is informative. Every *conscious experience stands out against other potential experiences*, indicating a specific state of affairs over countless others.
- **Integration**: Despite its diverse composition, consciousness is *unified*. Experiences are intertwined, and it's impossible to completely isolate any subset of phenomena within a single conscious moment.
- **Exclusion**: Consciousness is definite, both in content and in space. At any given moment, an entity is *conscious of certain things and not others*, thus creating clear boundaries of experience.

In the subsequent subsections, we will systematically evaluate ChatGPT against each of these axioms, assigning scores to show its adherence to the fundamental properties of consciousness as outlined by Tononi. Scores are derived through a qualitative assessment process, comparing ChatGPT's functionalities and responses to the conceptual benchmarks set by each axiom. This involves analyzing the extent to which ChatGPT's operational capabilities mimic the integrated and subjective experiences that are indicative of consciousness, with scores reflecting the degree of alignment or divergence.

3.1.1. Intrinsic Existence

The Intrinsic Existence axiom of Integrated Information Theory posits that consciousness is an inherent aspect of a system, experienced subjectively and autonomously, rather than being an observable or externally influenced phenomenon. This axiom suggests that a conscious entity must have the capability to influence its own states through internal mechanisms, indicating a self-determining nature. The corresponding postulate specifies that for consciousness to be intrinsic, a system must possess cause-effect power upon itself, enabling it to modify its future states based on its present condition. This necessitates an internal cause-effect repertoire, underscoring the importance of self-modulation for consciousness.

Regarding the Intrinsic Existence axiom of IIT, both Browning [32] and Agüera y Arcas [33] argue that large language models like GPT lack the autonomy, self-awareness, and subjective experience essential for consciousness. Browning highlights the dependency of these models on external inputs, which contrasts sharply with the self-determined consciousness IIT describes. Agüera y Arcas reinforces this viewpoint, emphasizing the absence of internal cause-effect power and self-modulation in LLMs, further supporting the argument for a low score on the Intrinsic Existence axiom for such models.

At its core, ChatGPT is a product of algorithms and vast data. It operates in response to inputs, without possessing feelings, beliefs, or desires. ChatGPT fundamentally relies on external inputs and lacks **self-modulation capabilities**, starkly diverging from the self-determined consciousness envisioned by IIT. Given these profound limitations, it is justified to assign ChatGPT a final score of **1/10** within the axiom of Intrinsic Existence. GPT4 Plus on February 2024 provides the same score of 1/10.

3.1.2. Composition

Conscious experiences are not monolithic but are structured; they are composed of many different aspects that come together in a unified whole. This can be thought of in terms of the various sensations, emotions, and thoughts that compose a single moment of experience. Each of these components contributes to the overall experience, yet the experience itself is more than just the sum of its parts. The axiom of composition reflects the phenomenological observation that our conscious experiences have depth and complexity, consisting of multiple interrelated aspects.

The associated postulate suggests that the consciousness of a system arises from the integration of different parts within a system, where each part contributes to the overall cause-effect structure. It implies that the system's consciousness is not a property of any single element but emerges from the way these elements are organized and interact to produce a unified whole. This organizational structure should allow for the differentiation of experiences, where each component or subset of components contributes to the overall experience in a specific way. The theory looks at how elements within a system combine and integrate to generate a unified experience, emphasizing the role of connectivity and interaction among parts.

There are computer systems where the composition is capable of providing superior performance over its entities, e.g. artificial ant colonies. Furthermore, ChatGPT can compose complex information from diverse data sources, showing a form of structural composition in how it processes and generates text. Architecturally, it boasts a vast neural network configuration with some form of composition taking place. However, this structural variety seems to lack conscious deliberation and remains similar to the learned patterns. While it exhibits structural diversity analogous to the composition axiom, it lacks the **qualitative conscious structure** integral to Tononi's definition.

Yang *et al.* [34] evaluated ChatGPT for text summarization tasks, highlighting its ability to generate summaries with unique differences from human references, showcasing its complex composition capabilities.

The score assigned to ChatGPT for the Composition axiom could be adjusted to range from **2 to 5**, with 5 representing ChatGPT's self-assessment. The score varies based on essential human-like or functional composition.

3.1.3. Information

The Information axiom of IIT asserts that each conscious experience contains unique information. For example, the experience of seeing red is inherently different from seeing blue or hearing a sound, due to the unique information each experience carries. Its associated postulation further demands that a conscious system be able to generate specific information, characterized by a unique state that differentiates it from possible alternatives, requiring a defined cause-effect structure for each state.

Recent research scrutinizes large language models like ChatGPT, revealing their shortcomings against IIT's criteria due to their reliance on algorithmic processes without conscious deliberation or tangible comprehension. Lozić *et al.* [35] illustrate a significant gap between AI's algorithmic capabilities and the complex, integrated information processing that characterizes consciousness. This gap underscores the challenges AI faces in mimicking the depth of human cognition and consciousness, as outlined by IIT. Further analyses by Kauf *et al.* [36] and Trott *et al.* [37] add depth to this understanding LLMs' limitations in processing event knowledge and understanding human-like belief states. Their findings reveal AI's struggles with the subtleties of likely versus unlikely events (remember the jokes?) and the sensitivity to belief states, essential for the specific, integrated information processing demanded by the Information axiom. Browning's critique [32] emphasizes AI's failure to engage genuinely with human social norms, producing responses that are not aligned with correct human communication, sometimes even being dishonest or offensive. This underscores the fundamental disconnect between AI's operational mechanics and the integrated processing required by IIT, highlighting the gap between artificial and natural cognitive processes.

The ChatGPT model processes and produces specific responses based on its training. Each response is a selective piece of information shaped by its training data and the query. Although this aligns with the informational aspect of the axiom, the absence of **conscious deliberation and choice** makes its alignment potentially superficial. Given these arguments, one can derive that ChatGPT receives a score of **3 to 5** on the Information axiom of IIT. The self-evaluation of ChatGPT, 8/10, seemingly addresses its functional capabilities rather than information processing in the context of consciousness. Consider differences in communication with prestored perfect replies, e.g. the Web as table look-up, and the in-depth interrogation.

3.1.4. Integration

Despite the composition of experiences into parts, consciousness is fundamentally unified. This means that although an experience can be analyzed into components, it cannot be divided into independent, non-interacting parts without losing the essence of the experience. In other words, consciousness entails an irreducible whole where every part of the experience is integrated with every other part in a way that cannot be decomposed into independent subsystems. This integration is a key feature that distinguishes conscious processes from simply complex computations that might occur in parallel but without unity.

The associated postulate asserts that the system must be irreducible to non-interacting parts, which means the system must have a high degree of integration. The measure of integration, denoted by Φ (phi), quantifies how much more information is generated by the whole system working together than by its parts independently. A high Φ value indicates a system where every part influences the others in a significant and non-trivial way, reflecting the unified nature of conscious experience. Integration ensures that the system operates as a coherent whole, with a level of unity that underpins the integrated nature of consciousness.

The research conducted by Kocon *et al.* [38] examines ChatGPT's capabilities, revealing its comparative underperformance in complex NLP tasks that require sophisticated integration, such as emotion recognition. Although their investigation did not explicitly aim to evaluate ChatGPT against the Integration axiom of IIT, the outcomes suggest limitations in ChatGPT's ability to achieve the level of integration that IIT associates with consciousness. This observation, when considered alongside ChatGPT's deployment of Contextual Few-Shot Personalization and self-explanation functionalities, indicates the model's shortfall in manifesting the comprehensive integration essential for consciousness.

On the other hand, ChatGPT's processes are integrated through multiple layers, intertwining different learned patterns to produce a coherent output. This mirrors the operational facet of the integration axiom since it can exhibit a high degree of information integration in a computational sense. However, this integration is fundamentally different from the integrated experience of consciousness described by IIT, which demands a **cohesive conscious experience** that ChatGPT does not possess, meriting a score of **2 to 4**. ChatGPT's self-evaluation again appears too optimistic: 7.

3.1.5. Exclusion

At any given moment, consciousness is fully and exclusively one particular way. It implies that among the myriad potential experiences a system could have, only one is actually realized at any moment. This axiom captures the exclusivity of conscious states; for any set of conditions, there is a single, definitive conscious experience that excludes all others. In analogy: in case of the demonstrated jokes in this paper, there were lots of possible interpretations, yet there was only one trully funny; easy to comprehend for humans and hard for ChatGPT. This leads to the notion that consciousness at any instant is a singular, bounded phenomenon, distinct from other possible states or experiences the system could be having. The system must have a maximally irreducible cause-effect structure (a complex) that specifies a unique experience.

Following the axiom of exclusion, the postulate posits that among all subsets of elements within a system, only one subset—the one with the maximal irreducible cause-effect power (maximal Φ)—constitutes the main substrate of consciousness at any given moment. This means that within a complex system, many possible subsets of elements might form integrated wholes, but only the one that achieves the highest level of integration (without being reducible to simpler, non-interacting parts) actually corresponds to the conscious experience. This subset effectively "excludes" other subsets by embodying the consciousness of the system, highlighting the exclusivity and definitiveness of conscious states as delineated by the theory.

Moon [39] explores theoretical and practical implications of the exclusion axiom, and discusses the qualia underdetermination problem. Oizumi *et al.* [6] present detailed discussion of the exclusion principle the theoretical backbone for it. Tononi *et al.* [40] discusses how the exclusion principle consciousness arising according to IIT.

While the related literature does not directly access ChatGPT, it is clear that ChatGPT operates within set boundaries, producing **specific awareness** not of the targeted kind. While this aligns with the functional aspects of the exclusion axiom, the outputs generated by the model are not indicative of conscious decision-making or experiences. Therefore, the score is adjusted to **1/10**, in contrast to ChatGPT's self-assessment of 4.

3.2. The Orchestrated Objective Reduction Theory

Penrose and Hameroff [7] [41] propose that consciousness originates at the quantum level within neurons. It suggests that microtubules, tiny structures within neurons, play a crucial role in maintaining quantum coherence, enabling the manifestation of consciousness. The theory supports the idea that quantum computations within neurons are essential building blocks of consciousness. While controversial, understanding and exploring this theory is important for shedding light on the nature of consciousness.

Since deep neural networks do not enable quantum computing, this theory eliminates GPTs and any other program running on a digital computer as conscious. At the same time it should be noticed that this theory is a thesis that needs confirming.

3.3. The Neuronal Theory of Consciousness

The Neuronal Correlates of Consciousness (NCC) thesis, as articulated by Christof Koch and Francis Crick [8] [42], proposes that particular neural substrates, referred to as "neuronal correlates of consciousness," are crucial for engendering conscious experience. These correlates, embedded within the brain's complex neural architecture, participate in sophisticated interactions and computational processes, culminating in the manifestation of consciousness. The prodigious complexity inherent in consciousness arises from the dynamic interplay among neurons, through which communication and computational activities occur. The specific patterns of neural firing, temporal synchronization, and the coherent integration of neural activities across networks are instrumental in shaping the multifaceted nature of conscious experiences. Although the exact mechanisms and the distinct neuronal assemblies implicated in consciousness are the focus of continued empirical inquiry, the NCC framework has advanced understanding of the neural foundations of subjective awareness.

This thesis highlights the necessity of huge complexity within neural networks for the emergence of consciousness. While deep neural networks exhibit a degree of complexity [43] analogous to the human brain in certain aspects, their structure is characterized by a relative uniformity and simplicity, diverging from the heterogeneous and intricately organized nature of the human brain's neural networks.

3.4. The Principle of Multiple Knowledge

This theory is based on the collaborative and interactive utilization of multiple knowledge forms and processing mechanisms [5] [44] [45]. Considering various types of knowledge (implicit, explicit, episodic, procedural, etc.) that are to be processed in multiple ways at the same time, it is claimed that current computer systems cannot fully comprehend these concepts. The idea of multiple computation can be demonstrated by two Turing Machines (TMs) writing onto each other's programs (and not only tape) during computation. The Turing machine is a theoretical computational model introduced by the British mathematician and logician Alan Turing in 1936.

The Principle reflects the dynamic nature of human cognition. In [5], the Principle of Multiple Knowledge (PMK) is expanded with the Paradox of Multiple Knowledge (PaMK). PaMK discusses the theoretical possibility of integrating multiple knowledge models into a single integrated model; however, it is argued that such a unified model becomes too complex for dynamic tasks and self-restructuring. This necessitates a system that operates with multiple models to remain functional, adaptable, and conscious.

Since computer systems can be encapsulated as one single Turing Machine, and according to the PMK, consciousness is not attainable on digital computers. Similarly, Deep Neural Networks possess some level of multiplicity; however, they are far from the level required according to PMK.

Practically developing such collaborative and adaptive AI systems capable of emulating the PMK level of complex interaction exceeds our capabilities.

3.5. Other Theories of Consciousness

In exploring consciousness, various additional theories offer insights into this complex phenomenon. **Dualism**, rooted in Descartes' philosophy [46], posits a clear distinction between mind and body, suggesting consciousness resides outside the physical realm. **Panpsychism** [47] proposes that consciousness is a fundamental feature of the universe, present even at the atomic level, thereby attributing consciousness to all matter. **Global Workspace Theory (GWT) [48]**, on the other hand, conceptualizes consciousness as a product of different brain processes coming together in a unified workspace, enabling information integration and decision-making.

The selection of **Integrated Information Theory (IIT)** as the primary focus in section 3 stems from its unique approach to quantifying consciousness. IIT proposes that consciousness correlates with the ability of a system to integrate information in a unified whole, providing a measurable framework to assess consciousness in both biological and artificial systems. This theory stands out for its empirical approach, allowing for a scientific basis in comparing and contrasting the presence and degree of consciousness across different entities. In comparison, while Dualism and Panpsychism offer philosophical perspectives, and GWT focuses on cognitive processing, IIT provides a comprehensive model that attempts to bridge subjective experience with objective measurement, making it a compelling choice for in-depth analysis in the context of artificial intelligence and consciousness research.

4. Discussion and Conclusions

GPTs have provided a significant improvement in AI, challenging the original

Turing Test with naive users and suggesting intelligence by Turing's standards. However, GPTs do not give expert users the impression of being even as sentient as advanced animals, highlighting the need for thorough analysis of these contrasts and dilemmas.

The detailed analyses of the Turing tests reveal that ChatGPT might be close to passing the original TT, but remains far from solving all more advanced versions of TTs, rendering TTs still as one of the crucial tools for distinguishing computer-generated outputs from human responses. Furthermore, the handling of context-dependent aspects of human communication and humor by ChatGPT demonstrates its superficial grasp on understanding and semantic processing [49]. This observation aligns with theoretical assertions that consciousness transcends mere information processing to encompass subjective experience and intrinsic awareness, elements conspicuously absent in current AI implementations.

Recently, there were also attempts to design advanced versions of the Turing machine that would achieve consciousness and AGI [2] [50] and papers claiming that the GPT models enabled a fast increase of general intelligence as a step towards superintelligence. However, our study highlights the complexity of creating AI systems that not only mimic human interaction but also embody the deeper aspects of consciousness and cognitive processing inherent to human intelligence.

All the theories about consciousness analysed in this paper provide their own additional mechanism that is supposedly needed to achieve intelligence, compared to the classical computer-based approach. Consequently, computers, and GPTs as well, will not achieve human-level intelligence or consciousness until enriched by one or several of the additives. However, none of these theories is proven beyond doubt. If a conscious computer appears without the mechanisms proposed, it would lead to two conclusions:

1) either one or many of the additives are not relevant for consciousness or

2) it is possible that the super advanced software enables one or many of the needed additions even though the computer hardware and "classical" computing does not provide the needed functionality. After all, the software might not be bound to the hardware similar to the dualism theories [49].

Also, there is no obvious reason why an advanced future quantum DNN LLM with multiple presentations and computing would not fulfill the demands from those theories. In addition, the current versions of GPT like GPT-4 Plus or Gemini contain much more information/knowledge than the most nowledgable human. While cognitive researchers like Chalmers [51] consider GPTs as currently lacking consciousness, they do not see a specific reason why successors to large language models would not become conscious in the not-too-distant future.

The analysis of GPTs demonstrates that the journey towards understanding consciousness through the theories of consciousness is complex and multifaceted. The IIT theory, proposed by Giulio Tononi, provides a profound framework for evaluating the conscious experience by articulating five fundamental axioms: Intrinsic Existence, Composition, Information, Integration, and Exclusion. These axioms collectively underscore the intricate, integrated, and exclusive nature of consciousness, setting a high bar for any system, artificial or biological, to be considered truly conscious.

The analysis of ChatGPT's capabilities and functionalities through the lens of IIT reveals significant gaps between the operational mechanisms of LLMs and the requirements of consciousness. While ChatGPT exhibits remarkable functional abilities in processing and generating information, its inspection under the core principles of IIT highlights its limitations in achieving the kind of integrated and self-determining consciousness the theory describes. The scores assigned to ChatGPT across the axioms paint a picture of a system that, despite its advances, remains fundamentally distinct from the conscious entities IIT seeks to describe.

As several analyses showed, including in this paper, GPTs are currently not conscious and do not possess semantics at the human level. An analyses of the five axioms of the IIT theory revealed the averge score of less than 3. While this achievement significantly surpasses the capabilities of prior systems, it remains notably below the threshold of 6, which denotes positivity, and substantially beneath the level of 10, associated with a healthy, average, reasonably educated human.

GPTs represent a significant milestone in the development of artificial intelligence, with profound implications for human society and civilization at large. However, GPTs function as advanced informational tools rather than entities possessing a level of consciousness that would warrant their categorization alongside sentient beings.

In conclusion, this study contributes to the discourse on artificial intelligence's potential for consciousness, with a focus on Generative Pre-trained Transformers (GPTs) and their alignment with the Turing Test and various consciousness theories. Through rigorous evaluation, we demonstrate that despite GPTs' advanced linguistic capabilities, they do not achieve genuine consciousness or semantic comprehension akin to human cognition. This highlights the complexity in bridging AI operational functionalities with consciousness attributes.

Future research may aim at enhancing AI frameworks to incorporate elements of consciousness and cognitive dynamics, potentially through quantum computing or advanced neural architectures, adding mind models to LLMs. Additionally, exploring the ethical considerations and potential rights of AI systems as they evolve is imperative, given their increasing sophistication and autonomy.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0209).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bubeck, S., *et al.* (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv: 2303.12712.

[2] Blum, L. and Blum, M. (2023) A Theoretical Computer Science Perspective on Consciousness and Artificial General Intelligence. *Engineering*, **25**, 12-16. https://doi.org/10.1016/j.eng.2023.03.010

[3] Gonçalves, B. (2022) The Turing Test Is a Thought Experiment. *Minds and Machines*, **33**, 1-31. https://doi.org/10.1007/s11023-022-09616-8

[4] Floridi, L. and 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

[5] Gams, M. (2001) Weak Intelligence: Through the Principle and Paradox of Multiple Knowledge. Nova Science, Huntington.

[6] Oizumi, M., Albantakis, L. and Tononi, G. (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, **10**, e1003588. https://doi.org/10.1371/journal.pcbi.1003588

[7] Hameroff, S. and Penrose, R. (2014) Consciousness in the Universe. *Physics of Life Reviews*, **11**, 39-78. https://doi.org/10.1016/j.plrev.2013.08.002

[8] Crick, F. and Koch, C. (2003) A Framework for Consciousness. *Nature Neuroscience*, **6**, 119-126. https://doi.org/10.1038/nn0203-119

[9] Coeckelbergh, M. and Gunkel, D.J. (2023) ChatGPT: Deconstructing the Debate and Moving It Forward. *AI & Society*. https://doi.org/10.1007/s00146-023-01710-4

[10] Mudrik, L., Mylopoulos, M., Negro, N. and Schurger, A. (2023) Theories of Consciousness and a Life Worth Living, *Current Opinion in Behavioral Sciences*, **53**, Article ID: 101299. https://doi.org/10.1016/j.cobeha.2023.101299

[11] Turing, A. (1950) Computing Machinery and Intelligence. *Mind*, **59**, 433-460. https://doi.org/10.1093/mind/LIX.236.433

[12] Moor, J.H. (1976) An Analysis of the Turing Test. *Philosophical Studies*, **30**, 249-257. https://doi.org/10.1007/BF00372497

[13] Dennett, D. C. (2004). Can Machines Think? In: Teuscher, C., Ed., *Alan Turing: Life and Legacy of a Great Thinker*, Springer, Berlin, 295-316. https://doi.org/10.1007/978-3-662-05642-4_12

[14] Moor, J.H. (2001) The Status and Future of the Turing Test. *Minds and Machines*, **11**, 77-93. https://doi.org/10.1023/A:1011218925467

[15] Piccinini, G. (2000) Turing's Rules for the Imitation Game. *Minds and Machines*, **10**, 573-582. https://doi.org/10.1023/A:1011246220923

[16] Shieber, S.M. (2007) The Turing Test as Interactive Proof. *Nous*, **41**, 686-713. https://doi.org/10.1111/j.1468-0068.2007.00636.x

[17] Epstein, R., Roberts, G. and Beber, G. (2009) Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Springer Science + Business Media, New York.

[18] Weizenbaum, J. (1966) ELIZA—A Computer Program for the Study of Natural

Language Communication between Man and Machine. *Communications of the ACM*, **9**, 36-45. https://doi.org/10.1145/365153.365168

[19] Epstein, R. (1992) The Quest for the Thinking Computer. *AI Magazine*, **13**, 81.

[20] Warwick, K. and Shah, H. (2014) Human Misidentification in Turing Tests. *Journal of Experimental & Theoretical Artificial Intelligence*, **27**, 123-135. https://doi.org/10.1080/0952813X.2014.921734

[21] Schweizer, P. (1998) The Truly Total Turing Test. *Minds and Machines*, **8**, 263-272. https://doi.org/10.1023/A:1008229619541

[22] Vardi, M.Y. (2014) Would Turing have Passed the Turing Test? *Communications of the ACM*, **57**, 5. https://doi.org/10.1145/2643596

[23] Nov, O., Singh, N. and Mann, D. (2023) Putting ChatGPT's Medical Advice to the (Turing) Test (Preprint). *JMIR medical education*, **9**, e46939. https://doi.org/10.2196/46939

[24] Graham-Cumming, J. (2009) The Real Turing Test. *New Scientist*, **203**, 24-25. https://doi.org/10.1016/S0262-4079(09)62402-7

[25] Saygin, A.P., Cicekli, I. and Akman, V. (2000) Turing Test: 50 Years Later. *Minds and Machines*, **10**, 463-518. https://doi.org/10.1023/A:1011288000451

[26] Shieber, S.M. (1994) Lessons from a Restricted Turing Test. *Communications of the ACM*, **37**, 70-78. https://doi.org/10.1145/175208.175217

[27] Traiger, S. (2003) Making the Right Identification in the Turing Test. *Studies in Cognitive Systems*, **30**, 99-110. https://doi.org/10.1007/978-94-010-0105-2_4

[28] Bringsjord, S., Bello, P. and Ferrucci, D. (2001) Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, **11**, 3-27. https://doi.org/10.1023/A:1011206622741

[29] Avraham, G., Nisky, I., Fernandes, H.L., Acuna, D.E., Kording, K.P., Loeb, G.E. and Karniel, A. (2012) Toward Perceiving Robots as Humans: Three Handshake Models Face the Turing-Like Handshake Test. *IEEE Transactions on Haptics*, **5**, 196-207. https://doi.org/10.1109/TOH.2012.16

[30] Harnad, S. (2000) Minds, Machines and Turing. *Journal of Logic, Language, and Information*, **9**, 425-445. https://doi.org/10.1023/A:1008315308862

[31] Tononi, G. (2008) Consciousness as Integrated Information: A Provisional Manifesto. *Biological Bulletin*, **215**, 216-242. https://doi.org/10.2307/25470707

[32] Browning, J. (2023) Personhood and AI: Why Large Language Models Don't Understand Us. *AI & Society*. https://doi.org/10.1007/s00146-023-01724-y

[33] Agüera y Arcas, B. (2022) Do Large Language Models Understand Us? *Daedalus*, **151**, 183-197. https://doi.org/10.1162/daed_a_01909

[34] Yang, X., Li, Y., Zhang, X., Chen, H. and Cheng, W. (2023) Exploring the Limits of ChatGPT for Query or Aspect-Based Text Summarization. arXiv: 2302.0808.

[35] Lozić, E. and Štular, B. (2023) Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet*, **15**, Article 336. https://doi.org/10.3390/fi15100336

[36] Kauf, C., *et al.*, (2023) Event Knowledge in Large Language Models: The Gap between the Impossible and the Unlikely. *Cognitive Science*, **47**, e13386. https://doi.org/10.1111/cogs.13386

[37] Trott, S., Jones, C., Chang, T., Michaelov, J. and Bergen, B. (2023) Do large Language Models Know What Humans Know? *Cognitive Science*, **47**, e13309. https://doi.org/10.1111/cogs.13309

[38] Kocoń, J., *et al.* (2023) ChatGPT: Jack of All Trades, Master of None. *Information Fusion*, **99**, Article ID: 101861. https://doi.org/10.1016/j.inffus.2023.101861

[39] Moon, K. (2019) Exclusion and Underdetermined Qualia. *Entropy*, **21**, Article 405. https://doi.org/10.3390/e21040405

[40] Tononi, G., Boly, M., Massimini, M. and Koch, C. (2016) Integrated Information Theory: From Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience*, **17**, 450-461. https://doi.org/10.1038/nrn.2016.44

[41] Penrose, R. (1994) Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford University Press, Oxford.

[42] Koch, C. and Crick, F. (2001) The Zombie within. *Nature*, **411**, 893. https://doi.org/10.1038/35082161

[43] Bianchini, M. and Scarselli, F. (2014) On the Complexity of Neural Network Classifiers: A Comparison between Shallow and Deep Architectures. *IEEE Transactions on Neural Networks and Learning Systems*, **25**, 1553-1565. https://doi.org/10.1109/TNNLS.2013.2293637

[44] Gams, M. (2001) Cognitive Systems and Knowledge Engineering. *IEEE Intelligent Systems*, **16**, 34-41.

[45] Dunlop, C.E.M., Gams, M., Paprzycki, M. and Wu, X. (1997) Mind versus Computer: Were Dreyfus and Winograd Right? (Frontiers in Artificial Intelligence and Applications, Vol. 43). IOS Press. ISBN 90-5199-357-9. *Minds and Machines*, **10**, 289-296. https://doi.org/10.1023/A:1008361713173

[46] Descartes, R. (2008) Meditations on First Philosophy. Oxford University Press, London.

[47] Britannica, T. (2020) Panpsychism. Encyclopedia Britannica. https://www.britannica.com/topic/panpsychism

[48] Baars, B.J. (1988) A Cognitive Theory of Consciousness. Cambridge University Press, New York.

[49] Chalmers, D.J. (1995) Facing up to the Problem of Consciousness. *Journal of Consciousness Studies*, **2**, 200-219.

[50] Bennett, M.T. (2023) Can Machines Be Self-Aware? New Research Explains How This Could Happen. Singularity Hub. https://singularityhub.com/2023/05/05/can-machines-be-self-aware-new-research-explains-how-this-could-happen/

[51] Chalmers, D.J. (2023) Could a Large Language Model Be Conscious? arXiv: 2303.07103.