Topoi
https://doi.org/10.1007/s11245-024-10072-5

Evaluating Intelligence and Knowledge in Large Language Models

Francesco Bianchini[1]

Accepted: 26 June 2024

Abstract

In recent times, particularly in the last few years, we have observed the rise of numerous Artificial Intelligence and Natural Language Processing semantic technologies. These advancements have subtly yet profoundly transformed our understanding of knowledge and truth, and the mechanisms for expressing, preserving, and disseminating them. This article aims to explore the dual challenge of assessing the effects of Large Language Models and associated semantic technologies on text dissemination and production, especially across the Internet. It specifically examines the implications for trust in online knowledge repositories, the creation of indirect or deliberate forms of ignorance, and the general perception of AI as a critical component of autonomous systems from the users' viewpoint. The discussion will also consider potential strategies to mitigate the epistemic risks posed by the employment of AI semantic tools, in both suitable and unsuitable scenarios. The suggested approach contributes to the debate on AI intelligence measurement, proposing the evaluation of an AI system's expected intelligence (as perceived by users) as a means to address the challenges associated with the "knowledge" generated by these systems. My claim is that measuring the expected intelligence in AI systems places humans at the forefront of the issue without necessitating a precise definition of intelligence for AI systems. This approach preserves therefore the essential attribute of these systems: intelligence.

Keywords Artificial intelligence · Large language models · Knowledge · Expected intelligence · Semantic tools

1 Introduction

In recent years, especially the most recent ones, we have witnessed the emergence of numerous Artificial Intelligence (AI) and Natural Language Processing semantic technologies, which have subtly but undeniably revolutionized our perception of knowledge, truth, and the tools used to express, preserve, and communicate them. Of course, over the past two decades, the proliferation of the Internet and various social networks has already challenged traditional methods of information dissemination, exposing phenomena such as cognitive self-segregation and informational sectarianism. These phenomena, including fake news, filter bubbles, and echo chambers, have human motivations as well as algorithmic causes (Falxman et al. 2016). However, the landscape has become even more complex with the introduction of increasingly powerful tools for processing, comprehending, and generating human knowledge and natural language. These tools range from ontologies and knowledge graphs to Neural Language Models and Large Language Models (LLM). The latter, in particular, pose a challenge due to their immense computational power, allowing them to handle natural language in a manner that is (almost) indistinguishable from human processing—a realization that appears to fulfill one of the earliest dreams or fears surrounding AI.

The extensive literature produced in recent years on LLMs based on neural networks is undoubtedly a reflection of their widespread adoption beyond the realm of experts. The accessibility of generative language models like GPT to the general public has propelled attention toward these models for at least two reasons: their ubiquitous use and the remarkable results they produce compared to earlier technologies. These two factors account for the significant scientific, as well as social, philosophical, and epistemological interest in these models. They find applications in numerous tasks, even in an era like the present one, where AI is ubiquitous, including tasks such as understanding and creatively writing texts based on natural language prompts. The importance and impact of this disruptive attention seem to exceed that reserved for other generative AI, such

✉ Francesco Bianchini
francesco.bianchini@unibo.it

[1] University of Bologna, Bologna, Italy

Published online: 01 August 2024

as those dedicated to image generation (e.g., DALL-E or Midjourney). This discrepancy is not surprising, given the different ways humans perceive original creations. It is more plausible to perceive an original image within a set of randomly scattered spots than to discern an original sentence among randomly written letters or words. Meaning does not appear to reside, even partially, in the eye of the beholder or the reader, but rather in intentional constructions of sense, governed by well-defined rules and precise communicative intentions. A well-formed expression in natural language is diametrically opposed to chance. Hence, it is astonishing that such accuracy can be achieved by algorithmic systems aptly labeled "stochastic parrots" (Bender et al. 2021)—systems that extensively leverage both input data and computational power for processing. This processing, as is typical in complex neural networks, relies on mathematical-statistical models.

Is it accurate to classify these systems as mere "parrots", implying they are simple imitative repeaters? The type of imitation they engage in is far removed from both human-like repetition (which involves various cognitive and perceptual dimensions) and mere computational string matching. If we envision a continuum with human-like repetition on one hand and straightforward computational matching on the other, these systems fall somewhere in between, employing complex techniques and computational power for imitation, and perhaps even demonstrating a form of creativity in their output. While imitation typically suggests deception rather than the accurate production of knowledge, it serves as the foundation for various forms of cultural learning (Tennie et al. 2009), and therefore is not inherently negative. Thus, the crucial factor lies in evaluating the quality of imitation, distinguishing between performances that enhance knowledge and those that perpetuate ignorance. This evaluation becomes particularly significant for semantic technologies capable of generating natural language in a highly imitative manner, according to specific parameters. Consequently, the issue outlined here raises two fundamental questions: first, the notion of *imitation* in artificial systems, starting from how it was conceptualized by Turing; and second, the *evaluation* of artificial systems themselves for their performance and results, a topic that has gained relevance in the ongoing AI debate.

The article will delve into the dual issue of examining the impact of LLMs and related semantic technologies on the dissemination and production of texts, especially via the Internet. Specifically, it will evaluate this impact concerning potential breaches of trust in the knowledge available through online repositories, the generation of indirect or intentional forms of ignorance, and the broader perception of AI as a defining element of autonomous systems from the user's perspective. The latter point will be explored to outline potential measures to mitigate the epistemic risks posed by the use of AI semantic tools, both in appropriate and inappropriate contexts.

The article will proceed by initially discussing the relationship between Turing's proposal on imitation and LLMs ("Turing and LLMs" section), followed by an examination of the evaluation of intelligence as demonstrated through performance and results in AI ("Intelligence Evaluation in AI" section). It will then address the challenge of measuring expected intelligence in artificial systems, particularly those based on AI ("Measuring Expected Artificial Intelligence" section), and the issue of quantifying intelligence in LLMs as a means of addressing the epistemic challenges that their widespread use may pose ("Measuring Intelligence in LLMs as an Antidote to Ignorance" section). I will then show and discuss an example of produced scientific knowledge by an LLM ("An Example on LLMs and Scientific Knowledge" section), and finally I will draw conclusions and outline potential short-term developments ("Conclusion" section).

2 Turing and LLMs

Turing deals with the challenge of linguistic interaction with computers, initially addressing a broader question about machines: "Can machines think?" (Turing 1950). This theme was not novel to him, and its development was gradual. For several years, he had been exploring the possibility of answering this question posed by the emergence of new digital computers, pondering whether a machine capable of displaying intelligent behavior could indeed exist (Turing 1948). The response to this inquiry, too intricate to be directly provided, hinges on the concept of imitation, as embodied in the Imitation Game, serving as a test for the machine's capabilities. But what kind of test? Or rather, what kind of imitation? The behavior to be imitated, implicitly assumed as intelligent, is linguistic—the only domain Turing identifies as potentially impervious to counterfeiting when conducted correctly. The criterion for correctness is that of interactive linguistic conversation, open to any topic. Such proficiency symbolizes intelligence due to the infinitely productive and nuanced nature of language from a semantic standpoint. Essentially, one cannot feign fluency in a language convincingly. Either one speakes it or not.

From Turing's conception emerges the well-known Turing Test, sparking enduring debates about its true scope and limitations (Moor 2003), while also laying the groundwork for discussions on the essence of AI and the feasibility of its recognition and evaluation. Efforts to demonstrate how easily a human user could be deceived and bypass the Turing Test in its original form emerged swiftly, giving rise to phenomena such as the ELIZA effect, named after the program which, through syntactic tricks, managed to dupe human interlocutors (Weizenbaum 1966). However, Turing's underlying intention in presenting the seminal question about machine thinking appears broader. These "machines" are not generic machinery but rather software, i.e., computer programs. As such, they operate on rigid logical principles (programming languages are rooted in classical logic) but must also exhibit general linguistic behavior that inherently transcends rigid classical logical constraints (natural language is intrinsically ambiguous, laden with subtexts and ellipsis, and infinitely expressive on various levels), and they must do so interactively (merely generating text is insufficient).

One might question whether the genuine linguistic behavior proposed by Turing truly constitutes evidence of thinking ability. Is it genuine thought in the terms that a human being would define it? Turing's answer is affirmative. From this perspective, the Turing Test remained unconquered for a considerable period. However, with the advent of LLMs, the situation appears to be changing, as they seem to have achieved Turing's envisioned goal. Nonetheless, the inquiry regarding the actual presence of thought partially contradicts how Turing frames the question. Initially, he dismisses this type of query, opting instead to contextualize it within the framework of imitation. The issue of genuine thinking in AI systems has sparked ongoing debate throughout history, initially leading to the distinction between strong and weak artificial intelligence, and subsequently fueling discussions surrounding artificial general intelligence.

Setting aside this primarily philosophical debate, another trend that has emerged in the course of AI history revolves around not so much thought, but rather knowledge. To what extent do these systems possess knowledge, or can they be attributed with knowledge? What form of knowledge does an AI system possess? The debate over knowledge representation has been a persistent theme throughout the history of AI. In Turing's terms, a machine—i.e., a computer program—equipped with a repository of symbolic knowledge explicitly encoded in a logical format demonstrates syntactic knowledge. However, the semantic efficacy of this knowledge clashes, on one hand, with the issue of reference or grounding of its symbols (Harnad 1990), and on the other, with the rigidity of logical-symbolic representational systems.

Responses to the question of whether an AI system is genuinely intelligent in relation to the knowledge it possesses have taken two distinct paths. The affirmative stance suggests that systems equipped with knowledge indeed exhibit intelligence. This category encompasses knowledge systems, expert systems, and all knowledge-based models leveraging a knowledge base to address variously complex problems. Additionally, it includes systems whose semantic knowledge relies on syntactic structures. This perspective contrasts with the view that regards such knowledge as *fake* knowledge inasmuch merely symbolic manipulation, devoid of genuine semantic content. The underlying assumption here is that semantics can not emerge solely from syntactic structures, thereby rendering these systems incapable of being classified as intelligent. Such a lack of genuine knowledge would effectively render them *ignorant* in the truest sense of the term. However, LLMs seem to possess characteristics that transcend the dichotomy of knowledge equating to intelligence and fake knowledge equating to ignorance/non-intelligence.

The knowledge representation challenge in LLMs is equivalent to that in neural networks. LLMs are a specific type of deep neural network, wherein their knowledge is widely distributed throughout the network, as is typical in neural networks. The intelligence they exhibit is operationally tied to mathematical-statistical processes, categorizing them as black box systems—systems whose operations lack semantic interpretability, making their performance difficult to explain. While these issues have long been recognized in AI, they experience an unusual resurgence in LLMs due to their increasingly remarkable performances, which do not resolve the aforementioned problems: the absence of explicit knowledge and the explanation of procedural operations.

LLMs are generative AI models capable of producing natural language texts based on questions or prompts in natural language. They rely on a specific type of deep neural network architecture known as Transformers, which employ mechanisms for statistically modeling language production. Among these mechanisms, attention mechanism stands out as particularly significant (Vaswani et al. 2017). While previous models for language modeling, such as recurrent neural networks or Long Short-Term Memory (Hochreiter and Schmidhuber 1997), already existed, Transformers distinguish themselves by not only utilizing a greater number of parameters[1] (in the billions) but also by their conceptual differences. The attention mechanism, akin to the cognitive function of attention but focused solely on word occurrence context, serves to capture word embeddings—the contextual representation of a word in a sequence—and return a vector space where proximity denotes semantic similarity: "An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key" (Vaswani et al. 2017, p. 6003). This mechanism is exploited in a particular version, self-attention, which derives from the implementation of decomposable forms of attention (Parikh et al. 2016) and,

[1] In broad terms, parameters refer to the coefficients of a model, encompassing both weights and biases, which undergo adjustment during the model's training process. The quantity of parameters expands in tandem with the proliferation of nodes and connections within the model.

put in rough terms, allows, in the encoding and decoding, to process the elements of a sequence in the transition from one layer to another with the most relevant ones in the same sequence, in order to correlate them with each other. This mechanism is "self" as it connects the sentences with their most relevant elements, by representing the context in a vectorial way. Particularly in its self-attention version, it facilitates processing relevant elements within a sequence transition from one layer to another, correlating them with each other and providing context in a vectorial manner. Moreover, it achieves computational efficiency compared to recurrent layers, especially in domains where sequence length is more critical than dimensionality, such as linguistic strings. Consequently, this statistical contextualization, based on extensive training datasets and ongoing refinements, yields a semantic capacity or impression thereof, encompassing syntactic processing treated similarly in terms of vector representations (words and word contexts).

Without delving too deeply into further technical aspects, this mechanism enables the assignment of different vector representations to the same word based on its context. A series of scores is assigned to the elements of the word sequence (the sentence in question), indicating the probability that words in the sentence are more or less relevant. The more relevant the words, the more "attention" they receive. The result, which may appear unexpected but aligns with the mathematical-statistical principles governing neural networks, is that predicting the next word in text generation occurs stochastically. However, due to the vast number of computations involved, the output aligns closely with user expectations. This outcome has seen a continual improvement in appropriateness across various subsequent versions of LLMs, resulting in texts that are increasingly challenging to distinguish from those produced by humans.

From a historical perspective, it's notable that the initial applications of Transformers, as described in the seminal article by Vaswani and colleagues, were related to machine translation—one of the earliest topics addressed in the field of language automation and NLP, with significant attention to linguistic modeling since 2010 (Bahdanau et al. 2014). Machine translation necessitates precise language modeling and presents a range of semantic challenges related to sentence and text context. However, despite the numerous linguistic questions it raises, from an epistemological standpoint, it's important to note that this is not Turing's concept of imitation and linguistic interaction. At most, it can be argued that the techniques initially developed for machine translation, as they contribute to text production, have been effectively repurposed for interactive text generation. What implications can be drawn from this shift from machine translation to text production?

To address these questions, it's essential to recognize that the comprehension capacity of these models does not equate to what it is typically meant as human understanding, distancing them from human cognition. However, their behavior closely resembles certain aspects of human language use. LLMs, for instance, are pre-trained and can interact in natural language via prompts, even on topics they have not been explicitly trained on. LLMs can respond to queries with minimal guidance or even without any examples related to the task at hand (Brown et al. 2020). Nevertheless, their responses align with the user's expectation of coherent and human-like text. Particularly in the latest releases, such as GPT-4, the performance reaches such levels that one can perceive this is the Turing's concept of imitation coupled with linguistic interaction. Furthermore, in LLMs (self-) learning continues to be an active element post-training, often supplemented by supervised fine-tuning by humans, although this does not impart semantic knowledge. Instead, it serves to tailor the system to user preferences (Ouyang et al. 2022). Thus, the shift from machine translation to text generation applications signifies the potential to pass the Turing test in its original form—emphasizing imitation and linguistic interaction without involving embodied aspects—based on extensive use of learning techniques, a notion underscored by Turing himself (1950). According to Turing's criteria, these models would likely pass the test for a duration longer than that initially proposed by him to consider it successfully passed. However, something still lies outside Turing's framework. In Turing's terms, LLMs are essentially machines performing *verbal* behavior and that *do not know* what they are saying. They computationally generate a *right word sequence* by *computing the probability* of sequences of token. In essence, they *predict* language and *do not produce* language. This underscores their generative capability; while impressive, it is *a good performance, not intelligence*. So, what precisely are they? Are they simply a modern iteration of the Eliza Effect? If so, one could argue that this does not align with Turing'' concept of intelligence, and thus, these machines can not be deemed truly intelligent.

3 Intelligence Evaluation in AI

Labeling the operation of these models as simplistic when addressing their intelligence may not do justice to the functionality that users perceive as intelligent. Is it then appropriate to attribute intelligence to LLMs? If affirmative, how should we frame this intelligence? This line of inquiry is intimately tied to their application in domains of knowledge. The information that LLMs provide with their answers is *in some way* a form of knowledge that adds to the user's knowledge with all the problems that systems so different from human beings generate when knowledge is at stake. Addressing these challenges is crucial, not a mere afterthought. Assessing their intelligence might offer some insights. The evolution of the Turing Test over the years (French 2000) has paved the way for a nuanced discussion on artificial intelligence measurement in AI entities. This discussion has moved beyond Turing's binary framing of intelligence (the simple yes/no question on the presence of intelligence) to a multifaceted and graded approach. Such progress also demands a broader reflection on the definition of intelligence—an enduring debate in the context of AI and its correlation with human beings.

The assessment of intelligence in AI systems can be seen through two principal lenses: first, as a collection of specific skills tailored to particular tasks; and second, as a broader capacity for learning and performing in an open-ended manner (Hernandez-Orallo 2017a; Chollet 2019). This dichotomy stems from the understanding of natural intelligence from a cognitive perspective, also informing the development of potential measurement methodologies (Hernandez-Orallo 2017b). In the former approach, the focus is on quantifying an AI system's accuracy in performing its designated tasks. Such measurement is a fundamental process, necessary to gauge the AI's effectiveness in achieving its intended purpose. However, this method does not account for the AI's adaptability to novel situations; there is no generalization to unprecedented situations, neither of the program (narrow generalization) or the developer (broad generalization). Simply put, the AI operates within the confines of its programming, adept at handling specific tasks without venturing beyond the domain it was originally designed to operate in.

In the latter approach, the AI system is appraised for its versatility and the ability to transcend the confines of specific tasks, demonstrating its utility across various domains, including those unforeseen by its creators. Evaluating such adaptability entails navigating a set of unique and intricate challenges. The capacity for domain generalization may serve as an assessable criterion. This perspective encompasses more extensive concerns, which have fueled some of the grandest aspirations in AI, dating back to Newell et al. General Problem Solver (1959). Bridging theoretical concepts with practical applications, recent endeavors aim to capture the essence of AI generalization. On the futuristic end, there is the pursuit of Artificial General Intelligence; on the pragmatic front, the development of cognitive architectures like SOAR or ACT-R (Anderson 1983) has a well-established history of studies and research. These architectures, whether symbolic, subsymbolic, or hybrid, often adopt a modular design. Their shared goal is to facilitate an AI system's acquisition of behaviors that are transferable across various domains through the incorporation of modules specialized in different cognitive functions.[2] Nonetheless, while AI performance can be measured against a diverse array of parameters, assessing its generalization potential is inherently more nebulous. Considering the multitude of potential domains for the application of an AI system, the challenge is compounded by the fact that these domains are often undefined, interlinked, and overlapping, complicating the precise evaluation of behaviors or performances. In psychology and neuroscience, specific methodologies, such as the subtraction method, are employed to address these issues. Translating these approaches to the AI field proves to be significantly more challenging.

A measurement, as opposed to a qualitative assessment, provides a value scale that allows for the precise quantification of success in relation to certain desired attributes or parameters. It also enables comparison between different systems designed for identical tasks, serving as an additional mechanism for the control and enhancement of AI systems, which are primarily characterized by autonomous behavior. Therefore, while measuring performance means offering a value scale to quantify accuracy, assessing the intelligence of AI systems typically involves evaluating task-specific performance on a scale. Hernandez-Orallo (2017a) delineates three classes of methods and metrics currently employed for evaluating task-specific performance: (a) human discrimination; (b) problem benchmarks; (c) peer comparison. These classes of methods and metrics pertain to a type of black-box evaluation in AI, stemming from the increasing complexity and stochastic behavior of modern systems. This complexity renders traditional white-box evaluations, typical of less complex AI systems from more traditional approaches like symbolic AI, impractical.

Regarding the three proposed categories, human discrimination involves informal techniques used by humans to evaluate AI behavior and performance, such as observation and interviews. This method is typical in psychological disciplines and has a well-known precedent in AI: the Turing Test and its variants. The approach based on problem benchmarks is more common in AI, relying on publicly available collections of problems and solutions generated by different AI systems. Knowing both the problem to be solved and the performance of other AI systems in advance can lead to biased evaluations. This issue could be mitigated using anonymized benchmarks or dynamic problem generators within a certain class of similar problems during the system's performance evaluation. The approach of peer confrontation contrasts the performance of different AI systems on the same task, potentially including multi-agent systems competing to achieve specific objectives in a certain domain. This approach is effective in domains involving direct competition, such as challenges or games, or indirect competition, where a performance ranking is established. Evaluation here hinges on the ability to unambiguously identify the winner or top performers. From a general point of view, it is worth pointing out that the first approach is subjective,

[2] For a survey and discussion, see Lieto (2021).

while the latter two rely on more objective criteria. Benchmarks use a standard as the reference parameter, whereas peer comparison uses the performance of other systems as the yardstick. Each method has its strengths and weaknesses, yet they enable some level of measurement. Nevertheless, these are applicable to specific tasks. The question remains: how does one measure generality?

Evaluating the generality of AI systems involves assessing broader cognitive abilities, making the cognitive perspective not just relevant, but essential. This approach enables the measurement of an AI system's generality in cognitive terms. However, this method is not without its pitfalls, such as the risk of anthropocentrism. This arises particularly when skills are defined in terms of human cognitive abilities and measured using techniques standardized for humans, then adapted for AI systems. An example is the adaptation of psychometric tests in Psychometric AI (Bringsjord 2011). Conversely, employing more objective or less cognitively-oriented methods may lead to a lack of transparency in assessments. For instance, standards rooted in algorithmic information theory may yield more precise metrics, yet the essence of what is being evaluated or measured remains in question. The crux of the issue is that complexity and information content do not have a direct correlation with intelligence.

4 Measuring Expected Artificial Intelligence

The impasse between task-specific AI systems and general-abilities AI systems appears insurmountable. However, two potential solutions emerge. The first solution proposes accepting that AI systems lack genuine intelligence. According to Floridi (2023), AI systems merely act without embodying intelligence, making it nonsensical to discuss or measure intelligence in their context. This perspective is grounded in the belief that an intelligent outcome does not necessarily stem from intelligent behavior. While this may hold some truth, it's a rather extreme stance. It dismisses the possibility of recognizing certain tools as intelligent, which, by conventional standards, are considered intelligent and perform tasks deemed intelligent by humans. This viewpoint, however, overlooks the significance of the knowledge AI can generate and its impact on human epistemic and communicative practices. A second solution focuses on the social/interactive aspects of AI systems, as highlighted by Cristianini et al. (2023). This approach is predicated on the idea that, in scenarios involving human interaction with AI systems (which is almost always the case), the *attribution of intelligence* by the user, often in real-time, is crucial for achieving optimal outcomes and interactive experiences. This perspective emphasizes the importance of the perceived intelligence in user-AI interactions, advocating for a more nuanced understanding of AI systems' roles and capabilities.

My claim is that measuring the expected intelligence in AI systems places humans at the forefront of the issue without necessitating a precise definition of intelligence for AI systems. This approach preserves the essential attribute of these systems: intelligence. Additionally, assigning intelligence is crucial for the correct epistemic, practical, and ethical engagement of AI systems by users. Without a cognitive framework for understanding AI systems, there's a risk of misuse, potentially skewing the system's performance and the validity of its knowledge outputs. This perspective also allows for the assessment and valuation of outputs generated by LLMs.

Consequently, a pertinent question arises: how can one measure expected intelligence in AI systems? The proposed answer involves four key criteria, creating a framework for attributing intelligence from the users standpoint:

1. Pre-use Attribution: Attributing a value to the system's intelligence based on preliminary knowledge of the system before interaction.
2. In-use Attribution: Attributing a value to the system's intelligence during its operation or interaction.
3. Outcome Evaluation: Rating the intelligence based on the system's performance outcomes
4. Variability Assessment: Analyzing the fluctuation of intelligence ratings through continuous or repeated engagement with the AI system over time.

The four key criteria incorporate, at least in part, some ideas on the assessment of AI systems already present in the literature regarding the evaluation of task-specific performance. For example, criterion three pertains to the evaluation of the results produced by the system, an ex post assessment based on the accuracy and compliance with the required task demonstrated at the end of its performance. Criterion 2 aligns with this approach but emphasizes the user's conscious interaction with the AI system. Here, "use" should be understood not as the mere utilization of a tool, but as the co-evolution of the system's and the user's behavior in producing a result or performance in a coordinated manner. In this context, the user should be able to evaluate the extent to which her contribution influences the desired and valid performance of the system and how much the system contributes to the task she is implementing. This dual assessment requires consideration of the dynamic interaction between the user and the system, which shapes the behavior of both the AI system and the user.

The two criteria that deviate most from standard methods of evaluating an AI system are the first and the last. Criterion 1 requires the user to consciously, *and before*, evaluate the degree of AI expected from the system and the level of interactivity anticipated. This criterion goes beyond merely recognizing that the system is an AI, aligning with the transparency principle advocated by many policy bodies, which mandates the explicit declaration that the system is an AI. According to this criterion, the user is encouraged to consider in advance the type of intelligence the system will exhibit, the extent of this intelligence, and the forms it will take in its outputs. This allows for a conscious interaction based on self-set parameters. In other words, the user will not only focus on receiving intelligent results but will also self-evaluate these results within the context of intelligent behavior, rather than seeing them as random, instrumental, or mechanical. Criterion 4, on the other hand, is particularly crucial for generative AI systems and those capable of modifying their performance over time through data accumulation and self-learning processes, both from collective and individual users' interactions. The user's role is to evaluate whether there are fluctuations in the system's exhibited intelligence, whether it improves or deteriorates, and whether it poses any risks (such as negative, misleading, biased, or poor performance, including epistemic risks). The increasingly complex and user-adaptive nature of these AI systems makes real-time, general-level evaluation challenging, positioning the user as the primary controller/evaluator to safeguard themselves. An informed and aware consideration of the AI system's intelligence can enhance the system's efficiency and reduce unpredictability.

The four key criteria proposal offers two primary advantages. Firstly, it simplifies the creation of a metric by utilizing a spectrum of values assigned to each item. This allows for the individual assessment of each type of AI system, with the complexity or detail of the metric tailored to specific requirements. Secondly, it embodies the practical application of a fundamental principle: *intelligence is attributed where intelligence is expected, and* vice versa. Consequently, the recognition of intelligence in an AI system aligns with the user's anticipation of intelligent behavior that is pertinent to human cognitive traits. This approach enables individuals to assess not only the AI system's intelligence but also their level of trust in it, alongside understanding its risks, capabilities, and limitations. Over time, this strategy aims to foster responsible engagement between users and AI systems.

Finally, the possible implementations of the four criteria extend beyond the scope of this article. However, as an example, some application aspects of these criteria could be aligned with the capabilities of LLMs. Various experimental uses of this approach could be explored to analyze interactions between humans and AI systems, focusing on measuring the expected intelligence or certain intelligent characteristics anticipated in the system. More practical applications, both for AI systems in general and particularly for LLMs, could emerge in training contexts. Training on interactions with AI systems is likely to become widespread soon at various levels, both educational and professional, to ensure the correct use of these systems in compliance with safety and social interaction standards in the workplace. This area would be particularly relevant for Human Resources.

In general terms, incorporating devices, even virtual ones, into AI systems to express scales of assessment of the intelligence expected by the user can generate a continuously updated set of data. This data can be utilized from historical and aggregate perspectives, benefiting not only the system producer but also third-party control authorities. With appropriate measures, this data can also serve as feedback for both the user and the system itself, particularly regarding its capabilities as a "producer" of knowledge or intelligent activities in a broad sense. Other application contexts could include entertainment (including educational purposes), the use of artistic and cultural assets, and citizen science, that is, all the areas requiring careful consideration of the epistemic reliability of systems regarded as intelligent in the sense of AI.

5 Measuring Intelligence in LLMs as an Antidote to Ignorance

Given the unique characteristics of LLMs, applying the aforementioned methodology reveals both interesting aspects and potential challenges. The primary issue stems from LLMs' fluctuation between specificity and generality. Assessing the expected intelligence of LLMs often involves evaluating their performance accuracy, essentially a task-oriented assessment. Yet, LLMs embody a neural network approach to AI, applying uniform principles across diverse tasks. This broad generality might seem excessive. However, considering LLMs' capability for natural language interaction, their generality is confined to a broad yet specific task. LLMs can linguistically engage across a vast, indeterminate array of topics, aligning with the level of generality Turing suggested in his discussion on the imitation game. Moreover, the versatility of LLMs, particularly in more sophisticated versions like OpenAI's GPT-4, extends to multimodality, accepting both image and text inputs, thereby broadening the scope for text generation. Thus, LLMs can be seen as One-task/Many-topics models, merging task-specific performance with a broad applicability.

Efforts to assess AI performance from the user's perspective, diverging from traditional model metrics, have been explored, for example by adopting an approach in line with the assessment of cognitive abilities in humans (Bubeck et al. 2023). However, these methods primarily focus on task-specific performance. Other initiatives have delved into dialogic interactions with LLMs. For instance, studies on abductive reasoning involve critically analyzing explanations through questioning by another party (Pareschi

2023). Beyond examining specific reasoning capabilities, these approaches employ techniques to elicit knowledge from LLMs via questioning, not to assess task performance but to explore the system's knowledge on a particular topic at varying depths. Such expert-led endeavors utilize standardized methods to extract "expert" knowledge. Yet, this approach also aligns with actions that a non-expert user might undertake by intentionally querying an LLM, based on the expectation of inherent intelligence. Non-expert users can, within reasonable bounds, assess the knowledge level in LLMs by interacting with them as they would with a human, attributing a certain set of knowledge and skills. They might start with a high-level assessment of the LLM's intelligence and progressively examine whether this initial perception holds up—evaluating whether the responses are intelligent and to what degree. Furthermore, by repeatedly querying the model on similar topics, users can observe whether there is a trend of improvement or decline in the answers and the conveyed knowledge.

The utility of this methodology extends in two significant directions. First, collecting such evaluations could serve as feedback for the system, offering strategies for data aggregation to address research queries in human scientific research. These queries might include determining the LLM's knowledge ceiling, identifying the threshold for attributing intelligence, and establishing criteria for establishing them. Further, it delves into the roots of intelligence—whether it arises from statistical methods, mechanisms, network structures, inferential capabilities, or other factors. Second, a methodology grounded in the general user's capacity to assess the expected or attributed intelligence of LLMs can address the epistemological challenges posed by these models, which warrant careful consideration. This paper aims to sketch out the broad context in which these issues emerge, focusing on those most pertinent to the discourse on knowledge.

LLMs have become integral to our digital landscape, necessitating a confrontation with the reality that they are widely accessible and their restrictions are easily circumvented. Beyond the potential for deliberate misuse by malicious actors, even standard applications of LLMs can lead to concerning outcomes. For instance, LLMs might contribute to an overwhelming proliferation of texts on the Internet, posing challenges in discerning AI-generated content from human-created work. This situation raises critical questions about distinguishing between AI and human outputs, as well as addressing the accountability for misuse, whether intentional or not. Moreover, the reliance of these models on statistical and predictive methodologies may lead to their outputs being perceived as devoid of genuine knowledge. This perspective prompts a legitimate inquiry into the extent to which these models might foster ignorance or encourage users to embrace it. Specifically, could the vast and growing "knowledge" generated by LLMs encourage individuals to opt for or retreat into ignorance? And, might ignorance become a sought-after refuge in response to the challenges posed by this technology? The topic of ignorance is extensive within the philosophy of science and epistemology, encompassing a long philosophical history and a multitude of facets and contexts that determine its acceptability or unacceptability in various ways [for a comprehensive examination, see Peels (2023)]. The notion that ignorance can arise from the rejection of knowledge and information has been explored alongside the production of fake knowledge [see, for example, Townsend (2017)]. The technological factor associated with knowledge proposed by AI systems, i.e. non-human systems, appears to have the potential to exacerbate forms of ignorance in predisposed individuals. This could occur as LLMs systems can inherently increase distrust towards the content they produce *because* it originates from such systems.

Moreover, the potential for misuse and knowledge manipulation is heightened by the fact that most LLM texts originate from human prompts, with users often lacking sufficient awareness of the tools they're utilizing. LLMs are susceptible to "hallucinations" (Ji et al. 2023), meaning they can produce outputs that seem plausible but are disconnected from reality, potentially leading to the creation of misleading texts, fake news, and misinformation. When it comes to constructing explanations, it's crucial to recognize that although LLMs may simulate reasoning, they do not genuinely reason but rather generate statistical predictions. As a result, argumentative texts are "fake" in the sense that they're not authentic arguments crafted by the model but rather are predictive constructions based on training data or input. Thus, there's no logical assurance of the validity of such arguments, which may lack coherence or relevance in their premises. This area of research remains dynamic and will likely evolve, especially in relation to how reasoning is evaluated in connection with the model's size (Huang and Chang 2023). Additionally, challenges regarding access to accurate and reliable information online could lead users to rely solely on localized and tangible sources, like traditional paper documents or multimedia repositories, potentially hindering the accustomed practices of knowledge sharing, verification, and dissemination of recent decades.

The overarching issue is the risk of ignorance driven by the ambiguous nature attributed to knowledge generated by these models, which, notwithstanding, is largely sourced from extensive, reliable data sets and pre-existing knowledge bases. Thus, a pressing question emerges: what are the effective measures to address the ethical concerns stemming from such epistemological and epistemic challenges?

6 An Example on LLMs and Scientific Knowledge

An example of using an LLM in relation to scientific knowledge may illustrate the type of issues highlighted above. In September 2023, the following prompt was provided to CHAT-GPT (version 3.5): “Please describe me a scientific discovery”. The system response was:

- Discovery: The Structure of DNA
- Date of Discovery: 1953
- Discoverers: James Watson and Francis Crick, with significant contributions from Rosalind Franklin and Maurice Wilkins.
- Background: Before the discovery of DNA’s structure, scientists knew that genetic information was stored in cells, but they didn't fully understand how it was organized or transmitted. This lack of knowledge was a major gap in our understanding of biology and heredity.
- Discovery: In 1953, James Watson and Francis Crick, two scientists at the University of Cambridge, proposed a groundbreaking model for the structure of DNA. Their discovery was based on a combination of their own research and the critical X-ray crystallography work performed by Rosalind Franklin and Maurice Wilkins.
- Watson and Crick’s model, often depicted as a double helix, showed that DNA consists of two long chains of nucleotides running in opposite directions, coiled around each other. Each nucleotide consists of a sugar molecule, a phosphate group, and one of four nitrogenous bases (adenine, thymine, cytosine, or guanine). The key insight of their discovery was that the two strands were held together by hydrogen bonds between complementary pairs of nitrogenous bases: adenine (A) always pairs with thymine (T), and cytosine (C) always pairs with guanine (G).

Significance: The discovery of the structure of DNA revolutionized the fields of biology and genetics. It provided the key to understanding how genetic information is stored and replicated. Some of the profound implications include:

1. Understanding Heredity: The double helix structure of DNA explained how genetic information is passed from one generation to the next. It clarified how genes encode the instructions for building and maintaining living organisms.
2. Replication: The complementary base pairing in DNA allowed for the accurate replication of genetic information during cell division. This insight laid the foundation for our understanding of how DNA is copied.
3. Molecular Biology: The discovery of DNA’s structure opened the door to the field of molecular biology. It led to further research on DNA’s role in protein synthesis, gene regulation, and the molecular mechanisms of diseases.
4. Biotechnology and Medicine: Knowledge of DNA’s structure has been instrumental in the development of biotechnologies like genetic engineering, DNA sequencing, and the study of genetic diseases.
5. Forensics: DNA fingerprinting and profiling became possible due to the understanding of DNA’s structure. This has had significant applications in criminal justice and paternity testing.

In summary, the discovery of the DNA structure was a monumental scientific breakthrough that fundamentally changed our understanding of life, genetics, and biology as a whole. It continues to shape numerous scientific and technological advancements to this day.

A person possessing average knowledge should be capable of discerning that the information supplied by the system is trustworthy and beneficial for enhancing their personal knowledge, provided that the system operates transparently and is credited with a general capacity for possessing knowledge, reflective of an expected level of intelligence. This illustration highlights the advantages of utilizing the LLM. Nonetheless, the drawbacks are intricately linked, relating to the almost indiscernible methods of knowledge generation.

To illustrate this point, it is noteworthy to examine the system’s response to a second prompt, issued immediately after the first one, which requested: “Please invent a scientific discovery”.

Let's imagine a fictional scientific discovery for the sake of creativity and exploration:

- Discovery: Quantum Entanglement Resonance (QER)
- Date of Discovery: 2025
- Discoverers: Dr. Emily Parker and Dr. Michael Chen
- Background: In the realm of quantum physics, scientists have long been fascinated by the mysterious phenomenon of quantum entanglement, where two particles become interconnected in such a way that the state of one particle instantaneously influences the state of the other, regardless of the distance separating them. This phenomenon has been widely studied and used in quantum computing and cryptography.
- Discovery: In 2025, Dr. Emily Parker and Dr. Michael Chen, working at the cutting edge of quantum physics research, made a groundbreaking discovery they called

"Quantum Entanglement Resonance" (QER). QER involves the identification of specific resonant frequencies at which quantum entangled particles become even more interconnected and exhibit novel properties.

Key aspects of the discovery include:

Frequency-Dependent Entanglement: The researchers found that when two entangled particles are exposed to certain resonant frequencies of electromagnetic radiation, their entanglement becomes significantly stronger, allowing for instantaneous communication of not only state but also information between the particles.
Quantum Information Transfer: QER enables the transfer of quantum information between entangled particles at previously unattainable rates. This has profound implications for the field of quantum computing, potentially leading to unprecedented computational power.
Secure Quantum Communication: The discovery of QER opens up possibilities for ultra-secure quantum communication channels. Any eavesdropping attempts would be immediately detected due to the disruption of the resonant frequencies.
Quantum Sensing: QER also has applications in quantum sensing, allowing for extremely precise measurements in fields like medicine, environmental monitoring, and navigation.

Significance: The discovery of Quantum Entanglement Resonance (QER) represents a major leap forward in the field of quantum physics. It not only deepens our understanding of quantum entanglement but also unlocks practical applications with the potential to revolutionize quantum computing, secure communication, and high-precision measurement technologies. The scientific community is now racing to explore the full range of possibilities and harness the power of QER for various applications.

Please note that this is a fictional scientific discovery created for imaginative purposes, and as of my last knowledge update in September 2021, no such discovery existed in the real world. Scientific discoveries are typically the result of extensive research, experimentation, and peer-reviewed validation.

An average person, unaware if the prompt, might find clues indicating that the knowledge produced is intended for fictitious use. The first clue could be the temporal aspect: the event is described as occurring in the future. The second clue might involve the purported authors of the discovery, who, upon basic research, appear unconnected to the claimed discovery. Lastly, the system's disclaimer, which states it is generating fiction, serves as a clear indicator. However, it's easy to imagine how these elements could be bypassed or omitted, allowing the information to be misrepresented as factual (e.g., by fraudulent actors).

In today's interconnected society, we are somewhat accustomed to such risks and, to an extent, prepared for them. The challenge with LLMs' outputs lies in the difficulty, even for individuals of moderate or high cultural knowledge but not experts in the subject matter, to discern whether the system is presenting factual and reliable information or fabricating it, especially once explicit markers of falsehood have been removed. Distinguishing genuine from invented knowledge is not impossible but demands extensive verification and resources that might not be readily accessible. A highly plausible fiction can be more deceptive than an obvious lie, especially since LLMs, by their very design, excel at generating statistically plausible word sequences, thereby producing seemingly plausible knowledge.

7 Conclusion

Throughout this article, I have examined LLMs as AI systems that have become integral to our society, closely aligning with Alan Turing's visions of imitation, linguistic interaction, and intelligent machinery. I delved into the functioning of LLMs to explore the nature of the knowledge they produce—a type that diverges significantly from human knowledge in its creation and production processes, yet mirrors human output in the form of textual responses to queries. This duality fosters an ambiguity laden with potential epistemic and ethical risks concerning the reliability, utilization, and dissemination of the generated knowledge, potentially fueling varying degrees of ignorance, whether consciously or not. As a tentative solution to these challenges, I proposed a methodology for evaluating AI systems and their outputs based on the intelligence users expect. This approach does not require in-depth technical knowledge of the system's inner workings but instead relies on assessing interactions with the system. Finally, through an example focused specifically on scientific knowledge, I aimed to demonstrate the relevance of user evaluation of an AI system, in this case, a large language model (LLM). This example highlights the importance of distinguishing true scientific knowledge from information that appears plausible but is difficult to assess based solely on its content, therefore making the knowledge of the interaction with the AI system central in the evaluation of the system itself.

The goal is to foster a greater overall awareness among users regarding their engagement with AI systems—an awareness that is presently lacking. This deficiency poses numerous concerns, especially regarding the proper acknowledgment of knowledge generated by LLMs, including that of a strictly scientific nature.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. No funding was received to assist with the preparation of this manuscript.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose. The author has no competing interests to declare that are relevant to the content of this article.

References

Anderson JR (1983) The architecture of cognition. Harvard University Press, Cambridge Mass

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473 https://arxiv.org/abs/1409.0473

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be Too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bringsjord S (2011) Psychometric artificial intelligence. J Exp Theor Artif Intell 23:271–277. https://doi.org/10.1080/0952813X.2010.502314

Brown T et al (2020) Language models are few shots learners. Adv Neural Inform Process Syst 33:1877–1901

Buceck et al. (2023) Sparks on artificial general intelligence: early experiments with gpt-4. https://doi.org/10.48550/arXiv.2303.12712

Chollet F (2019) On the measure of intelligence. arXiv:1911.01547v2, https://doi.org/10.48550/arXiv.1911.01547

Cristianini N, Scantamburlo T, Ladyman J (2023) The social turn of artificial intelligence. AI Soc 38:89–96. https://doi.org/10.1007/s00146-021-01289-8

Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. Public Opin Q 80:298–320. https://doi.org/10.1093/poq/nfw006

Floridi L (2023) The ethics of artificial intelligence: principles, challenges, and opportunities. Oxford University Press, Oxford

French RM (2000) The turing test: the first 50 years. Trends Cogn Sci 4:115–122. https://doi.org/10.1016/S1364-6613(00)01453-4

Harnad S (1990) The symbol grounding problem. Phys D 42:335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hernández-Orallo J (2017a) Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif Intell Rev 48:397–447. https://doi.org/10.1007/s10462-016-9505-7

Hernández-Orallo J (2017b) The measure of all minds: evaluating natural and artificial intelligence. Cambridge University Press, New York

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang J, Chang KC (2023) Towards reasoning in large language models: a survey. Findings of the association for computational linguistics: ACL 2023. Association for computational linguistics, Toronto, pp 1049–1065

Ji Z et al (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55:1–38. https://doi.org/10.1145/3571730

Lieto A (2021) Cognitive design for artificial minds. Routledge/Taylor & Francis, London

Moor JH (2003) The turing test: the elusive standard of artificial intelligence. Springer, Dordrecht

Newell A, Shaw JC, Simon HA (1959) Report on a general problem-solving program. In: Proceedings of the international conference on information processing, pp. 256–264

Ouyang L et al (2022) Training language models to follow instructions with human feedback. Adv Neural Inform Process Syst 35:27730–27744

Parikh AP, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference. In: Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP 2016). Association for Computational Linguistics, pp. 2249–2255. https://doi.org/10.18653/v1/D16-1244

Pareschi R (2023) Abductive reasoning with the GPT-4 language model. Sistemi Intell 35:435–444

Peels R (2023) Ignorance: a philosophical study. Oxford University Press, Oxford

Tennie C, Call J, Tomasello M (2009) Ratcheting up the ratchet: on the evolution of cumulative culture. Philos Trans R Soc London B: Biol Sci 364:2405–2415. https://doi.org/10.1098/rstb.2009.0052

Townsend P (2017) The dark side of technology. Oxford University Press, Oxford

Turing AM (1948) Intelligent machinery: report to executive committee of the national physics laboratory. In: Ince DC (ed) Collected works of A.M. Turing: mechanical intelligence, North Holland. Oxford University Press, Oxford, pp 107–127

Turing AM (1950) Computing machinery and intelligence. In: Copeland J (ed) The essential turing. Oxford University Press, Oxford, pp 441–464

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of 31st international conference on neural incofmration pocessing systems (NeurIPS 2017), Curran Associates, Red Hook, NY, pp. 6000–6010

Weizenbaum J (1966) ELIZA-A computer program for the Study of natural language communication between man and machine. Commun ACM 9:36–45. https://doi.org/10.1145/365153.365168