ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

edited by

Claudio Ternullo, Matteo Antonelli

Isonomia *Epistemologica*

Isonomia – Epistemologica

Volume XII

ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

Volume 1
Il realismo scientifico di Evandro Agazzi
Mario Alai (a cura di)

Volume 2
Complessità e riduzionismo
Vincenzo Fano, Enrico Giannetto, Giulia Giannini, Pierluigi Graziani (a cura di)

Volume 3
Oltre la fisica normale
Isabella Tassani (a cura di)

Volume 4
Mettere a fuoco il mondo
Elena Casetta, Valeria Giardino (a cura di)

Volume 5
Metaphor and Argumentation
Francesca Ervas, Massimo Sangoi (a cura di)

Volume 6
Forecasting the Future
Stefano Bordoni, Sara Matera (a cura di)

Volume 7
Teaching and Learning Mathematics
Laura Branchetti (a cura di)

Volume 8
Animali razionali
Pierluigi Graziani, Giorgio Grimaldi, Massimo Sangoi (a cura di)

Volume 9
Reasoning, Metaphor and Science
Flavia Marcacci, Maria Grazia Rossi (a cura di)

Volume 10
Rational animals
Giorgio Grimaldi, Marialuisa Parise (a cura di)

Volume 11
MQ 90. Dualismo, entanglement, olismo. Un dibattito ancora aperto
Isabella Tassani (a cura di)

Volume 12
Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)
Claudio Ternullo, Matteo Antonelli (a cura di)

ISONOMIA - Epistemologica Series Editor
Gino Tarozzi gino.tarozzi@uniurb.it

ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

Edited by

Claudio Ternullo
Matteo Antonelli

ISSN 2037-4348

Direttore scientifico: Gino Tarozzi
Direttore editoriale: Pierluigi Graziani
Università degli Studi di Urbino Carlo Bo
Dipartimento di Scienze Pure e Applicate

Redazione: Via Veterani, 36 – 61029 Urbino (PU)
http://isonomia.uniurb.it/

Design by info@giticom.it

In copertina: *Riflessioni* by Miriam Borgioli, 2025.

Riflessioni è un'illustrazione che tenta di raccogliere in sé i principali temi di questo volume: Intelligenza artificiale, filosofia della scienza e della matematica. Per distaccarmi dall'immaginario ormai estremamente diffuso dell'IA e delle scienze odierne rappresentate con robot o visioni sci-fi o cyberpunk ho deciso di rivolgermi all'iconologia passata. Le allegorie sono figure molto presenti in tutta la storia dell'arte, soprattutto nel rinascimento, Cesare Ripa ne fa una raccolta in un trattato del 1593 *Iconologia overo Descrittione Dell'imagini Universali cavate dall'Antichità et da altri luoghi,* per l'illustrazione di copertina sono partita proprio da qui, dalla descrizione che il trattato dà della Filosofia e della Scienza. Come molte discipline entrambe sono raffigurate come donne di bell'aspetto, la prima tra le varie caratteristiche spiccano una posa pensosa e l'essere rivestita di stracci che mostrano ampie porzioni di pelle poiché: "Povera e nuda vai Filosofia" come diceva Petrarca; la seconda viene descritta invece con diverse caratteristiche tra cui le ali sulla testa e lo specchio. Le ali rappresentano il fine intelletto, attributo che ritroviamo anche nella figura della Logica, lo specchio invece: "Lo specchio dimostra quel, che dicono i Filosofi, che Scientia sit abstrahendo, perche il senso nel capire gli accidenti, porge all'intelletto la cognitione delle sostanze ideali, come vedendosi nello specchio la forma accidentale delle cose esistenti si considera la loro essenza" (Cesare Ripa, *Iconologia*; Tea libri, 2020, p. 398).
Il secondo grande riferimento presente è un omaggio al grande incisore M.C. Escher, artista apprezzato prima da scienziati e matematici che dalla critica d'arte per la sua personale ricerca che affronta temi come infinito, strutture matematiche e prospettiva. In particolare mi sono ispirata a una delle opere più surrealiste della sua produzione, ovvero *Buccia* una xilografia su legno di testa policroma del 1955, dove l'esterno del corpo è come una buccia che nasconde la parte più nobile dell'essere umano. In questa rappresentazione ho quindi giocato con i vari elementi di iconologia cercando di creare una moderna allegoria, una donna pensosa di fine intelletto che riflette sulle questioni proposte, essa è composta da un nastro per la macchina di Turing, una stringa di 1 e 0 che si rivela allo specchio, un codice binario da cui nasce se non tutta buona parte delle nostre tecnologie.

Table of contents

Evaluating and measuring intelligence in Neural Language Models: a methodological approach

Francesco Bianchini
University of Bologna
francesco.bianchini@unibo.it

1. Introduction

In recent years, artificial intelligence (AI) systems have evolved at an increasingly rapid pace, encompassing multiple levels and perspectives. Alongside the natural technical advancements characteristic of this field, there has been significant progress in how AI systems and tools are accepted and integrated by users. As technology has advanced, users have developed a broader awareness of these tools. However, this awareness remains superficial and incomplete for many. For some users, this awareness manifests as a recognition of the existence of new AI tools. Others have come to appreciate their potential by using them for tasks of personal or professional interest. A smaller subset – typically more experienced users – has grappled with the actual limitations of these systems. This gradual acquisition of awareness, likely more widespread than at any previous stage in AI's history, has contributed to the broad diffusion of these systems[1]. These tools can be broadly categorized into two distinct yet partially overlapping groups: voice assistants and generative AI.

Voice assistants have emerged as tools rooted in decades of research in natural language processing[2]. They have significantly extended the horizon of interaction between humans and AI systems. Generative AIs, on the other hand, were developed with goals distinct from simply disseminating AI tools.

[1] Kelly *et al* (2023).
[2] Pieraccini (2012).

Francesco Bianchini, "Evaluating and measuring intelligence in Neural Language Models: a methodological approach", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 75-100

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

Voice assistants act as interfaces connecting the networked world with productive, commercial, and informational domains. They have also broadened access to these resources for many individuals, including those previously unable to read or write. By simplifying access to internet content through natural language, voice assistants enable easier interaction with the vast array of online resources. They achieve this by providing a user-friendly interface to access the network's content while collecting information from nearly anyone who interacts with them. Their primary goal is not to provide responses in an entirely human-like manner, but to accurately interpret *users' intentions* and give precise answers. In contrast, the focus of generative AIs lies in producing outputs – whether text, images, or multimedia content – that are not only relevant but also convincingly human-like. They are built to return a result that is as appropriate as possible from the point of view of interaction to be *cognitively* understood as human *by the user*. The purpose of a conversation with a generative language system is to create interactions that are indistinguishable from those with a human, both in originality and style. While both voice assistants and generative AIs rely heavily on language, their objectives diverge. Voice assistants openly function as AI tools, with their artificial nature visible to the user. Generative AIs, however, strive for an interaction so seamless that their artificiality fades entirely from the user's perception.

Therefore, compared to voice assistants, generative AI models dedicated to language present distinct characteristics that warrant closer examination, particularly in terms of the intelligence they exhibit. This article will explore Neural Large Language Models within the framework established by Turing (Section 2), address the challenges of evaluating and measuring the intelligence of AI systems in contemporary contexts also by formulating a new methodological approach (Section 3), and analyze generative AI models for language through this specific lens (Section 4). Finally, in the conclusion (Section 5), observations will be offered on the potential challenges and developments expected within this field in the near future.

2. Turing was right

Neural language models (NLMs) are artificial neural networks specifically designed for natural language processing (NLP) tasks. Among these, Large Language Models (LLMs) have gained prominence in recent years, representing a key area within generative AI. LLMs are built on the

Transformer architecture[3], which leverages an attention mechanism originally developed for machine translation, a foundational domain in NLP[4]. The Transformer introduces a self-attention mechanism[5], enabling the model to process text sequences by relating different positions within a sequence. Through iterative applications of self-attention, the model forms a holistic representation of the sequence. This approach enhances encoding and decoding processes, offering significantly faster performance compared to recurrent neural networks. Crucially, self-attention facilitates contextual understanding, allowing the model to represent a word's meaning dynamically based on the specific text or sequence in which it appears. As with other neural networks, these representations are vector-based, and computations occur through transformations across multiple intermediate layers.

The technical aspects of these models are essential for understanding their place within the broader category of statistical-predictive systems. This is why they have been described, in the context of language generation, as "stochastic parrots"[6]. Due to their capacity for generating conversational language, these models also potentially align with the concept of "thinking machines" as defined by Turing (1950) prior to the advent of AI. Turing envisioned machines capable of conversing with humans in a way that would make it indistinguishable whether they were interacting with a human or a specially programmed digital computer. As is well known – and extensively discussed in the literature on what is now called the Turing Test[7] – Turing did not specify the exact nature of such machines. He hypothesized they would likely need the ability to learn but offered no guidance on how these thinking programs should be constructed[8].

[3] Vaswani *et al.* (2017).
[4] Bahdanau *et al.* (2015).
[5] Vaswani *et al.* (2017).
[6] Bender *et al.* (2021).
[7] Cfr. Moor (2003).
[8] In fact, there is a significant gap between what Turing envisioned and contemporary LLMs, both in terms of objectives and underlying concepts. Turing's goal in his 1950 paper was to provide an operational means of addressing the question "Can machines think?" while avoiding philosophical entanglements. Linguistic interaction was one of the devices he employed to construct the hypothetical scenario, specifically to create a neutral ground for comparing human beings and appropriately programmed digital computers. This setup originated from the imitation game played between an interrogator on one hand, and a man and a woman on the other. Over time, however, and regardless of whether this was faithful to Turing's original intent, the focus on linguistic interaction became central. It eventually came to define the standard interpretation of the Turing Test, giving rise to a wide-ranging and productive debate.

These kinds of questions emerged a few years later with the advent and subsequent development of AI. However, Turing insists that such machines must be able to play the imitation game, regardless of the specific characteristics of the game itself: "it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man"[9]. If we consider the technical aspects of LLMs, they do not seem to align with what Turing had in mind. At best, they are learning machines, but the sense in which they "learn" is somewhat vague and does not easily lend itself to comparison with Turing's intended claim. Nevertheless, when we examine the actual functioning of pre-trained generative models based on Transformers, the scenario envisioned by Turing appears strikingly relevant. Natural language interaction with these models occurs through prompts – questions or suggestions posed to the program – where inputs can range from multiple examples (few-shot learning) to none at all (zero-shot learning). The progressive refinement of these systems yields natural language outputs that are largely indistinguishable from human-generated text or at least equally comprehensible. From this perspective, LLMs are already capable – and will likely become even more so – of passing the Turing Test in its classical form.

According to Turing (1950: 449), this outcome would not be surprising, given his prediction that within fifty years of his seminal article on computational machines and intelligence, a computer would be able to play and win the imitation game at least thirty percent of the time. In making this prediction, Turing does not concern himself with the specific characteristics of the system capable of winning the game, aside from a general reference to computational resources. What truly matters is that the interaction occurs in a human-like linguistic format. That said, not in all discussions on thinking machines did Turing disregard their cognitive characteristics[10]. However, his 1950 text focuses primarily on the possibility of natural language exchanges between humans and machines. This exchange serves a dual purpose: 1) it places both entities on neutral ground and 2) allows for discussions on any topic, functioning as a kind of generalist methodology. The latest generation of LLMs increasingly align with Turing's vision. They engage in *linguistic* exchanges, exhibit *generality* in the range of topics they can cover, and generate text that is *understandable* in a human-like manner. Even minimal interaction with the most advanced LLMs demonstrates that these

[9] Turing (1950: 437).

[10] Turing addressed the issue of the characteristics that a system capable of learning must have, for example, in a 1951 work, focusing on the role of memory and the way in which it can become increasingly complex and "human". See Turing (1951).

characteristics are met, making it clear that the Turing Test, in its original form, is easily passed.

The push toward simulating human thought is more evident in Turing's 1951 text, suggesting that he himself considered the Turing Test insufficient for assessing the presence of intelligence in a machine, at least when it comes to human-like intelligence: "my contention is that machines can be constructed which will *simulate* the behavior of the human mind very closely"[11]. Although this text is less frequently cited, it introduces a crucial concept for the later development of AI: simulation. It also reinforces the idea that the simulation of human intelligence was already a central theme for Turing. This perspective aligns with how we might evaluate neural network-based LLMs. While they can engage in human-like conversational interactions, they ultimately exhibit only verbal behavior. They are machines – programs – they do not understand the content of their own outputs. Instead, they generate coherent word sequences by computing probabilistic relationships between tokens. In essence, LLMs predict language with remarkable accuracy, but they do not embody intelligence in the sense we typically use the term. In some ways, they could be seen as a modern version of machines that elicit an Eliza effect – a phenomenon named after the ELIZA program developed by Weizenbaum (1966). This interpretation, however, does not fully capture Turing's vision[12]. And the key question remains: do these models actually exhibit intelligence?

An answer of this kind risks being too simplistic. LLMs undoubtedly exhibit a *form* of intelligence; after all, their outputs, whether text, code, or other content, are difficult to dismiss as unintelligent. The question of intelligence in LLMs has been already approached from multiple perspectives. For some scholars – for instance, Millière and Buckner (2024) – the concept of intelligence is too elusive to be meaningfully applied to LLMs. Others have examined the metaphors used to describe generative AI systems, particularly LLMs, to assess the implications of characterizing them as intelligent, especially from an anthropomorphic standpoint[13]. More recently, LLMs have also been employed to investigate various dimensions traditionally associated with intelligence, including different forms of understanding[14] and the relationship between LLMs and the brain[15].

[11] Turing (1951: 472, emphasis added). There is also a reference to simulation in Turing (1950), but only to compare the adult mind with the child mind.
[12] On imitation and LLMs see Boisseau (2024).
[13] Mitchell (2025).
[14] Miracchi & Titus (2024).
[15] Lamarre *et al.* (2022).

A more precise and relevant (for the aim of this article) question might be: can we meaningfully attribute intelligence to them? Or rather, when we speak of intelligence in reference to LLMs, what exactly are we referring to? This question arises precisely because their outputs compel us to recognize a form of intelligence; otherwise, we risk losing sight of what we consider cognitively valuable – content that can be used in epistemic contexts or at least remains intelligible within a cognitive framework. The challenge, then, is: how should we evaluate their intelligence?

3. The quest for evaluating AI intelligence

Since the earliest developments in AI, the challenge of evaluating intelligence in artificial systems has taken on a dual form. On one hand, it has followed in the footsteps of Turing and the Turing Test, generating numerous variations and fueling a decades-long debate[16]. On the other hand, various AI approaches have been examined to determine which best aligns with the goal of creating human-like or cognitively plausible intelligence, at least to some extent[17].This second line of inquiry, often intertwined with the evolution of cognitive science, rests on the assumption that artificially replicating human cognition is, by definition, a valid means of simulating intelligence. In other words, if intelligence is a defining characteristic of human beings, then reproducing their cognitive mechanisms, functions, and properties should lead directly to the simulation of intelligence. The main issue with this perspective is its excessive anthropocentrism. This concern has been particularly noted in relation to classical AI and its symbolic approach[18]. However, even in more recent developments in AI – shaped by the embodied turn in cognitive science and the rise of bio-inspired computational architectures – traces of anthropocentrism persist, in line with a view that underlines a partial continuity between classical approaches and new approaches to cognitive science[19].

Of course, the evaluation of intelligence of AI also forms part of a broader and long-standing debate concerning the nature of intelligence itself, a debate that has not always been approached from a human-centered perspective. In the context of AI, intelligence has, for instance, been investigated as a property of systems, often linked to rationality as a defining

[16] Moor (2003).
[17] Cfr. for example, Boden (2006).
[18] Preston (1991).
[19] Shapiro (2019).

feature of intelligent behavior[20]. More recently, scholars have explored the nature of both human and machine intelligence in relation to creativity[21], as well as to capacities such as perception, understanding, and abstraction within learning processes[22].

Recent developments have introduced new methods for evaluating intelligence in artificial systems, shifting the focus toward measuring intelligence rather than treating it as a simple yes/no question. These methodological approaches recognize that the issue is tied to a broader, unresolved question: what is intelligence?

Moreover, the challenge of attributing intelligence has gained increasing importance in recent years, driven by the widespread proliferation of AI systems. Over the past fifteen years, AI has transitioned from a specialized technological discipline, primarily confined to niche applications, to a widely accessible software technology used by the general public. The advent of LLMs has further accelerated this diffusion, leading to the growing, often unreflective, integration of AI systems into everyday life. This raises critical questions about how users perceive both the performance and the outputs of these systems. In particular, this development brings forth a range of ethical and societal concerns, spanning multiple domains, including culture, education, information dissemination and communication, marketing, and commerce, among others. The central thesis of this work is that the ways in which intelligence is attributed to and evaluated in AI systems are increasingly relevant for their appropriate deployment and integration into society. Furthermore, this form of evaluation is, at its core, an epistemic issue with significant epistemological dimensions. In the following pages, key guidelines are outlined for constructing such an evaluation framework, referring to Bianchini (2024) for a more detailed discussion.

The problem of attributing and evaluating intelligence goes deeper than the simple Eliza Effect mentioned earlier. It is not merely about the possibility of being "fooled" by systems that employ tricks to create the illusion of intelligence in their behavior or outputs. In other words, it is not just a contemporary manifestation of the broader human tendency to attribute intentionality or understanding as part of cognitive processing. This debate has been central to the philosophy of AI for decades[23] and remains active, particularly in discussions surrounding human-artificial system interaction,

[20] Russell (1997).
[21] Boden (2016).
[22] Mitchell (2019).
[23] Cfr. Dennett (1987) and Searle (1983).

especially in robotics[24]. However, the discourse on the attribution of intentional attitudes – while fundamental to the philosophy of mind and crucial in human-robot interaction – primarily concerns unreflective attribution. That is, it examines the natural human tendency to ascribe intentionality, and by extension, intelligence somehow, to non-human entities, particularly artificial systems, thereby granting them an appearance of cognitive/intelligent capacity. Recently, scholars have questioned whether such attribution occurs as widely as traditionally assumed. Some argue that certain forms of anthropomorphizing may be more myth than reality[25]. Nevertheless, even if such attributions are less pervasive than once believed, their persistence underscores the significance of this issue in AI research. It remains crucial not only for understanding human interactions with AI systems but also for assessing these systems both as practical tools and as subjects of theoretical analysis from the point of view of intelligence.

In this regard, it is necessary to take a further step and consider the interaction with AI systems, particularly in relation to evaluating their intelligence. As previously mentioned, the classic attribution of intentionality is largely considered an automatic cognitive act rather than a conscious assessment. A conscious attribution, however, is based on expected intelligence and can serve as the foundation for new approaches to measuring intelligence. The notion of *expected intelligence*, which is closely tied to an interactive approach to AI, provides a basis for evaluating the intelligence of an artificial system through the initial assumptions made by the user interacting with it. In this context, expected intelligence refers to the largely conscious tendency to engage with an artificial system from which an epistemically and/or cognitively relevant response is anticipated. In other words, expected intelligence functions as a precondition for recognizing, and thus evaluating, an artificial system as autonomous system. Without this consciously held precondition, the system's behaviors and outputs would not necessarily be interpreted within a meaningful framework and might instead be regarded as mere occurrences or mechanical reactions to specific stimuli. In AI, and particularly in fields such as robotics, the interactive approach relies on the concept of expected intelligence both to explain the behavior of artificial agents and to guide their design in relation to cognitively capable users, namely human beings. This concept thus serves as the starting point for evaluating attributed intelligence[26].

[24] Wykowska (2024).
[25] Coghlan (2024).
[26] Bacaro & Bianchini (2024).

The ability to assess the attributed intelligence – or lack thereof – of a system is crucial not only for understanding AI itself but also for evaluating its broader social and technological impact. This evaluation plays a significant role in addressing the Collingridge Dilemma[27], which highlights a fundamental challenge in technology governance: some technologies are difficult to predict in terms of their societal impact until they become widely adopted, yet by that time, they are often difficult to control or modify, particularly in terms of their standardized use. AI systems developed over the past decade fit this dilemma perfectly, especially those that are easily accessible and widely used. LLMs provide a clear example. Their rapid and widespread adoption is largely due to their impressive capabilities, yet their long-term impact remains under scrutiny. The widespread diffusion of LLMs, whose consequences remain difficult to fully anticipate, has given rise to a broad spectrum of ethical issues. These range from the potential amplification of misinformation and the reinforcement of biases to the economic and social impacts of their deployment, as well as concerns about reliability, particularly regarding the data on which these models are trained. While some of these challenges are common to all systems based on deep neural networks, they become especially critical in domains where text production and the use of knowledge are foundational, such as education, or where data usage, transparency, and reliability are essential prerequisites for application, as in the medical field[28]. Developing a conscious evaluation of the expected intelligence of such systems – beyond merely assessing their efficiency and accuracy – could offer a means of mitigating the challenges posed by the Collingridge Dilemma, particularly where predictive limitations arise, and the related ethical issues.

Let us now examine in more detail how to evaluate the expected intelligence of an AI system. In the first place, this issue can be seen as equivalent to measuring the intelligence of an artificial system deemed intelligent. The systematic analysis of intelligence measurement in AI has gained attention only relatively recently and has led to two primary characterizations[29]: a) intelligence as a set of task-specific skills; b) intelligence as a general ability to learn and perform open-endedly. In the first case, the focus is on measuring the accuracy of an AI system's performance. Here, no generalization occurs – neither within the system itself (narrow generalization) nor through developer-implemented methods (broad

[27] Collingridge (1980).
[28] Ong *et al.* (2024).
[29] Hernández-Orallo (2017), Chollet (2019).

generalization). In the second case, the aim is to assess how well specific abilities can be generalized across multiple domains. This approach is reminiscent of Newell, Shaw, and Simon's (1959) General Problem Solver and is further developed in modern cognitive architectures such as SOAR and ACT-R.

The first approach – measuring task-specific performance – appears particularly well-suited for evaluating AI systems. This is because it allows for the construction of a measurable value scale, typically based on accuracy. Such measurements are often carried out relative to a predefined standard or as an average over multiple performances. In this framework, assessing AI intelligence usually entails evaluating task-oriented performance on a scale, where a "good" or "poor" performance is determined by specific parameters. This process is inherently deliberate and guided by a well-defined objective. Hernández-Orallo (2017) identified three types of methods and metrics aligned with this perspective, focusing on: 1) human discrimination; 2) problem benchmarks; 3) peer comparison. The first approach is inherently subjective and remains within an anthropocentric framework. The other two involve comparison either with a predefined standard or with the average performance of other systems or human participants performing the same task. In this sense, they can be considered more objective and provide effective parameterization, even if they are limited to highly specific tasks, such as categorization in a neural network or user preference profiling.

The challenge arises with generality – specifically, the evaluation of AI systems' intelligence across different domains and from an indeterminate perspective. In particular, how can we assess intelligence based on abilities, focusing on broader cognitive aspects? The risk here is falling into anthropocentrism, searching for cognitive traits within AI systems. While this approach aligns with cognitive science's historical research programs[30], it differs from evaluating a system's expected intelligence, an issue that remains neutral regarding whether AI systems possess cognitive qualities. On the other hand, adopting a neutral formal standard for evaluating AI intelligence – such as one based on algorithmic information theory[31] – risks resulting in an opaque assessment. This is because objective measurement elements would primarily relate to different dimensions of algorithmic complexity. However, complexity and information are not *directly* equivalent to intelligence. In other words, while intelligence can be seen as a property of

[30] Boden (2006).
[31] Chaitin (1987).

complex systems, it does not follow that every complex system capable of processing information is necessarily intelligent.

To address the challenge posed by generalist approaches to AI – particularly in assessing their adaptability across multiple contexts, a key hallmark of intelligence – three distinct theoretical responses can be considered. First, one might argue that AI systems are not intelligent at all but merely instruments of action[32]. This perspective rests on the assumption that an intelligent outcome is not always an intelligent behavior, or the result of an intelligent behavior. While this "eliminativist" stance on AI intelligence may seem too radical, it has the merit of distinguishing between intelligence as an intrinsic property of the system and the attribution of intelligence to the system itself.

A second possible response focuses on the social and interactive aspects of AI systems[33]. The study of human-AI interaction has a long history, and interactional perspectives have gained increasing relevance, partly due to the rise of embodied approaches, such as enactivism, within cognitive science, particularly in relation to artificial systems. Without committing to a specific theory of cognition, a general assumption in this view is that, in most cases involving human users and AI systems, the attribution of intelligence by the human user, often in real-time, is crucial for achieving optimal results and effective interaction. In other words, without the presumption of a shared cognitive framework, which falls within the broader concept of intelligent behavior, meaningful interaction becomes unlikely. Instead, the AI system risks being reduced to a mere tool used by the human operator.

Finally, a third possible response arises from the debate on the attribution of mental states to artificial systems[34]. The tendency of humans to ascribe mental states –particularly to robotic artifacts – is one possible explanation for the way we interact with certain AI systems. This attribution is not necessarily limited to robotic systems; it can also extend to other artificial entities perceived as intelligent. Within this perspective, the debate remains open regarding the ontological status of these attributed mental states and the various approaches to assigning intentionality to AI systems[35]. Nevertheless, while attributing mental states can serve an explanatory role in understanding human-AI interaction, it does not necessarily address the issue of intelligence itself. Intelligence, in this sense, remains conceptually distinct from the cognitive elements we might identify when evaluating these systems. In other

[32] Floridi (2023).
[33] Cristianini *et al.* (2023).
[34] Thelmann *et al.* (2022).
[35] Larghi & Datteri (2024).

words, the attribution of intelligence to an AI system appears to be independent of what we believe is happening within the system, even from an attributional standpoint.

If, on the one hand, we wish to avoid overly deflationary positions regarding the intelligence of AI systems, and, on the other, set aside considerations about how these systems are designed or aligned with recognized cognitive systems – primarily humans – the behavioral perspective remains the most central approach for the conscious attribution of intelligence[36]. This perspective, which can be seen as partly inheriting Turing's legacy, allows us to analyze the attribution of intelligence from the user's standpoint, emphasizing its role as an essential requirement for the epistemic, applied, and ethical functioning of AI systems.

The attribution of intelligence from the user's perspective can be developed along four dimensions[37]:

Before interaction – Based on the user's preliminary knowledge of the AI system.

During interaction – While actively using or engaging with the system.

Post-interaction – Evaluating the system's performance and the outcomes it produces.

Over repeated interactions – Assessing intelligence attribution over time, considering potential variability in perception.

In all these cases, the system's behavior is evaluated in broad terms. This evaluation can concern both performance on a specific task, especially when repeated with varying results, and the system's behavior from a more general perspective. The latter involves determining whether the system demonstrates general capabilities beyond isolated tasks, indicating a broader implementation of intelligence.

Finally, different metrics can be devised to best capture the four dimensions of evaluation, depending on the specific context. Without aiming for exhaustiveness, at least two broad categories of applicable metrics can be identified.

The first category includes metric formats based on scalar dimensions within a defined range: for example, Likert-type scales. These can vary in granularity depending on the level of detail desired (e.g., five-point, seven-point, or ten-point scales). A higher level of detail may be appropriate for assessing the attribution of intelligence in scientific or experimental settings,

[36] For a behavioral perspective on evaluation in terms of prediction see Cevolini, Esposito (2022).

[37] For a more detailed description see Bianchini (2024).

while lower-resolution scales can support self-assessment by users regarding their interaction with an AI system. In such cases, the aim may be to promote user self-awareness and responsibility, to implement nudging strategies, or to generate aggregate rating data that can inform system design or incremental improvements. It is also worth noting that the four dimensions allow for a temporal assessment of the attribution of intelligence within the interactive process, whether it is increasing, decreasing, or remaining stable. This temporal perspective can help identify the phases of interaction in which perceived or attributed intelligence is heightened or diminished. For instance, a decreasing attribution over time may indicate that the system is perceived as displaying a weak degree of "artificial intelligence", and thus as being less reliable or accurate in relation to user expectations.

The second type of metric could instead leverage the direct relationship with the user, considered as a median point of reference. From this midpoint, the user would assign scores indicating whether the intelligence attributed to themselves is greater or lesser than that attributed to the system at various stages of the interaction. As with the first type, these metrics could vary in granularity depending on their intended purpose. The goal of this approach is to place the user even more centrally in the process of attribution, encouraging them to assess their own intelligence in comparison to that of the system. This can have several theoretical implications for research on human-AI interaction, as well as practical benefits. For example, it may promote more conscious and constructive use of the system, help identify weaknesses in the interaction, and foster more responsible usage, especially in contexts where there is a risk of user deskilling (among which the educational one).

4. Measuring expected artificial intelligence and the case of LLMs

Beyond the potential metrics that could be developed using these four dimensions – aimed at refining the measurement of AI intelligence across different application domains – this proposal seeks to capture a fundamental practical principle: *intelligence is attributed when it is expected, and it is expected when it is attributed*. This principle applies particularly to AI systems, which are defined within the broader field of artificial *intelligence* and are characterized by their capacity to implement intelligent behavior autonomously, another key criterion of AI.

This discussion has significant methodological implications for investigating AI in relation to human intelligence and cognition. Since the inception of AI, researchers have explored the possibility of constructing AI

systems as a means of understanding human cognition and its processes[38]. However, the principles underlying this undertaking can be generalized. The assumption underlying our approach is that expected intelligence – attributed to a system by human observers – is coupled with *something* underlying that enables intelligent behavior. This something, in turn, serves as a preliminary condition for recognizing intelligence. Such an assumption carries two important implications. First, it justifies treating the system as intelligent, meaning we must expect it to perform actions we consider intelligent; otherwise, we risk falling into deception or misconception. Second, it places a demand on human intelligence itself: the system's behavior must be authentically intelligent, rather than a collection of superficial tricks that undermine the legitimacy of considering it truly intelligent.

The crucial question, then, is: where do we draw the line between authentic intelligence and mere imitation? To avoid anthropomorphism or the assumption that intelligence must emerge from specific internal mechanisms modeled on human cognition, we can turn to the concept of expected intelligence as a measurable phenomenon. This allows us to address the boundary between intelligence and non-intelligence in a more gradual and pragmatic way, aligned with real-world interactions between humans and AI systems. This behavioral perspective has the further advantage of avoiding a human-centric commitment to what constitutes intelligence. In other words, the processes that give rise to intelligence in an AI system do not necessarily have to mirror those found in human cognition.

On the other hand, this perspective carries the risk of leading to an overly anarchic approach to the attribution of intelligence. If intelligence could be ascribed to virtually anything without a clear justification, the concept itself might lose its meaning. Therefore, it seems necessary to also consider the issue from the opposite standpoint. To avoid such conceptual chaos – where intelligence could be arbitrarily attributed without a solid basis – there must be some criterion to justify the attribution. This criterion could take the form of a mechanism, a technique, a dynamic interaction, a mathematical or statistical function, or any other structured method. While this criterion does not necessarily need to be predetermined – allowing for a certain degree of flexibility or a standby approach – it must still exist in some form to preserve the coherence of the notion of artificial intelligence as applied to the system in question. In practice, the loss of this notion is not what we observe in the real world. Instead, the attribution of intelligence, at least to some degree, to AI systems is something we do continuously.

[38] Cordeschi (2002).

Leaving aside analytical metrics, let's attempt to transform the four dimensions by which we define the evaluation and measurement of an AI system's intelligence into a methodological approach. It will then be considered its epistemological significance. The steps of this methodological process could be as follows:

1. *Assuming* the possibility of using or interacting with an AI system.
2. *Expecting* intelligence in the system.
3. *Attributing* intelligence (hypothetically) to the system.
4. *Attributing* or "*finding*" intelligence (actually) to the system.
5. *Identifying* the "reason" of intelligence in the AI system (the research-oriented step).

The first four steps can be applied whenever one encounters an AI system or a system presented as such. Confirming step 4 in this process amounts to recognizing the system as genuinely intelligent and potentially assigning a measurement index to this characteristic. Step 5 is optional and relevant primarily when situating the system within a particular AI framework, or multiple convergent AI approaches, for research, regulatory, ethical, or legal purposes.

In more detail, the four dimensions previously described can be mapped onto this methodological process as follows. The *"before"* dimension corresponds to steps 1-2-3, as it involves moving from the initial assumption to a hypothetical attribution of intelligence. The *"during"* dimension spans steps 2-3-4, since it covers the transition from the evaluation of expected intelligence to its actual attribution to the system. The *"after"* dimension pertains to steps 3-4 and specifically involves the ex-post assessment of the shift from hypothetical to actual attribution. Finally, the *iterative* dimension encompasses steps 1–2–3–4, as the evaluation process is designed to be repeated over time.

In general, steps 1 to 5 can be understood as a form of reverse engineering through interaction. That is, rather than beginning with the acknowledgement of predefined cognitive capabilities, one could start from direct engagement with the system. More specialized competencies of experts would come into play at a later stage of analysis. This approach would enable even non-experts to engage with AI systems in an informed manner, using an initial heuristic evaluation and measurement method to navigate their interactions effectively. It is important to note that as AI systems become increasingly integrated into daily life and accessible to all types of users, this methodological framework will be crucial. It will serve not only as a means of maintaining oversight of

AI systems but also as a foundation for sustainable and informed interactions with them – an essential aspect of the society of the coming decades.

The applicability of this method is broad within the field of AI and extends to all users of AI systems, including those interacting with content profiling tools, voice assistants, medical and educational technologies, autonomous vehicles, and even autonomous weapons. In all these cases, both general users and experts – though not necessarily AI specialists – can engage with AI systems and analyze their interactions from an intelligence-based perspective.

Among the most prominent AI systems today are generative AI systems, particularly neural large language models, already mentioned in the earlier sections. LLMs possess distinctive characteristics that make them especially well-suited for evaluation through the methodological framework outlined above. Their performance can be assessed in a task-oriented manner across various domains, and they belong to the broader neural network paradigm, which is explicitly designed to handle diverse tasks. This inherent generality, however, presents a challenge: it is often too expansive to be meaningfully evaluated as a single entity. Nevertheless, LLMs demonstrate linguistic competence across a vast array of subjects, suggesting that language might represent the appropriate level of generality at which to assess their intelligence. Moreover, many contemporary models are multimodal, capable of processing text, images, and code as inputs. Essentially, these systems perform a specific task, language processing, but in a way that connects to a wide range of topics. In this sense, they can be seen as task-specific systems exhibiting a form of general intelligence – namely, linguistic intelligence in the broad sense. For this reason, LLMs appear to strike the right balance between specialization and generality for assessing AI intelligence: they are neither so narrowly focused as to reduce their cognitive potential to a single capability nor so broadly defined as to make their intelligence indistinguishable from mere computational complexity.

From the perspective of LLMs, the five steps introduced are easily applicable. The growing confidence in these systems parallels their rapid diffusion, which aligns perfectly with Collingridge's dilemma[39]. This, in turn, underscores the need for a more conscious and responsible use of "intelligent" tools. Let us now explore how, at an epistemological level, these AI systems can be evaluated in relation to textual production.

First, there is now broad consensus that LLMs should be regarded as intelligent tools, not merely in the generic sense of being AI systems, but in

[39] Collingridge (1980).

the more substantive sense of enabling the production of outputs recognized as intelligent. This directly leads to step 2: the expectation that the system will generate texts that are coherent, meaningful, relevant to user queries, and cognitively adequate for human understanding. This step – akin to a "Turing step" – is generally satisfied, particularly by the most advanced LLMs, which can respond to a vast range of natural language requests across an indeterminate number of topics. Step 3 follows: the hypothetical attribution of intelligence to the system itself, rather than just its outputs. This step is a generalization, where the system's intelligence is empirically inferred from the quality of its textual productions and extrapolated into a broader hypothesis of general intelligence. Step 4 involves confirming this attribution of intelligence, which can be assessed using the four temporal dimensions of interaction previously mentioned. These dimensions can also be quantified to allow for a more gradual evaluation of intelligence, moving beyond a binary distinction between intelligence and non-intelligence. For instance, intelligence can be evaluated through a metric that assesses the comprehensibility, relevance, and coherence of the generated texts, features typically associated with intelligence. Similarly, the fourth dimension, repeated use, can help determine the system's reliability: whether it produces false information, when it starts generating hallucinations (i.e., plausible but incorrect content based on its training data, now a well-documented characteristic of LLMs[40]), and the extent to which it exhibits standardization or stylistic repetition. This longitudinal evaluation can also assess whether errors are present and how they evolve over time.

The value of steps 1–4 lies in their ability to provide all users with a framework for evaluating AI systems, fostering a bottom-up approach that enables meaningful interaction with intelligent systems. This evaluation allows users to assess the system's potential actions, activities, or behaviors from the perspective of intelligent understanding. Such an approach not only helps in interpreting the capabilities and limitations of LLMs but also extends to other AI systems. Consider, for example, interactions with fully autonomous vehicles, AI-driven medical applications, or even autonomous weapons. A precise understanding of their "intelligent" behavior is essential for ensuring safe and effective interaction, especially in high-stakes scenarios where errors could lead to disastrous and irreparable consequences.

The application of steps 1–4 can yield particularly interesting and practical outcomes for users not primarily concerned with research purposes, especially in the case of LLMs. While the attribution of intelligence in other

[40] See Farquhar *et al.* (2024).

AI systems often serves as a prerequisite for assessing their reliability, LLMs introduce a distinct epistemic dimension. Consider, for instance, navigational systems or self-driving vehicles. We trust their intelligence insofar as we delegate to them tasks that we would typically perform using our own cognitive abilities. This trust largely hinges on the extent to which we attribute intelligence to these systems, especially since, in most cases, we lack detailed knowledge of the technical mechanisms underlying their autonomous operation. Take the extreme example of a monorail transporting passengers between terminals in an airport without a human operator. It is relatively easy to trust such a system because we can readily imagine the limited and well-defined nature of the task, which seems to require only a modest level of intelligence, if any, by everyday standards. We can roughly grasp how its autonomy functions and feel comfortable attributing it with just enough intelligence to fulfill that role. By contrast, navigation systems or autonomous vehicles involve a far greater number of variables, and their functioning depends on mechanisms that are more opaque and harder to conceptualize. In these cases, the attribution of intelligence is closely tied to the reliability we are prepared to grant them, perhaps based on direct experience or observed behavior. A mistake or failure would diminish our trust, effectively lowering the degree of intelligence we attribute to the system. As a result, we may become reluctant to rely on it again unless significant improvements and verifiable changes are made.

Let us now consider the case of LLMs. In this context, we cannot merely observe their behavior as with other AI systems; rather, we must assess the products they generate through interaction to judge their intelligence. Unlike in other systems, intelligence here is not conflated with reliability, something we may be willing to compromise on, as long as we are aware of it and the system remains useful, but is instead tied to usability itself. In the case of LLMs, the pragmatic dimension gives way to a cognitive-epistemic one. If we did not regard LLMs as intelligent, that is, as capable of producing coherent, comprehensible texts aligned with our prompts and responsive to the real world, we would have no reason to use them. To attribute such capabilities, however, we must expect LLMs to produce "intelligent" texts: texts that possess semantic interpretability, epistemic content, and – crucially – some trace of the evidentiary or inferential structure that would allow us to confirm or contest their claims. Only by attributing a degree of intelligence to an LLM can we evaluate its textual outputs according to these criteria, much as we routinely do with human interlocutors. If an LLM fails to meet these standards, we cease to use it. If it succeeds, then, even if only with regard to its outputs, we are implicitly attributing to it a minimal cognitive

common ground. This common ground may shift depending on context, users, or over time. However, the more robust and recognizable this shared cognitive basis becomes, the more intelligence we attribute to the model, and the more inclined we are to engage with it. Importantly, this attribution does not require that the LLM possess intelligence of the same kind as human beings. Even outputs containing hallucinations may offer useful information, despite their misleading nature. We recognize the value in such texts because we attribute to the system a degree of intelligence, albeit a limited one, sufficient to distinguish them from mere juxtapositions of words devoid of meaning or relevance.

The key point with LLMs is that, unlike other AI systems, their adoption has been significantly more widespread and rapid. Moreover, unlike other forms of AI, it is difficult to define a fixed set of instructions to learn how to use them correctly. Instead, it is through use and interaction that users gradually learn how to operate them effectively. For this reason, attributing intelligence to these systems becomes a necessary first step, one that users must continually take to engage with them appropriately. This consideration also applies to domain experts who may not be directly involved in AI research. For instance, professionals such as lawyers or physicians can rely on LLMs to support their work, but they must be able to assess the degree of intelligence these systems display in their respective fields. This is essential to avoid risks such as bias or epistemic injustice. In such cases, knowing how the systems work is not sufficient. Proper use of these tools depends on the textual knowledge they produce in interaction, more specifically, on the user's ability to interpret their output appropriately and to formulate prompts competently, in line with the capabilities attributed to the system.

Step 5 of this methodological approach addresses more advanced research interests and involves experts working with AI in various capacities. The question of what underlies the intelligence observed in these systems is both a matter of practical design and implementation and a theoretical issue within an epistemological framework. Thus, answering the question "what is intelligence due to in this reverse engineering process?" has both practical and conceptual implications. For instance, determining whether intelligence in the system arises from statistical-predictive methods, mechanisms, network topology, structural design, architecture, inferential and/or representational abilities, or a combination of these factors can provide insights in multiple ways. It can inform the development of more efficient AI systems, enhance our understanding of intelligence and cognition, and help explain why AI systems are often perceived as intelligent from different perspectives.

An analytical consideration of the "reason" behind intelligence in AI systems can contribute to addressing several key challenges. It can aid in solving the problem of AI explainability[41]; it can help overcome anthropomorphism in the analysis of AI systems by identifying techniques that, while distinct from human reasoning, are nonetheless effective within specific programming domains, such as certain cases of supervised learning[42]; it can tackle semantic issues like symbol grounding, which remain relevant even in the latest AI systems, particularly in neural LLMs[43]. Additionally, this approach can support the development of models capable of inferring or deriving others' intentions and beliefs, so provided with a form of Theory of Mind[44]. Finally, and perhaps most importantly in relation to LLMs, such a methodological perspective can help determine the epistemic reliability of the texts these models generate. Specifically, it can assess to what extent we can trust the knowledge embedded in their outputs, both in particular cases and in general, thereby allowing us to evaluate their strengths and weaknesses as "epistemic authorities"[45].

A final consideration must be given to the risk of anthropomorphism, which increasingly concerns AI systems, especially generative ones, such as large language models (LLMs). Since the four evaluation dimensions outlined above, along with the proposed methodological process, are grounded in interaction between the AI system and the human user, and since the evaluation is carried out by the user on the basis of that interaction, the risk of anthropomorphic attribution is heightened. In other words, there is a growing tendency not only to interpret the behaviors and outputs of the system as anthropomorphic, but also to expect exclusively such behaviors, thereby selecting or misinterpreting those that fall outside this frame. This risk, however, is inherent in any process involving the attribution of cognitive features. The evaluation methodology proposed here should thus be understood in continuity with broader philosophical reflections on intentionality. Dennett himself – one of the most influential theorists of intentionality – argued that a necessary condition for attributing intentional states to a system is the presence of rationality, a rationality modeled on the human mind and shaped by evolutionary processes[46].

[41] Miller (2023).
[42] Watson (2019).
[43] Pavlick (2023).
[44] Nguyen & Gonzalez (2022).
[45] Ferrario *et al.* (2024).
[46] Dennett (1987, 1991).

The attribution of intelligence can be understood as a renewed form of attributing intentionality, one that focuses more on behavior and outputs than on the internal states of a system. However, the criteria used for such evaluation risk falling into the same anthropomorphic assumptions. How can we judge something to be intelligent except on the basis of what we already consider to be intelligent? Admittedly, knowing that we are dealing with an artificial system should prompt us to suspend judgment regarding the forms of intelligence we attribute, considering them with broader scope and a greater openness to possibilities beyond those supported by the "reasons" discussed in step 5 (e.g., similarity in structure or mechanisms with human beings). Yet even in this broader framework, the risk remains. If we combine the tendency toward anthropomorphism with automation bias, that is, the human predisposition to favor the outputs of artificial systems[47], we may similarly overestimate or over-rely on these systems' cognitive capacities. Just as automation bias can lead to an undue acceptance of machine-generated suggestions in decision-making, it can also foster, by analogy, an inflated attribution of intelligence to these systems. This risk becomes particularly pronounced when the system engages in human-like interaction, as in the case of linguistic exchanges with LLMs. This recurring challenge in the development of AI systems may be mitigated by cultivating greater user awareness and responsibility. As AI systems are continuously modified and improved, their performance becomes increasingly difficult to distinguish from human-like behavior, blurring the boundaries and increasing the likelihood of misattribution. One of the key aims of the four interactive dimensions proposed for evaluating the attribution of intelligence is precisely to foster this kind of awareness.

5. Conclusion

This paper aimed to address the evaluation of AI systems within the domain of neural network-based LLMs. The discussion of these models' intelligence began with an analysis of Turing's ideas, updated in light of the capabilities and behaviors of LLMs. The evaluation of intelligence was then reframed beyond the mere detection of its presence or absence in AI systems, particularly in LLMs. Recent developments in the debate on AI intelligence measurement were examined, highlighting the current focus on two main approaches: the evaluation of task-oriented systems, which excel in specific

[47] Skitka *et al.* (1999).

domains, and abilities-oriented systems, which demonstrate a broader and more general form of intelligence.

The proposal presented in this paper emerges from a reversal of perspective. Given the widespread adoption of AI systems, these technologies can now be considered accessible to a large number of users. Starting from the users and their *interactions* with AI, this paper argues for the necessity of fostering a conscious and informed use of these systems. This begins with examining the intelligence attributed to AI in relation to the intelligence expected within the user's knowledge context. Such conscious use can be guided by measuring the characteristics of AI intelligence along four temporal dimensions of interaction: before, during, after, and iterative. These dimensions can be translated into metrics to evaluate the various stages of the methodological approach proposed in the second part of this paper. This approach redefines the attribution and identification of intelligence by prioritizing user interaction over the intrinsic nature or design of the system. In this framework, understanding the underlying mechanisms of AI becomes the final step rather than the starting point, offering benefits across various fields while minimizing anthropocentrism and anthropomorphism. Neural LLMs serve as a prime example of widely adopted, interactive AI systems capable of generating behavior commonly perceived as intelligent. The proposed methodological approach has been applied to these models to illustrate potential research directions on LLMs and to explore the nature of intelligence in artificial systems.

References

Bacaro, M., & Bianchini, F. (2024), "Artificial Intelligence as Expected Intelligence", in F. Bianchini, V. Fano, P. & Graziani (eds.), *Current Topics in Logic and the Philosophy of Science. Papers from SILFS 2022 postgraduate conference*, College Publications, Rickmansworth, pp. 89-115.

Bahdanau, D., Cho, K., & Bengio, Y. (2015), "Neural machine translation by jointly learning to align and translate", conference paper in International Conference of Learning Representations 2015.

Bender, E. M., 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

Bianchini, F. (2024), "Evaluating Intelligence and Knowledge in Large Language Models", in *Topoi*, pp. 1-11. https://doi.org/10.1007/s11245-024-10072-5

Boden, M. (2006), *Mind as Machine. A History of Cognitive Science*. Oxford, Clarendon Press.

Boden, M. (2016). *AI. Its Nature and Future*. Oxford, Oxford University Press.

Boisseau, É. (2024), "Imitation and Large Language Models", in *Minds and Machines*, 34, 42. https://doi.org/10.1007/s11023-024-09698-6

Cevolini, A., & Esposito, E. (2022), "From Actuarial to Behavioural Valuation. The Impact of Telematics on Motor Insurance", in *Valuation Studies* 9 (1), pp. 109-39. https://doi.org/10.3384/VS.2001-5992.2022.9.1.109-139.

Chaitin, G. J. (1987), *Algorithmic information theory*, Cambridge, Cambridge University Press.

Chollet, F. (2019), "On the Measure of Intelligence", in *arXiv*:1911.01547v2, https://doi.org/10.48550/arXiv.1911.01547

Coghlan, S. (2024), "Anthropomorphizing Machines: Reality or Popular Myth?", in *Minds and Machines*, 34, pp. 1-25. https://doi.org/10.1007/s11023-024-09686-w

Collingridge, D. (1980), *The Social Control of Technology*, New York, St. Martin's Press.

Cordeschi, R. (2002), *The discovery of the Artificial. Behavior, Mind and Machines before and beyond Cybernetics.* Dordrecht/Boston/London: Kluwer Academic Publishers.

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

Dennett, D. C. (1987), *The Intentional Stance*, Cambridge Mass., The MIT Press.

Dennett, D.C. (1991), *Consciousness Explained*, Boston, Little, Brown and Co.

Farquhar, S., Kossen, J., Kuhn, L. *et al.* (2024), "Detecting hallucinations in large language models using semantic entropy", in *Nature*, 630, pp. 625–630. https://doi.org/10.1038/s41586-024-07421-0

Ferrario, A., Facchini, A., & Termine, A. (2024), "Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems", in *Minds and Machines*, 34, 30. https://doi.org/10.1007/s11023-024-09681-1

Floridi, L. (2023), *The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities*, Oxford, Oxford University Press.

Hernández-Orallo, J (2017), "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement", in *Artificial Intelligence Review*, 48, pp. 397-447. https://doi.org/10.1007/s10462-016-9505-7

Kelly, S., Kaye, S., & Oviedo-Trespalacios, O. (2023), "What factors contribute to the acceptance of artificial intelligence? A systematic review", in *Telematics and Informatics*, 77, https://doi.org/10.1016/j.tele.2022.101925.

Lamarre, M., Chen, C., & Deniz, F. (2022), "Attention weights accurately predict language representations in the brain", in *bioRxiv*, 2022-12.

Larghi, S., & Datteri, E. (2024), "Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance", in A. Aldini (eds), *Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. SEFM 2023 Lecture Notes in Computer Science*, 14568. Cham, Springer. https://doi.org/10.1007/978-3-031-66021-4_2

Millière, R., & Buckner, C. (2024), "A Philosophical Introduction to Language Models, Part II: The Way Forward", in arXiv: 2405.03207

Mitchell, M. (2019), *Artificial Intelligence: A Guide for Thinking Humans*, New York, Farrar, Straus and Giroux (FSG).

Mitchell, M. (2025), "The metaphors of artificial intelligence", in *Science*, 386, 6723, DOI: 10.1126/science.adt6140

Miller, T. (2023), "Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI", in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Miracchi Titus, L. (2024), "Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy", in *Cognitive Systems Research*, 101174.

Moor, J. H. (ed.) (2003), *The Turing Test. The Elusive Standard of Artificial Intelligence*, Dordrecht, Springer. https://doi.org/10.1007/978-94-010-0105-2

Newell, A., Shaw, J. C., Simon, H. A. (1959), "Report on a general problem-solving program", in *Proceedings of the International Conference on Information Processing*, pp. 256–264.

Nguyen, T. N., & Gonzalez, C. (2022), "Theory of Mind from Observation in Cognitive Models and Humans", in *Topics in Cognitive Science*, 14, pp. 665–686. 10.1111/tops.12553

Ong, J. C. L., Chang S. Y., Wasswa, W., Atul, J. B., Nigam, H. S., Lita, S. T. C. *et al.* (2024), "Ethical and regulatory challenges of large language models in medicine", in *The Lancet Digital Health*, 6, 6, e428 - e432. 10.1016/S2589-7500(24)00061-X External Link

Pieraccini, R. (2012), *The Voice in the Machine. Building Computers That Understand Speech*, Cambridge Mass., The MIT Press.

Preston, B. (1991), "AI, anthropocentrism, and the evolution of 'intelligence'", in *Minds and Machines*, 1, pp. 259–277. https://doi.org/10.1007/BF00351181

Russell, S. (1997), Rationality and Intelligence, in *Artificial Intelligence*, 94, pp. 57-77.

Searle, J. (1983), *Intentionality: An Essay in the Philosophy of Mind*, New York, Cambridge University Press.

Shapiro, L. (2019), *Embodied cognition*, 2nd edition, New York, Routledge.

Skitka l. j., Mosier, K. L., & Burdick, M. (1999), "Does automation bias decision-making?", in *International Journal of Human-Computer Studies*, 51, 5, pp. 991-1006, https://doi.org/10.1006/ijhc.1999.0252

Thellman, S., de Graaf, M., & Ziemke, T. (2022), "Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings", in *Journal of Human-Robot Interaction*, 11(4). https://doi.org/10.1145/3526112

Turing, A. M. (1950), "Computing Machinery and Intelligence", in *Mind*, 59, pp. 433–460, (reprinted in J. Copeland (ed.), *The essential Turing,* Oxford, Oxford University Press, 2004, pp. 441–464).

Turing, A. M. (1951), "Intelligent Machinery. A Heretical Theory", in J. Copeland (ed.), *The essential Turing,* Oxford, Oxford University Press, 2004, pp. 472–475.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017) "Attention is all you need", in *Proceedings of 31st International Conference on Neural Information Processing Systems* (NeurIPS 2017), Curran Associates, Red Hook, NY, pp. 6000-6010.

Pavlick, E. (2023), "Symbols and grounding in large language models", in *Philosophical Transactions of Royal Society A*, 381, 20220041. https://doi.org/10.1098/rsta.2022.0041

Watson, D. (2019), "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence", in *Minds and Machines*, 29, pp. 417–440. https://doi.org/10.1007/s11023-019-09506-6

Weizenbaum, J. (1966), "ELIZA. A computer program for the Study of natural language communication between man and machine", in *Communications of the ACM*, 9, pp. 36-45. doi:10.1145/365153.365168

Wykowska, A. (2024), *Intentional Stance Towards Humanoid Robots. Lessons Learned from Studies in Human-Robot Interaction*, Cham, Springer. https://doi.org/10.1007/978-3-031-65483-1