Chinese Chat Room: AI hallucinations, epistemology and cognition[1]

The purpose of this paper is to show that understanding AI hallucination requires an interdisciplinary approach that combines insights from epistemology and cognitive science to address the nature of AI-generated knowledge, with a terminological worry that concepts we often use might carry unnecessary presuppositions. Along with terminological issues, it is demonstrated that AI systems, comparable to human cognition, are susceptible to errors in judgement and reasoning, and proposes that epistemological frameworks, such as reliabilism, can be similarly applied to enhance the trustworthiness of AI outputs. This exploration seeks to deepen our understanding of the possibility of AI cognition and its implications for the broader philosophical questions of knowledge and intelligence.

1. Introduction

The rapid expansion of *large language models* (LLMs)[2] has brought both opportunities and challenges across various fields, from everyday communication to complex decision-making in critical domains such as healthcare and law. Among the most pressing concerns is the phenomenon of AI *hallucinations* – instances where these systems generate false or misleading outputs that are confidently presented as factual information. As LLMs continue to gain widespread use, understanding the mechanisms behind these hallucinations becomes crucial not only for improving model reliability but also for addressing broader epistemological questions.

This paper examines AI hallucinations by exploring how they parallel human cognitive errors and how, despite such parallels, they operate within an entirely different framework of knowledge production. By applying philosophical concepts like reliabilism, the study seeks to clarify what it means for AI systems to *know* something and to what extent we can rely on them to produce trustworthy knowledge.

1 To appear in Special Edition of "Studies in Logic, Grammar, and Rhetoric", *Frontiers of Artificial Intelligence – Philosophical Explorations.*

2 Statements made in this paper about LLMs refer to all large language models, such as various GPT versions, Mistral, Claude, etc., but the public opinion is usually shaped on the basis of ChatGPT, an interface for various GPT versions by Open AI.

2. Defining AI hallucinations

Artificial intelligence *hallucinations*, previously referred to as confabulations or delusions, are unintended responses generated by artificial intelligence that contain false or misleading information that is presented as a fact. For example, an artificial intelligence (AI) system may claim that the capital of France is Zagreb, or that Mona Lisa was painted in the 19th century. Such text may be nonsensical or unfaithful to the provided source input (Ji et al., 2020). AI hallucinations are concerning because they raise safety issues regarding real-world applications, such as in biomedicine or health,[3] or lead to potential privacy violations, especially regarding natural-language generation. Ji et al. (2020) illustrate this problem with a case of a medicine-taking instruction generated by machine translation that is hallucinatory and may have dangerous consequences.

The hallucinatory output of LLMs is not restricted only to textual data, since such hallucinations are analogous to other generative AI system outputs, for example, image generation techniques that produce anomalies like seven-fingered hands. The term itself was used originally in computer vision to describe adding details to an image (Maleki, Padmanabhan & Dutta, 2024). Dziri et al. (2023, p. 5272) highlight that hallucinations occur when a response's factual accuracy cannot be entirely confirmed based on the provided knowledge snippet, even if the information is correct in reality. This also includes personal opinions, experiences, feelings, and subjective assessments that cannot be directly traced back to the source material.

There are two main types of hallucinations: *intrinsic hallucinations*, where the generated output contradicts the source content, and *extrinsic hallucinations*, where the generated output cannot be verified from the source content (Ji et al., 2020). In the first case, the source content may have information about Mona Lisa created in 1503, but an AI system might claim it was, in fact, painted in 1815. In the second case, an AI system might claim that Leonardo Vinci really liked long walks, a fact that might be true, but cannot be verified from the source content.

From a technical standpoint, hallucinations occur from data or from the training and inference

3 See Hatem, Simmons & Thornton (2023) for more details.

part. In the first case, the so-called source-reference divergence (Ji et al., 2020) happens as an artifact of heuristic data collection or due to other machine-learning tasks that contain such divergence, where the model can be encouraged to generate text that is not grounded in the provided source. Some data might be sparse and incorrect and contribute to the model's faulty knowledge as well, especially in highly specialized areas. In the data collection step, working with large-scale datasets is difficult, and some selections of sentences in the data might be faulty or due to human error in the data itself. In the second case, during deep-learning processes, encoders that have the role of encoding text into (usually vectorized) representations, might learn wrong correlations between different parts of data in the dataset or perform wrong decodings as well.

In human psychology, a *hallucination* is usually a perception of some external stimulus that convinces a person that is real, while in fact it is not: a sensory perception without external stimulation of the relevant sensory organ (Waters and Fernyhough, 2016). Similar to various types of AI hallucinations, human hallucinations can be *true hallucinations* and *pseudohallucinations*. The former refers to experiences perceived as real and outside the body, while pseudohallucinations denote hallucination-like experiences occurring within the body. That is, the person can recognize it to be subjective and unreal. There is an analogy between these two processes, but one needs to be aware of how the concept and term of a human *hallucination* carries with it the presupposition of cognition and perception, that might erroneously shape the public's opinion about the current state of artificial intelligence.[4] Namely, attributing a human qualia-like quality can lead to attributing non-existent epistemological states.

Comparable to various psychological and medical ways of battling hallucinations, there have been ways of mitigating AI hallucinations, especially regarding various retrieval-augmented generation methods. Namely, some external knowledge sources or checks are used as an additional step along with the model's output. For example, an additional retrieval-augmented generation (RAG) system – that adds additional knowledge or checks for the model output – might look up some facts from an Internet encyclopedia, and add it to the prompt for the model, adding some novel data that is not present or correctly inferred in the model itself.

4 See Maleki, Padmanabhan & Dutta (2024) for more details about the term and use. A bigger discussion on the term is out of scope of this paper, but I will return to the issue in section 4.

Besides ethical issues, various epistemological problems and issues in the philosophy of mind arise from the notion of AI hallucinations. We will observe three major points. First, since knowledge is tied to cognition, we will start with observing how the standard natural-language understanding problem of Searle's Chinese room still applies to the current status of AI systems as *seemingly* intelligent agents. Next, we will analyze the epistemological uncertainty of AI-generated knowledge and compare various RAG-methods to externalist epistemological theories. Finally, we will compare the current development of AI systems and issues that lie within the ontology of human intelligence, in order to see whether the development of AI is diverging or converging towards its human role model.

3. Examining intelligence: AI, cognition, and human parallels

In various sources and social media, people can observe how intimate we are becoming with the chatbots. We claim that they "tell lies" and "act weird" (Cade, 2023) or that "ChatGPT believes" something.[5] Namely, it is natural to ascribe human mental states to something that might seem as intelligent as human beings. The standard problem of other minds, where our knowledge is always indirect, comes into play. However, the closer the computer is to our notion of *intelligence*, behaving more like a human being than usual, we are behavioristically more inclined to talk about intelligence.

Turing's test (1950) or *the imitation game* was originally conceived as a set of questions designed for a human tester to differentiate a human being from a computer in the form of conversation. Turing started his work by asking "can machines think" and emphasizing that this should begin with the definitions of the meaning of the terms "machine" and "think", but since these definitions might already be begging the question, he proposed the famous imitation game. Turing did not originally want to use it to test the intelligence of machines – no matter how we define what intelligence in that case is – but to replace that question with another – the question of how to recognize what an intelligent agent is – and aid the development of the philosophy of

5 After the submission of this paper, Piedrahita & Carter (2024) and Goldstein & Levinstein (2024) talked about similar issues.

artificial intelligence. Of course, such questions would also require not only conceptual but also ontological development (Krzanowski & Polak, 2022). Standard objections to the test include the linguistic centricity of the test that only relies on language instead of focusing on other cognitive capabilities (Gardner, 2011) and that modern AI research uses different and more accurate methods to test their AI programs: the same way planes are tested by giving them a task of flying, instead of comparing them to birds (Russell & Norvig, 2003).

It is no wonder that Searle's (1980) *Chinese Room* argument focused on the issue of what the notion of AI means when compared to human beings. Searle differentiates between *strong AI*, which actually possesses something like a mind and consciousness, and *weak AI*, which acts *as if* it possesses a mind and consciousness. In a famous thought experiment in which a computer is taught to produce an output or a translation when given a Chinese symbol without actually understanding what is going on, the whole system looks as if it understands Chinese, but there is no actual understanding going on: just, in Turing's words, a plain imitation game. Comparably, *artificial general intelligence* (AGI) is a type of artificial intelligence that performs as well as humans do in various tasks or even better, and is often considered as one of the definitions of strong AI. However, it is still unknown whether consciousness is required in order to perform such tasks in a satisfiable manner.

Cantwell-Smith (2019) similarly argues that AI systems lack the intrinsic capacity for genuine understanding, employing *judgment* as a term for the normative ideal to which we should hold full-blooded human intelligence, in contrast to *reckoning*, the type of calculative prowess at which AI systems excel today. Cantwell-Smith notes that in current neural architectures, the "knowledge" may be encoded in weights distributed across the entire network, and it is not evident nor straightforward to transform the network states into parameters or other states.

Similar to the computational complexity of various computing tasks, the notion of AI-completeness[6] has come into play, by analogously comparing various problems for which artificial general intelligence is required for them to be solved. Yampolskiy (2012) states that a program is AI-complete if it is able to be solved by a hypothetical Human Oracle, a machine

6 For philosophical issues in AI-completeness, see Šekrst, 2020.

that can decide various problems in constant time, i.e., respond to difficult problems instantly, along with the condition that any AI problem can be converted into this specific one by some polynomial-time[7] algorithm. The last condition mimics the case in computational complexity.[8] Namely, **P** problems are a class of decision problems[9] that can be solved by a deterministic Turing machine using a polynomial amount of computation time, while **NP** problems are a class of decision problems that have proofs verifiable in polynomial time by a deterministic Turing machine. ***P*** *vs.* ***NP*** *problem* in computer science asks whether problems that can be *verified* correctly can also be *solved* correctly. Let us illustrate this with an example: it is easy to check whether some number is an answer to your problem, but it is not easy to come up with that number itself. Major computations and algorithms in modern computer science depend on the fact that we still have no polynomial-time solving methods for various problems that take years or even millions of years to solve, including various cryptographic methods, that only rely on brute-forcing, i.e., checking each possible combination, that soon rise exponentially.

One might object by saying that the current status of large language models does not provide enough evidence for strong AI, but it might, as their complexity rises. That is certainly correct, but I am not arguing against actual understanding or the future possibility of consciousness.[10] What is problematic here is to talk about AGI or strong AI in a way that involves human concepts such as *cognition* or *consciousness* since we are talking about extremely different ontologies.[11]

What can we use without anthropomorphized concepts? One solution might lie in ***AI-complete*** problems, which, according to Yampolskiy (2012), include natural language understanding, problem-solving, knowledge representation and reasoning, and vision or image understanding. We can see how a complete AI system is not only a linguistic-based machine whose intelligence can be tested only regarding its language processing but includes other cognitive-like capabilities as well. It seems that talking about AI-completeness seems like a more precise and more

7 Polynomial-time solving, in essence, means the time and resources needed do not rise exponentially or worse, which often requires physically impossible amount of time and space.
8 For a general overview of philosophically interesting issues in computational complexity, see Aaronson (2013).
9 Decision problems appear in mathematical questions of decidability, and refer to those problems that can be answered with yes/no values to the proposed input.
10 One can, however, argue that this may be an instance of an AI hallucination.
11 For a great overview on how to talk about AI ontologies, see Krzanowski & Polak (2022).

philosophically inclined concept than questions regarding consciousness and various imitation games.

Taking that into account, the modern advent of large language models became extremely important not only for philosophy of artificial intelligence and mind but also for epistemology and language studies. Their name points out to their greatest advantage – they were trained on large language-based datasets and are able to produce output similar to human ones since the training dataset included human-produced output not only from official documents and encyclopedias but also from social media, mimicking the way humans reason and talk. People constantly tell large language models not to sound like machine-generated text, and in a certain feedback loop, with each training epoch, the models are getting better at this imitation game. However, with the advent of AI hallucinations, it seems that the imitation game is in one way closer to humans than we thought, sharing our fallibility as both epistemic and intelligent agents. On the other hand, using terms from the human psychology such as *hallucination* already establishes a paradigm of superficial ascribing of mental states to a large language model. What we also do is that we often ascribe *knowledge* as well. Let us observe whether in this case we can talk about justified true beliefs at all.

4. Knowledge and reliability in AI systems

If we consider the classical definition of knowledge as a *justified true belief*, without going into Gettier (1963) problems, it presupposes that artificial intelligence agents possess beliefs, which brings mental states into the picture, a stance which we are still not ready to take. However, it is easy to find various knowledge bases of LLMs and social-media discussions regarding what they *know*.[12] This is, like hallucinations, another issue of using cognitive concepts by analogy. Knowledge, belief, and similar epistemological concepts in AI systems are a part of a different ontology. However, there are interesting epistemological parallels in both types of ontologies that justify the usage of such terms.

12 Search engines show a vast number of results for phrases such as “ChatGPT knows” or various types of mental states such as “ChatGPT believes”, even in papers (see Hintze, 2023).

The rise of large language models also echoes the linguistic turn debates of the 1980s, particularly the question of whether knowledge is objective or constructed through language. For example, Lyotard's (1984) explores how language shapes our understanding of truth, emphasizing the role of narratives in constructing knowledge, and showing a shift to linguistic and symbolic production in the postmodern culture, especially with the development of artificial intelligence. In this context, LLM outputs prompt reflection on whether they represent genuine knowledge or simply linguistic constructions, re-opening long-standing discussions about the relationship between language and truth.

When a human being *knows* a certain proposition, they do it on the basis of certain evidence of past experience, whether it is internal to a person or external, which is a matter of *internalism* vs. *externalism* debate in the epistemology of justifications. Internalists claim that justification is required for knowledge and that the agent's internal states completely determine the nature of justification, while externalists deny at least one of these conditions by: 1) claiming that external facts such as causal chains of state affairs make the belief true, 2) using counterfactual dependence of states of affairs make the belief true, 3) focusing on the reliability of the belief-producing processes, 4) positing whether a belief is likely to be true or not (Poston, 2024).

Reliabilism about justification states that "a belief is justified if and only if it is produced by a reliable psychological process, meaning a process that produces a high proportion of true beliefs" (Goldman, 2012). I will not take any stance in the internalism or the externalism debate as a true one, but I will be taking reliabilism as an example that could be – in principle – applied to an artificial intelligent agent, so one can see what epistemological issues arise that are both issues in epistemology and philosophy of the artificial intelligence. Namely, it is a convenient theory for this comparison since people do not need to possess infallible or certainty-producing processes for the generation of justified beliefs, only *fairly reliable* ones, which excludes standard objections like Descartes' evil demon (Goldman, 2012).

So, can one talk about *reliable* processes of knowledge production in large language models? In machine-learning processes, a model is trained on data. In the case of LLMs, that data comprises a huge portion of the Internet. They are based on the transformer architecture (Vaswani et al.,

2017), which performs repeated transformations of the vector (embedding) representations in order to extract even more linguistic information. A *word embedding* is a representation of a word, typically, a real-valued vector that encodes the meaning such that words that are closer in meaning are closer in the vector space. A standard transformer architecture converts text into tokens, then into embedding vector representations, which are then drawn into the transformer layer.

I have mentioned that AI hallucinations can arise from data or from issues in encoding and decoding. Liu et al. (2024) state that there might be a distribution shift between the source training and test data, or that the lack of human supervision may produce unintended results. Jiu et al. (2022) analyze various cases of training errors. For example, imperfect representation learning where the encoder comprehends and encodes the text into meaningful representations could influence the degree of hallucinations. Also, there are issues in decoding, where the decoder takes the input from the encoder and generates the final target sequence might either use the wrong part of the encoded input or the design of the system itself might be prone to hallucinations.

In the cases where hallucinations come from the data itself, data might be sparse for some areas, influenced by people's writings about some topics or originally incorrect, if taken from some unreliable sources. AI hallucinations comprise not only factually wrong statements or fabricated information but also *qualia*-like statements for the model itself. In a famous occurrence, the Bing chat model proclaimed love to a journalist (Roose, 2023), and there was a case of a Google engineer claiming the model was talking about its self-awareness and consciousness (De Cosmo, 2022).[13]

One must now ask themselves: is natural-language generation and "understanding" process in large language models a *reliable* process? That is, can one talk about some kind of epistemology here, even though it is carried away in a different kind of ontology? Namely, issues in hallucinations have started to get mitigated by various strategies. Some of these include the

13 When I talk about hallucinations here, I am talking about *open-domain* (knowledge of the world) fabricated answers, and not simple errors or mistakes in different contexts.

mentioned *retrieval-augmented generation* (RAG) methods:[14] for example, using the web search to additionally check the model's output or producing various changes to the training or inference phases.[15] Tonmoy et al. (2024) show that there are two main paths in hallucination mitigation: *prompt engineering* and *developing models*. The first step includes modifying the system prompt given to a model, often by retrieval augmented generation and adding additional knowledge bases, or by self-refinement through feedback and reasoning and various prompt-tuning strategies. On the other hand, the model development can be optimized by introducing new encoding/decoding strategies, supervised fine-tuning, utilization of knowledge graphs, and modifications to loss functions that quantify the difference between the actual target values and the predicted outputs of a model.

According to Hughes hallucination evaluation model,[16] GPT 4 Turbo has a factual consistency rate of 97.5%, which means its hallucination rate is only 2.5%, while other large language models can produce hallucinations in 10-20% of the cases. This is just a sample metric that does not correctly correspond to the real state, but Open AI (2024) itself has produced comparisons between GPT-3 and GPT-4, stating that the initial version of GPT-4 scored 19% higher at avoiding *open-domain* hallucinations (stating incorrect information about the world) and 29% higher at avoiding *closed-domain* hallucinations (straying away from the current text or context). Taking into account these improvements that are on the higher end of the accuracy spectrum, the natural-language generation in large language models is fit to be claimed as a *generally reliable* process of producing justification. One might object saying that there are errors in reasoning and similar mistakes outside AI hallucinations, and one would be correct. However, the same issue is at stake with human epistemological agents as well: what I want to emphasize is that *only with hallucination-mitigation techniques, LLMs can in general be considered a reliable source that can possess their own kind of epistemology*. Without such improvements, it is difficult to talk about reliable processes at all, due to sparse data or big hallucination errors.

When a person queries a model, it can provide sources for its claims, or its claims are present in

14 A set of techniques for enhancing the reliability and accuracy of AI models where facts are fetched from external sources and used to modify the model's output.
15 See Šekrst (forthcoming) for a RAG-technique using justification logic to classify sources according to their knowledge-producing qualities.
16 Available at: https://huggingface.co/vectara/hallucination_evaluation_model (Vectara, 2024).

the data itself as reliable, foundationalism-mimicking baseline cases. Suppose the overall process of generation is optimized with mentioned hallucination-mitigation strategies. In that case, we are definitely talking about a reliable process of producing knowledge, especially taking into account that such strategies are becoming inherent parts of the model itself. It is interesting to note that large language models, the same way as humans can, perform a certain type of *belief revision*[17] when they provide an incorrect answer and are faced with a contradiction from the user. For example, the GPT model will often respond with "I apologize for the oversight" and take new information into account, where it may or may not continue reasoning correctly, based on the quality of data provided (cf. Arkoudas, 2023).

One of the most common issues in reliabilism is the generality problem, where any taken belief is a product of a causal process in the mind of the agent in question, at a particular time and place, but the question is: what is the type of that token? Goldman and Beddor (2021) illustrate this by saying that each process token can be typed in a broader and narrower way, for example, Smith sees a maple tree outside his house and forms a belief that there is a maple tree outside. The question is what is the general type used here: *vision*, *visual experience of a tree*, *visual experience of a maple tree outside*, etc., and singling out any of these seems arbitrary. I will not be offering a solution to this debate since it is out of the scope of this paper, but one information-oriented extension might be a fruitful setup for the development of AI epistemic agents. Goldman and Beddor (2021) consider Lyons's (2019) approach as an example of a fully developed one, in which psychological process types are information-processing algorithms that need to be relativized to parameters. To illustrate, *lighting condition* is a parameter that affects the accuracy and speed of a person's visual apparatus processing, and the process type is something like "visual recognition of objects based on the retinal stimulus of sort *S*, in lighting conditions *C,* with attention distributed in manner *M*".

In the case of a large language model, I would like to show how a similar approach could be taken. Every call to a model consists of a user prompt and – often underlying – *parameters*. One of these parameters that the user can control is the *temperature* parameter that influences the

17 Cf. in logic of belief revision (see Hansson, 2022), the set representing the belief state is assumed to be a logically closed set of sentences that can be changed by the introdution or removal of a belief-representing sentence. If such a logic would be used as a RAG tool for the generative models, hallucinatory statements could be removed from the underlying set.

model's output. A higher temperature results in more creative outputs of a model, but is also prone to more hallucinations, while a lower temperature is more predictable, but strays away from imitating human beings in a conversation. Comparable to information-oriented epistemology, for an AI model, identifying a picture of a maple tree would be a similar process type "computer-vision recognition of objects based on parameters x_1 …. x_n" or "talking about maples, taking into account the knowledge base *B*, with parameters x_1 … x_n".

Unlike human cognition, AI operates on algorithms and data processing, devoid of subjective experiences and *qualia*-states or intentions. Therefore, a more nuanced examination is required to determine whether the principles of reliabilism can be seamlessly applied to various AI systems. However, the purpose of this section was to show that there are curious parallel similarities and similar issues arising in both human epistemic agents and possible AI epistemic agents, especially taking into account that a similar information-processing framework might be used to capture nuances of both types of epistemologies in question.

5. Computation and explainability in AI

It has been established that in order to talk about any real kind of epistemology in AI systems, hallucination-mitigation techniques need to be incorporated in order to be close to an ideal epistemic agent. However, as it is the case in the complete description of a reliable process in human beings, a vast amount of background processing in the human brain is unknown to us, comparable to the issue of *black-box* computation[18] in deep learning. Namely, deep-learning procedures can be computationally too expensive or impossible to explain, similar to the issues in human cognition, giving rise to the approach of *explainable AI* (XAI), which focuses on making AI processes more transparent and understandable. Longo et al. (2024) illustrate this with a variety of methods to make AI models more interpretable and understandable. These include *ante-hoc* (intrinsic) explainability, which involves designing inherently interpretable models like decision trees, and *post-hoc* explainability, which creates explanations for trained

18 A black-box program or algorithm is a procedure in which the user cannot (easily) see the inner workings and decisions made in the background.

models using sophisticated techniques like LIME and SHAP.[19] XAI methods can be local, focusing on individual predictions, or global, explaining overall model behavior. Model-agnostic methods, applicable to any model, and model-specific methods, tailored to particular models, are both utilized. Attribution methods such as saliency maps highlight important input features, while concept-based learning algorithms link predictions to human-understandable concepts. Attention mechanisms in neural networks and rule-based approaches further aid in transparency. Evaluating XAI methods often involves various human studies, though synthetic data and virtual participants are proposed to enhance robustness and generalizability.

By employing explainability techniques, along with hallucination-mitigation strategies, epistemological reliabilism aligns well with the goals of XAI. Similar to the issues in generality problem, different explainability method can show different types of mechanisms in the background, but attribution methods like saliency maps that visualize the model's output (Longo et al. 2024) can reveal how specific inputs influence outputs, aligning with more detailed levels of generality, close to Lyons's (2019) detailed information-parameter processing loop. However, we are still far away from explainable AI in general since explainability issues are often computationally expensive or even physically infeasible, as it is the case with NP-hard problems.

What is interesting here is that humans are often incorporated into such studies as a part of the *human-in-the-loop* process, where a human expert guides the AI development decisions, ascribing their own background beliefs and justifications. Even in various hallucination-mitigation strategies, one is again often checking other external sources of information, coming to another human corrective mechanism. It is no wonder that a misuse of similar concepts from human-oriented philosophy, epistemology and cognitive science can lead people to believe there is or will be consciousness in large language models, no matter whether this will or not actually be the case.

Through epistemology, one perceives large language models as engaging in a process of

19 LIME stands for *Local Interpretable Model-agnostic Explanations* and it is a consistent model-agnostic explainer that tries to explain the prediction of any classifier by learning the model's inner workings locally around the given prediction. SHAP stands for *Shapley Additive Explanations*, and is a method that tries to explain individual predictions as well, based on game theory and optimal Shapley values (characterized by a collection of desirable properties).

knowledge generation that parallels human cognition, but within a computational framework. Large language models, trained on vast datasets drawn from the internet, encode and decode text, transforming it into meaningful representations similar to how humans process language and new information. However, akin to the generality problem in epistemology, the specifics of this process remain elusive, with the "black-box" nature of deep learning mirroring the mysteries of human cognition.

Moreover, the emergence of AI hallucinations pinpoints the shared vulnerabilities between large language models and human cognition, highlighting the need for mitigation strategies and the ongoing pursuit of explainable AI. As AI models evolve to address these challenges, incorporating techniques such as retrieval-augmented generation and explainability methods, they increasingly resemble human cognitive processes, yet with their own distinct computational characteristics, which constitute different but analogous ontologies.

Taking into account the still-unknown territory of human cognition, expanding the comparison between human and AI hallucinations could offer a rich avenue for enhancing our understanding of AI cognition and refining mitigation strategies. By diving deeper into the similarities and differences between human hallucinations, which arise from internal cognitive processes, and AI hallucinations, which stem from algorithmic errors or biases, one gains valuable insights into the underlying mechanisms of both phenomena. Furthermore, by exploring how these comparisons illuminate the intricacies of AI cognition – such as the role of context, prior experiences, and feedback mechanisms – we can advance our understanding of AI decision-making processes and contribute to the development of more robust and transparent AI systems, satisfying the goal of explainable AI movement.

6. Discussion and final remarks

As AI systems advance, they not only reshape our understanding of intelligence but also reflect shared vulnerabilities and challenges reminiscent of human cognition. In essence, the journey through AI hallucinations and cognition not only deepens our understanding of artificial

intelligence but also prompts profound reflections on the nature of *intelligence*, *justification* and *knowledge* themselves. Even though we are still not talking about the same kind of epistemology, knowledge, beliefs and understanding, we might see parallels in their similarity since they are modeled after human cognitive concepts. Until then, one should be careful since our cognitive vocabulary is undermined with background presuppositions, leading us to ascribe mental states where there are none or ascribing none where there might be something of a different ontological status. This is also tied to the usage of the term *hallucination*, implying background mental states.

The examination of epistemological frameworks such as reliabilism shows that while AI systems can exhibit patterns of reliability, their knowledge production seems fundamentally distinct from human cognition (cf. Cantwell Smith, 2019). The processes involved in generating "truth" for LLMs are mechanical and algorithmic, yet their errors resonate deeply with human-like imperfections. The paper underscored that AI systems, like humans, are fallible in their reasoning and output. However, the application of human epistemological concepts such as justified true belief to AI systems remains fraught with conceptual challenges – challenges rooted in the profound ontological gap between human and machine cognition.

One of the central contributions of this paper is its emphasis on the necessity of interdisciplinary approaches. Addressing AI hallucinations purely through a technical lens neglects the broader philosophical implications. By incorporating insights from epistemology, cognitive science, and the philosophy of mind, this paper makes the case that understanding and mitigating AI hallucinations requires a shift from isolated technical solutions to a more integrated, epistemologically grounded approach. This shift is not merely academic; it is essential for building trust in AI systems as they are deployed in increasingly critical real-world applications.

The conclusions drawn here are critical for the future development and societal integration of AI systems. As AI takes on larger roles in areas like healthcare, law, and education, the impact of hallucinations remains significant. This paper advocates for stronger epistemological frameworks

and highlights the importance of explainable AI to ensure that AI-generated *knowledge* – or, if we are philosophically careful, *information* – is both reliable and transparent, framing hallucinations not merely as technical flaws but as deeper epistemological and cognitive challenges.

References

Aaronson, S. (2013). Why philosophers should care about computational complexity. In B. J. Copeland, C. J. Posy & O. Shagrir (Eds.), *Computability: Turing, Gödel, Church, and Beyond*, 261–328.

Arkoudas, K. (2023). GPT-4 Can't Reason. arXiv. https://arxiv.org/pdf/2308.03762

Cantwell Smith, Brian (2019). *The Promise of Artificial Intelligence*. MIT Press.

De Cosmo, L. (2022). Google Engineer Claims AI Chatbot Is Sentient: Why That Matters. *Scientific American*, July 12 2022. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters

Dziri, N.; Milton, S.; Yu, M.; Zaiane, O. & Reddy, S. (2023). On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 5271–5285. https://aclanthology.org/2022.naacl-main.387.pdf

Gardner, H. (2011). *Frames of Mind*. New York: Basic Books.

Gettier, Edmund L. (1963). Is Justified True Belief Knowledge?. *Analysis*, 23(6), 121–123. doi:10.1093/analys/23.6.121

Goldman, A. (1998). Reliabilism. In *The Routledge Encyclopedia of Philosophy*. Taylor and Francis. https://www.rep.routledge.com/articles/thematic/reliabilism/v-1 doi:10.4324/9780415249126-P044-1

Goldstein, S., & Levinstein, B. (2024). Does ChatGPT have a mind? *arXiv*. https://arxiv.org/pdf/2407.11015

Goldman, A. & Beddor, B. (2021). Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*, https://plato.stanford.edu/archives/sum2021/entries/reliabilism

Hansson, S. O. (2022). Logic of Belief Revision,. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2022 Edition)*, https://plato.stanford.edu/archives/spr2022/entries/logic-belief-revision

Hatem, R., Simmons, B. & Thornton, J. E. (2023). A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. *Cureus*, 15(9). doi:10.7759/cureus.44720

Hintze, A. (2023). ChatGPT believes it is conscious. *arXiv*. https://arxiv.org/abs/2304.12898

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/35717

Krzanowski, R. & Polak, P. (2022). The meta-ontology of AI systems with human-level intelligence. *Philosophical Problems in Science* 73, 197–230. https://zfn.edu.pl/index.php/zfn/article/view/610

Longo et al. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106. https://doi.org/10.1016/j.inffus.2024.102301

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2024). Trustworthy LLMS: A survey and guideline for evaluating large language model's alignment. *arXiv*. https://arxiv.org/pdf/2308.05374

Lyons, J. (2019). Algorithm and Parameters: Solving the Generality Problem for Reliabilism, *The Philosophical Review*, 128(4), 463–509. doi:10.1215/00318108-7697876

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. *IEEE Conference on Artificial Intelligence (CAI)*, pp. 133–138. https://ieeecai.org/2024/wp-content/pdfs/540900a127/540900a127.pdf

Lyotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge* (G. Bennington & B. Massumi, Trans.). Manchester: Manchester University Press (Original work published 1979).

Open AI (2024). GPT-4 System Card. *OpenAI.com*. Mar 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf

Piedrahita, O. A., & Carter, J. A. (2024). Can AI believe? *Philosophy & Technology, 37*(89). https://doi.org/10.1007/s13347-024-00780-6

Poston, T. (2024). Internalism and externalism in epistemology. https://iep.utm.edu/int-ext

Roose, K. (2023). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *New York Times*. Feb 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Russell, S. & Norvig, P. (2003). *Artificial intelligence: A Modern Approach*. New Jersey: Prentice Hall.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3(3), 417–457. doi:10.1017/S0140525X00005756

Šekrst, K. (2020). AI-Completeness: Using Deep Learning to Eliminate the Human Factor. In S. Skansi (Ed.), *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives* (pp. 117–130). Cham: Springer International Publishing.

Šekrst, K. (forthcoming). Unjustified untrue "beliefs": AI hallucinations and justification logics. In K. Świętorzecka, F. Grgić, & A. Brozek (Eds.), *Logic, knowledge, and tradition: Essays in honor of Srećko Kovač*.

Tonymoy et al. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv*. https://arxiv.org/abs/2401.01313

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, LIX (236), 433–460. doi:10.1093/mind/LIX.236.433

Yampolskiy, R. (2012). AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI. *23rd Midwest Artificial Intelligence and Cognitive Science Conference*, MAICS 2012, Cincinnati, Ohio, USA, 21-22 April 2012.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In : I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems 30* (NIPS 2017)

Vectara (2024). *Hallucination Leaderboard*. Retrieved on May 31, 2024. https://github.com/vectara/hallucination-leaderboard

Waters, F. & Fernyhough, C. (2016). Hallucinations: A Systematic Review of Points of Similarity and Difference Across Diagnostic Classes. *Schizophrenia Bulletin*, 43(1). doi: 10.1093/schbul/sbw132