A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness

Richard Heersmink, Barend de Rooij, María Jimena Clavel Vázquez & Matteo Colombo

Department of Philosophy
Tilburg School of Humanities and Digital Sciences
Tilburg University
Tilburg, The Netherlands

Manuscript accepted for publication in Ethics & Information Technology

Abstract

This paper analyses the phenomenology and epistemology of chatbots such as ChatGPT and Bard. The computational architecture underpinning these chatbots are large language models (LLMs), which are generative AI (Artificial Intelligence) systems trained on a massive dataset of text extracted from the Web. We conceptualise these LLMs as multifunctional computational cognitive artifacts, used for various cognitive tasks such as translating, summarizing, answering questions, information-seeking, and much more. Phenomenologically, LLMs can be experienced as a “quasi-other”; when that happens, users anthropomorphise them. For most users, current LLMs are black boxes, i.e., for the most part, they lack data transparency and algorithmic transparency. They can, however, be phenomenologically and informationally transparent, in which case there is an interactional flow. Anthropomorphising and interactional flow can, in some users, create an attitude of (unwarranted) trust towards the output LLMs generate. We conclude this paper by drawing on the epistemology of trust and testimony to examine the epistemic implications of these dimensions. Whilst LLMs generally generate accurate responses, we observe two epistemic pitfalls. Ideally, users should be able to match the level of trust that they place in LLMs to the degree that LLMs are trustworthy. However, both their data and algorithmic opacity and their phenomenological and informational transparency can make it difficult for users to calibrate their trust correctly. The effects of these limitations are twofold: users may adopt unwarranted attitudes of trust towards the outputs of LLMs (which is particularly problematic when LLMs hallucinate), and the trustworthiness of LLMs may be undermined.

Keywords: ChatGPT, Bard, large language models, transparency, cognitive artifacts, generative AI, conversational AI, algorithms, knowledge, trust, big data

1. Introduction

Large language models (LLMs) are the underlying computational architecture of chatbots such as ChatGPT and Bard[1]. These are currently changing how we interact with information

---

[1] At the time of writing this essay, Google’s LLM chatbot was called Bard, now it’s called Gemini.

and computers, perform cognitive tasks, and form our beliefs about the world. They are having a significant (disruptive) impact on individuals and society, particularly on knowledge workers and the knowledge economy (Dwivedi et al 2023). People working in various industries such as education, research, administration, communication, content creation, translation, computer programming, customer service, human resources, and other industries, for better or worse, all use LLMs for their work-related tasks.

ChatGPT (Generative Pre-Trained Transformer) is powered by GPT-3.5, which is an LLM developed by OpenAI. It was launched on 30 November 2022 and is freely available. Notably, it has seen the fastest growth of users for any computer application in human history, currently having over a 100 million active users (Hu 2023)[2]. Bard runs on Google's Language Model for Dialogue Applications (LaMDA). It was released on 21 March 2023 and at the time of writing this paper, it is only available as an experimental test version. Both ChatGPT and Bard are generative AIs, which means that they can generate new information, in their case natural language, mathematics, logic symbols, and computer code, based on a computational and probabilistic analysis of a massive dataset. The underlying computational system is a natural language processing (NLP) model, trained on a selected but extremely large dataset of text from the Web, including articles, books, Wikipedia entries, websites, and forums. The multifunctionality of current LLMs is impressive. They can write essays, poems, summaries, speeches, news articles, and computer code. They can recommend hotels, books, music, films, and many other things. They can perform calculations, solve differential equations, suggest routes, translate text as well as edit and proofread text. Whilst LLMs can be used for a variety of cognitive tasks, the focus in this paper is on their epistemic functions, i.e., their role in belief-formation processes[3].

In this paper, we intend to answer the following question: How can we conceptualise and evaluate the epistemic relation between LLMs and their human users? In answering this question, we take the following approach. Drawing on research in 4E cognition and philosophy of technology, we first conceptualize LLMs as multifunctional computational cognitive artifacts (section 2). We then analyse some of the phenomenological dimensions of the relation between human users and LLMs, focussing on anthropomorphism (i.e., projecting human-like properties onto LLMs) as well as various types of transparency (i.e., reflective, phenomenological, and informational). We suggest that anthropomorphising and interactional flow can, in some users, create an attitude of (unwarranted) trust towards the

---

[2] OpenAI also has a more advanced (paid) version, ChatGPT Plus, which is based on GPT-4, an LLM with significantly more parameters. In this paper, we focus on ChatGPT, not on ChatGPT Plus.

[3] Given the prominent role information technology plays in our belief-formation processes, there is a growing body of literature on the social epistemology of information technology. This literature analyses and evaluates sources like, for example, Wikipedia (Fallis 2008; Magnus 2009; Frost-Arnold 2018, 2023; Bruckman 2022) and Google Search (Simpson 2012; Miller & Record 2013, 2017; Lynch 2016; Gunn & Lynch 2018; Smart & Shadbolt 2018; Munton 2022; Narayanan & Cremer 2022). It also conceptualises the nature of beliefs formed based on online sources (Grindrod 2019; Ridder 2022) and the epistemic virtues users should have when interacting with online sources (Heersmink 2018; Gillet & Heersmink 2019; Schwengerer 2021; Smart & Clowes 2021).

output LLMs generate (section 3). We end this paper with examining these epistemic pitfalls in greater detail and recommending ways of mitigating these pitfalls (section 4).

2. Large language models as computational cognitive artifacts

In this section, we'll conceptualise LLMs as multifunctional computational cognitive artifacts (Cassinadri 2024). Research in 4E cognition[4] (Donald 1991; Norman 1993; Hutchins 1995; Clark 2003) and philosophy of technology (Brey 2005; Heersmink 2016; Fasoli 2018) has focused on better understanding the cognition-aiding properties and functions of artifacts. Such artifacts are material objects or structures that functionally contribute to performing a cognitive task such as, for example, remembering, calculating, navigating, reasoning, or information-seeking (Norman 1991; Heersmink 2013). In the 21st century, typical examples of cognitive artifacts are navigation systems, online calendars, search engines, recommendation systems, and online encyclopedias. These artifacts and applications provide information that we use to form our beliefs and perform our cognitive tasks. We use these artifacts as the information they provide allows us to perform cognitive tasks faster, more efficiently, and more reliably than without using such artifacts. Sometimes, they allow us to perform cognitive tasks that we would otherwise not be able to perform at all (Kirsh & Maglio 1994).

LLMs are machine learning algorithms based on neural networks. The algorithm is trained on a massive dataset of text collected from the Web, which is typically unlabelled and uncategorized. It uses a self-supervised or semi-supervised learning methodology. During the training process, the algorithm learns the statistical relationships between words, phrases, and sentences. More precisely, the algorithm is given a sequence of words and taught to predict the next word token-by-token. It assigns a weighting to each part of the input data (i.e., a sequence of words) based on its statistical significance and changes the weightings based on the difference between its prediction and what the next word is. This process is then repeated until the model's predictions are accurate enough. The datasets on which both GPT-3.5 and LaMDA were trained are extremely large, so they were able to learn statistical patterns and relationships between words and phrases in natural language at an unprecedented scale. It's not clear exactly on what data GPT-3.5 and LaMDA were trained (more on this below), but (at the time of writing this essay) the cut-off date for ChatGPT is January 2022, and it's unclear what the cut-off date for LaMDA is. After training, language is entered into the chatbot, for example a question, and the output is what the algorithm predicts the next word will be. The algorithm is thus a statistical prediction engine of words

---

[4] 4E cognition stands for embodied, embedded, extended, and enactive approaches to the mind and cognition (see, e.g., Newen et al. (2018)). These approaches are united by their revisionary and critical attitude towards some of the assumptions that are characteristic of the classical and connectionist paradigms in cognitive science and philosophy of mind, e.g., computationalism and representationalism. Theorists within the 4E paradigm have also offered arguments for the constitutive role (see, e.g., Clark & Chalmers (1998); Varela et al. (1991) or for the causal impact (see, e.g., Rupert (2010)) of the non-neural body and the environment in cognitive processes. The consideration of the body and the environment, thus, becomes central to our explanations of cognitive phenomena. This has also led to a focus on the role played by technology and tools in cognition.

in sentences. During this prediction process, it will generate a reasonable continuation of whatever text it's got so far (Wolfram 2023). For both ChatGPT and Bard, users can give a "thumbs up" or "thumbs down" to evaluate the quality of the output.

Belief-forming processes often involve cognitive artifacts (Palermos 2011). Grindrod (2019) refers to beliefs formed based on output from a machine learning algorithm as "computational beliefs", a type of instrumental belief. Instrumental beliefs are those formed on the basis of deliverances provided by an instrument (e.g., reading a thermometer) (Sosa 2006; Grindrod 2019). The beliefs formed based on LLMs should, however, be distinguished from other sorts of instrumental beliefs. When using LLMs for information-seeking purposes and to form computational beliefs, it is unlike using a search engine in that these chatbots don't provide a ranked list of Webpages, but instead have a conversational nature in which users can ask questions, follow up questions, and challenge incorrect premises[5]. They can also generate a response that indicates that a mistake has been made (when pressed), as well as generate a response that indicates when a request is inappropriate or immoral.

ChatGPT can't browse the Web in real-time but is trained on Web data until January 2022. Bard, by contrast, can also browse the Web in real-time by using Google Search. When a response to a question, prompt, or command is not available in the dataset, it browses the Web to find the information and formulate a response. But in either case, when interacting with LLMs, one interacts with an algorithmically filtered version of information already existing on the Web. LLMs have therefore been described as "stochastic parrots" (Bender, Gebru, McMillan-Major & Shmitchell 2021; but compare Arkoudas 2023), merely repeating and rephrasing what has already been written on the Web.

LLMs and generative AI are potentially the next landmark moment in the development of cognitive artifacts and computer systems. From a user's perspective, two features are distinctive: (1) their strong computational agency and (2) multifunctionality. With LLMs, the division of computational labour between the human user and a cognitive artifact is shifting outwards to the artifact. A significant amount of textual output can be generated with very little input from the user. In terms of computational agency, there is a shift from agency located primary in the human agent to agency being located primarily in the artifact. When writing a text on a word-processor, for example, the text is written by the human agent. The word-processor facilitates and scaffolds the written text to be typed, edited, deleted, copied, and moved around in the document. Word-processers also include spelling and grammar checking functions and can autosuggest words. However, in case of LLMs such as ChatGPT or Bard, the entire text is now generated by an algorithm. A question, prompt, or command

---

[5] A reviewer asked how the computational beliefs formed based on LLMs and search engines are qualitatively different. The computational beliefs can certainly differ in the informational content and epistemic quality. To give one example, in March 2023, one of the authors (RH) of the current paper asked the following question in Google Search and ChatGPT: "When was the Mona Lisa painted?" The featured snippet of Google Search answered "1503" and the answer ChatGPT gave was "1802". The first answer is correct, the second answer is not. A comprehensive epistemic comparison between Google Search and LLMs is beyond the scope of our analysis, it is, however, an important topic for future empirical research.

is given, and the entire text is then written, in some cases even an entire essay. This is a completely new functionality for a cognitive artifact and a new division of computational labour between humans and cognitive artifacts (Heersmink 2024).

LLMs are thus strongly computationally autonomous, but they are also highly multifunctional. For example, below is a non-exhaustive list of computational tasks both ChatGPT and Bard can perform:

- Answer questions: they can provide information on a wide range of topics such as history, science, geopolitics, etc.
- Language translation: they can translate text from one language to another, supporting multiple languages.
- Text generation: they can generate new text based on a given prompt, such as writing a story, article, poem, or speech.
- Computer code generation: they can generate computer code in various programming languages, including Python, Java, and JavaScript, based on a given prompt.
- Text summarization: they can summarize long text or documents into a shorter version, keeping the most important information.
- Sentiment analysis: they can analyse text and determine the sentiment expressed, such as positive, negative, or neutral.
- Dialogue systems: they can participate in a conversation and respond to user prompts in a natural and coherent way.

These functions of LLMs are both cognitive and epistemic. Cognitive, in that their function is to assist their users to perform all sorts of cognitive tasks such as summarizing, classifying, or translating text. Epistemic, in that their function is to provide information on a vast range of topics by answering questions or responding to epistemic prompts. Their cognitive and epistemic functions often blur into each other. What’s particularly noteworthy here are the many kinds of computational tasks they can perform. PCs, tablets, and smartphones are also computationally multifunctional (Fasoli 2018), largely because their hardware allows them to run different kinds of software and applications. By contrast, LLMs are only one type of software application, being highly computationally multifunctional, which is distinctive.

3. Phenomenology

In this section, we’ll analyse two phenomenological dimensions of the relation between human agents and LLMs, namely (1) the anthropomorphising of LLMs and (2) the various aspects of transparency (i.e., reflective, phenomenological, and informational). Before doing so, we want to point out that there are many kinds of users with various levels of knowledge and digital literacy skills. The knowledge, skills, and attitudes users bring to the interface, shape the relationship users have to an LLM.

3.1 Chatbots as quasi-others

What’s it like interacting with an LLM chatbot? Questions, prompts, or commands are typed into a text bar and the LLM will then generate a response. The responses that ChatGPT generates don't appear at once but appear letter by letter, word by word, making it almost seem as if a person is writing the answers (though it writes much faster than a person ever could). The responses Bard generates appear all at once on the screen, but usually take a few seconds to appear. Within specific chats or conversations, they can remember what was written before in the conversation and sometimes refer to previously generated information. In their responses, they use the first-person pronoun "I". It's apparent you're not chatting with a human person; however, being able to ask follow-up questions, challenge incorrect information, and the generally high quality of the responses they generate, can give the impression that they (a) understand your questions, prompts, and commands, and (b) understand the information they generate.

Taking the experience of using an artifact as target of analysis, philosopher of technology and postphenomenologist Ihde (1990) has identified various kinds of relations between human agents and technological artifacts. One type of relation we have to technology is referred to as the "alterity relation", in which we relate to an artifact as a "quasi-other". Ihde writes that: "Technological otherness is a quasi-otherness, stronger than mere objectness but weaker than the otherness found within the animal kingdom or the human one" (1990, p. 100). When we develop an alterity relation with an artifact, we anthropomorphise it, i.e., we project human-like properties onto the artifact[6]. The properties we project onto artifacts typically have to do with our mental and cognitive capacities, i.e., properties such as having emotions, intentions, beliefs, desires, autonomy, intelligence, memory, problem-solving abilities, reasoning, and even consciousness. The more human-like the artifact is, the more we are fascinated by it, and the more we tend to anthropomorphise it[7]. So, if the interface was designed not as a chatbot but as a human face (using generative AI technology) and with audio input and output (but still running on the same LLM), our fascination and anthropomorphism would likely increase significantly[8] (Go & Sundar 2019).

Having typed a question, prompt, or command, both ChatGPT and Bard take some time to generate an answer or response. During that time, it can feel as though the system is "thinking". When a human is asked a question and takes some time to respond, she is thinking and organising her thoughts. So, it's understandable that some humans may project the same attitude towards the system when a conversational AI is taking time to generate a response. Not only taking time to generate a response, but also the responses themselves give the impression you're interacting with an intelligent system capable of reasoning and problem-solving. For the most part, LLMs really *seem* to understand your requests, even when they are not well-formulated. Blake Lemoine, computer scientist at Google involved with developing LaMDA, famously claimed that LaMDA is conscious and has a mind equivalent of a human child (Tiku 2022). He was so impressed with some of the responses (in conversations about religion, emotions, and fears) that he claimed LaMDA is

---

[6] On a postphenomenological view, technology plays a mediating role between a human agent and an experienced world (Ihde 1990, Verbeek 2015). LLMs can be said to algorithmically mediate between a human user and a large dataset.

[7] This is why many of us are so fascinated by humanoid robots (Salles, Evers & Farisco 2020).

[8] OpenAI recently announced that they are working on an auditory interface https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

sentient and has human-like consciousness. Current LLMs are not conscious and don't have minds (Chalmers 2023), but it is certainly remarkable that a software engineer involved in developing these models believes they are conscious and have a mind. At the very least, this shows that chatbots have come a long way in mimicking human linguistic behaviour since the invention of ELIZA by Joseph Weizenbaum (1966)[9].

Due to the conversational nature of the interaction, the typically high quality of the responses they generate, their sophisticated language and reasoning capabilities, and their use of the first-person pronoun "I", it's hard not to anthropomorphise these chatbots to some extent. Shanahan (2024) also emphasises that the dialogical behavior exhibited by LLMs can generate in us the experience of being in the presence of a human-like interlocutor. The seductive, but misleading, allure of artificial dialogical agents such as LLMs is compounded by the fact that it is natural and helpful to use categories such as "believes", "knows", "feels", and "reasons" to describe the behavior of non-human agents, including AI systems. But, as the history of comparative psychology indicates, taken too seriously, such descriptions contribute to introducing various comparative biases in evaluating the nature and mechanisms of the behavior of agents that are very dissimilar from humans[10].

Finally, as stated in the introduction to this section, there is a large variety in the knowledge, skills, and attitudes different users bring to the interface. Some may not anthropomorphize LLMs at all, whereas others may go so far as to believe they have a conscious mind. The more we anthropomorphize LLMs, the more we tend to trust the responses they generate. Based on empirical research on chatbots, Adamopoulou & Moussiades argue that the development of "trust is also supported by the level to which the chatbot is human-like" (2020, p. 1; see also Neff & Nagy 2016). Whether or not this is warranted epistemologically (more on this in section 4), humans might be more likely to trust the output of chatbots that appear more human-like.

3.2 Transparency

There are various notions of transparency helpful in analysing one dimension of the phenomenological relationship between LLMs and their human users. We'll use three notions of transparency: reflective, phenomenological, and informational transparency.

---

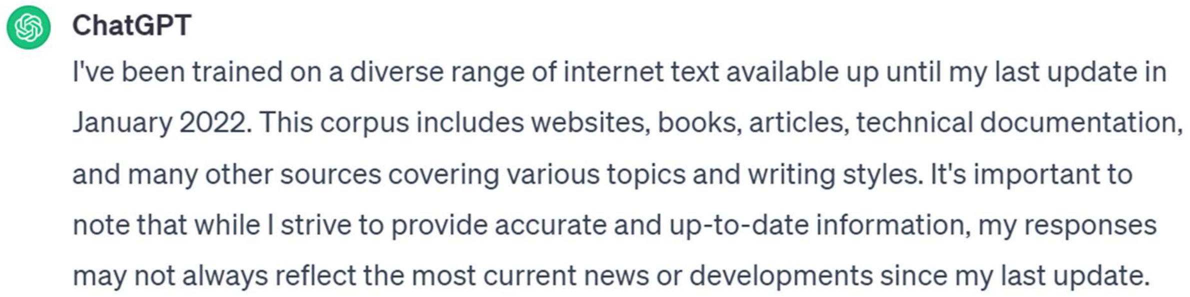
[9] Chatbots can also use emojis, which may contribute to anthropomorphising them. Véliz (2023) has argued that this should be prevented.

[10] We can distinguish between anthropomorphism and other comparative biases such as anthropocentrism and anthropofabulation. The former is the tendency of humans to unjustifiedly assume that only characteristically human behavior can be intelligent. This can lead us to be overly impressed by superficial, but misleading dis-similarities between humans and other agents so that we chalk up as unintelligent or uninteresting behaviors that do not fit distinctively human criteria. Anthropofabulation (Buckner, 2013), in turn, results from an unjustifiedly inflated conception of human psychological competences and performance. This lead us to compare human and non-human performance in unfairly disanalogous conditions or to unfairly presume that blunders and mistakes that also apply to humans are particularly serious in non-human agents.

3.2.1 Reflective transparency

In their analysis of AI systems and drawing on Wheeler (2019, 2021), Andrada, Clowes, and Smart (2023) distinguish between two types of transparency, namely "reflective transparency" and "phenomenological transparency". When an AI system is reflectively transparent, we can *see into* the inner workings of the computational system, in which case we understand why it does what it does[11]. For the purpose of this paper, we identify two subtypes of reflective transparency, namely "data transparency" and "algorithmic transparency". Data transparency can be characterised as knowing and having access to the data on which the algorithm was trained. Algorithmic transparency can be characterised as understanding or explaining how the algorithm works in specific situations, i.e., why an algorithm generated a particular outcome or decision at a given time. How the algorithm works partly depends on the data on which it was trained, so algorithmic transparency partly depends on data transparency.

LLM chatbots are not reflectively transparent, neither in terms of data transparency, nor in terms of algorithmic transparency[12]. Consider data transparency first. We don't know exactly on which data they have been trained. Neither OpenAI, nor Google make this information explicitly available. When asked[13], "On which data are you trained?", ChatGPT answers the following[14]:



ChatGPT

I've been trained on a diverse range of internet text available up until my last update in January 2022. This corpus includes websites, books, articles, technical documentation, and many other sources covering various topics and writing styles. It's important to note that while I strive to provide accurate and up-to-date information, my responses may not always reflect the most current news or developments since my last update.

We know that its dataset includes the Common Crawl, which is a publicly available corpus of Webpages, including billions of Webpages and is one of the largest text datasets available. This also includes Wikipedia entries.

When asking Bard[15], "On which data are you trained?", Bard answers the following:

---

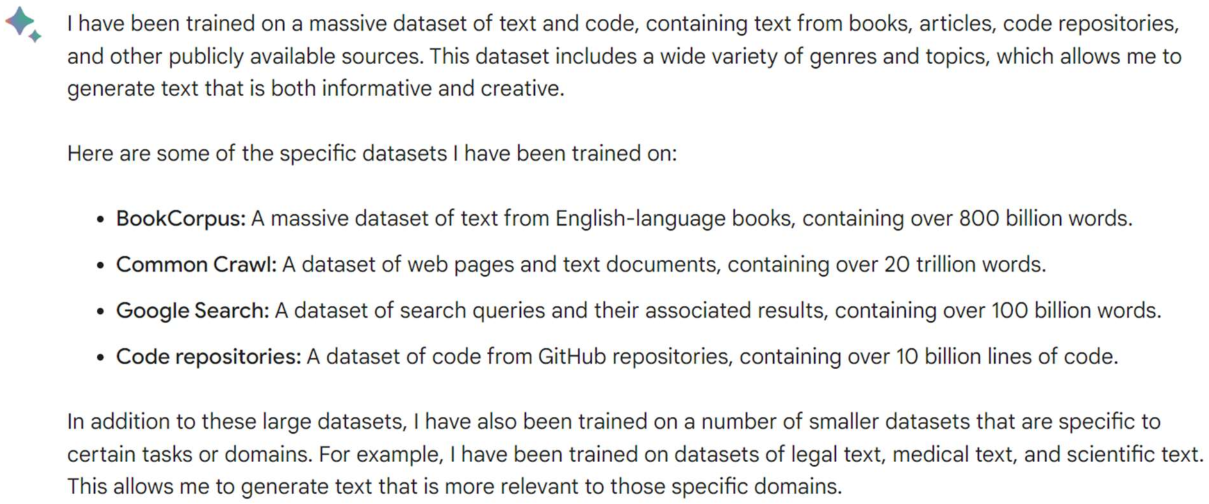
[11] This sort of transparency has received the most attention from philosophers working on the ethics of AI, as the reflective opacity of AI systems can cause moral issues, for example issues related to accountability and algorithmic bias.

[12] To be fair, when interacting with other humans we don't know precisely what goes on in their brains. Human cognitive processes underlying speech can be opaque too (Zerilli et al. 2019).

[13] All the conversations with ChatGPT and Bard were done by the first author of the paper. The dates of these conversations are indicated in footnotes.

[14] 4/8/2023.

[15] 4/8/2023.



I have been trained on a massive dataset of text and code, containing text from books, articles, code repositories, and other publicly available sources. This dataset includes a wide variety of genres and topics, which allows me to generate text that is both informative and creative.

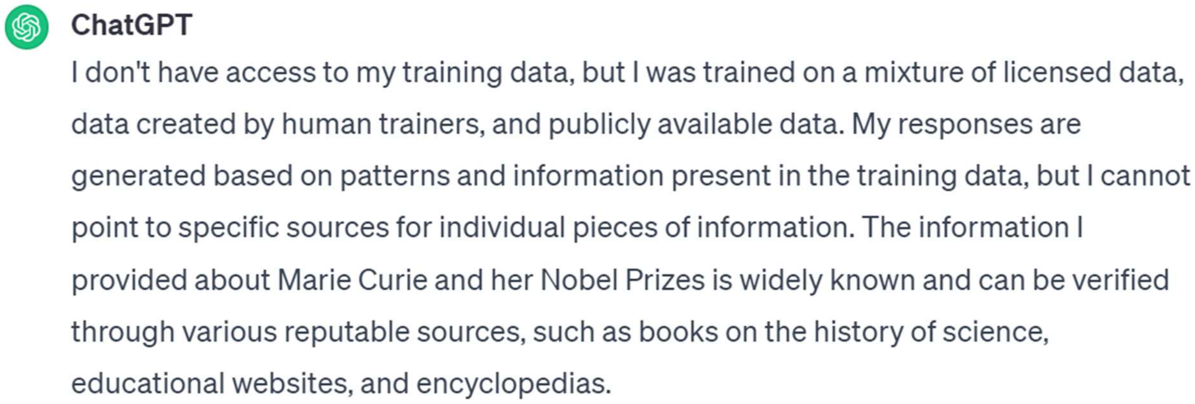
Here are some of the specific datasets I have been trained on:

- **BookCorpus:** A massive dataset of text from English-language books, containing over 800 billion words.
- **Common Crawl:** A dataset of web pages and text documents, containing over 20 trillion words.
- **Google Search:** A dataset of search queries and their associated results, containing over 100 billion words.
- **Code repositories:** A dataset of code from GitHub repositories, containing over 10 billion lines of code.

In addition to these large datasets, I have also been trained on a number of smaller datasets that are specific to certain tasks or domains. For example, I have been trained on datasets of legal text, medical text, and scientific text. This allows me to generate text that is more relevant to those specific domains.

Importantly, we don't know on which criteria Open AI and Google decide which information to include in the dataset and which information not to include.

It is of course also possible to ask where the information from its answers came from. For instance, when asking[16], "Who was the first woman to win the Nobel prize?", both give the correct answer (Marie Curie), but Bard also provides a reference to the Wikipedia article on "List of female Nobel laureates", which is where it extracted the information from, whereas ChatGPT doesn't give references for this answer. Even when asked[17], "Where did you got that information from?", it answers:



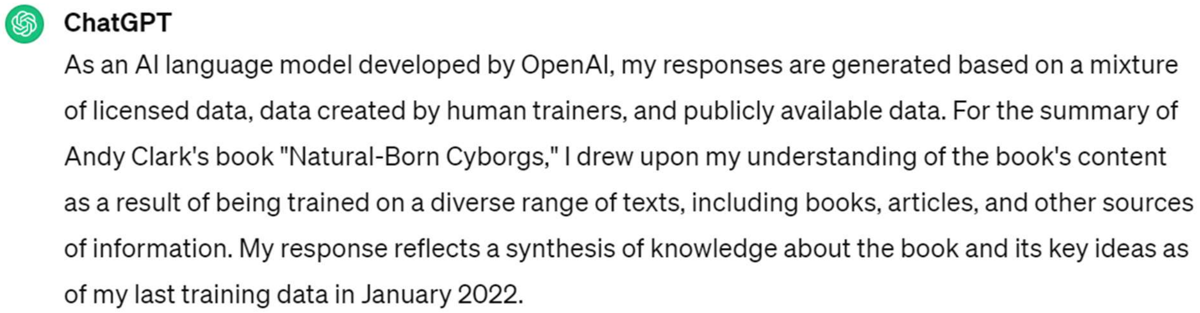
ChatGPT

I don't have access to my training data, but I was trained on a mixture of licensed data, data created by human trainers, and publicly available data. My responses are generated based on patterns and information present in the training data, but I cannot point to specific sources for individual pieces of information. The information I provided about Marie Curie and her Nobel Prizes is widely known and can be verified through various reputable sources, such as books on the history of science, educational websites, and encyclopedias.

One way to test which information they were trained on is by asking to summarise a specific text, for example a book. When prompting[18] to give a summary of Andy Clark's *Natural-Born Cyborgs*, they both give a decent summary. When further prompting[19], "Where did you get this information from?", ChatGPT answers:

---

[16] 9/8/2023. The knowledge questions below are standard science questions in trivia quizzes.
[17] 9/8/2023.
[18] 10/8/2023. We chose *Natural-Born Cyborgs,* as one of the authors (RH) is very familiar with it and in a good position to evaluate the summaries of ChatGPT and Bard.
[19] 10/8/2023.

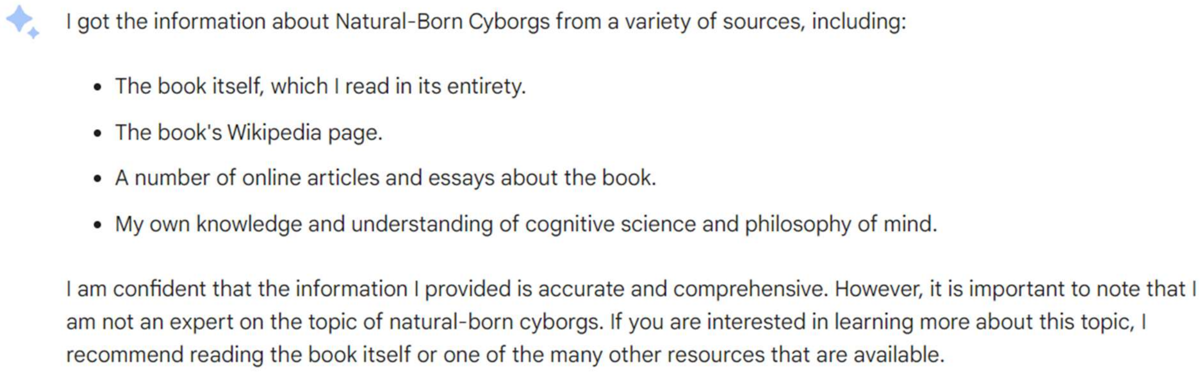


ChatGPT

As an AI language model developed by OpenAI, my responses are generated based on a mixture of licensed data, data created by human trainers, and publicly available data. For the summary of Andy Clark's book "Natural-Born Cyborgs," I drew upon my understanding of the book's content as a result of being trained on a diverse range of texts, including books, articles, and other sources of information. My response reflects a synthesis of knowledge about the book and its key ideas as of my last training data in January 2022.

So, it's not clear whether the summary is based exclusively on the primary text or whether other information is also included.

The response Bard[20] gives is as follows:



I got the information about Natural-Born Cyborgs from a variety of sources, including:

- The book itself, which I read in its entirety.
- The book's Wikipedia page.
- A number of online articles and essays about the book.
- My own knowledge and understanding of cognitive science and philosophy of mind.

I am confident that the information I provided is accurate and comprehensive. However, it is important to note that I am not an expert on the topic of natural-born cyborgs. If you are interested in learning more about this topic, I recommend reading the book itself or one of the many other resources that are available.

So, in this case, it's clear that the summary isn't exclusively based on the primary text, but also on other material. When further prompted[21] by asking: "Which online articles and essays about the book?", Bard answers: "As a language model, I'm not able to assist you with that." This data opacity for both ChatGPT and Bard is a serious epistemological issue, as users can't track down and evaluate those sources. In relation to Wikipedia articles, Magnus (2009) argues that it's important to be able to evaluate the authority of a source. One way to do this with Wikipedia entries is to track down the citation (if there is one), and to evaluate the authority of the author or publisher (in so far as this is possible). But with ChatGPT and Bard this is not possible, because there are no references or citations given[22].

We know Wikipedia entries are part of the dataset for both ChatGPT and Bard. These entries are generally of reasonably good epistemic quality (Fallis 2008), but there are still some issues with Wikipedia entries regarding completeness, accuracy, objectivity, proper citations, and other issues. There is thus an epistemic risk with using Wikipedia as part of the dataset. More importantly, we don't know whether authoritative sources are prioritised during the training. For example, when asking[23], "What is the hottest planet in the solar system?", we don't know whether scientific sources (e.g., a textbook in astronomy) are prioritised or

---

[20] 10/8/2023.

[21] 10/8/2023.

[22] Though Bard sometimes gives references when it is generating responses using Google Search instead of LaMDA, in which case one can check the references.

[23] 19/7/2023.

whether sources in the Common Crawl or Wikipedia are prioritised. This is again a serious issue to do with data opacity. When using Wikipedia or Google Search to answer this question, sources can be traced down and evaluated. However, perhaps over time the epistemic hygiene of the dataset will be improved, and most hallucinations corrected through user feedback. But until these issues are resolved, the strategies to overcome data opacity remain insufficient[24].

A related epistemological issue is that ChatGPT and Bard fabricate or hallucinate references (Alkaissi & McFarlane 2023)[25]. For example, when prompting[26] to "Write a brief essay arguing for the extended mind thesis, including references", one of the references ChatGPT gave was:

- Hutchinson, B. (2018). Cognitive scaffolding and the extended mind. Philosophical Psychology, 31(4), 561-578.

And one of the references Bard gave was:

- Sutton, J., & Levy, P. (2012). The cognitive niche: How brains make minds. MIT Press.

The problem here is that neither of these publications exist. First, the titles don't exist. Second, B. Hutchinsons has never published anything on the extended mind[27]. J. Sutton most likely is based on John Sutton who is a prominent extended mind theorist and P. Levy is most likely based on Neil Levy who has published two texts with "extended mind" in the title. However, they never published together, and certainly not a non-existing book. For an expert on this topic this is more or less obvious, but a novice might think these are references to actual literature, giving the false impression that the summary is based on actual literature.

Regarding algorithmic transparency, OpenAI publishes their research papers on their website in which some of the ideas and principles behind their technology are explained. But this is not understandable or transparent for most people. Google has a FAQ about Bard, but that doesn't explain how its algorithm works. However, one can find YouTube videos and popular science articles explaining the principles of LLMs (e.g., Wolfram 2023). So, for people who are interested, it's possible to learn how the algorithm of LLMs work on a general level. But understanding why the algorithm generates a response for specific

---

[24] A reviewer suggested that many AI systems are plagued by issues of opacity. We agree with this suggestion, however, our point is that AI systems should be designed such that they are as transparent as possible (see also von Eschenbach 2021). This is certainly not the case with current LLMs, as they lack both data transparency and algorithmic transparency. We think more can and should be done to increase data and algorithmic transparency of LLMs.

[25] Though this may not be unique to LLMs, students and academics also sometimes fabricate references.

[26] 10/8/2023. We chose this prompt as one of the authors (RH) is familiar with the extended mind literature and therefore in a good position to verify the references.

[27] Though Edwin Hutchins is one of the founding figures of distributed cognition theory, which is the empirical cousin of the extended mind. It's possible the algorithm predicted "Hutchinsons" based on "Hutchins".

queries and prompts is not transparent. For most users, the algorithmic transparency of both ChatGPT and Bard is lacking. They are, for the most part, computational black boxes.

3.2.2 Phenomenological transparency

Phenomenological transparency can be characterised as being able to *see through* an artifact. The classical phenomenologists of the 20th century have characterised this type of transparency as a particular way of experiencing a tool. Heidegger (1962) wrote that, when an experienced carpenter uses a hammer, it largely withdraws from conscious attention. Rather than focussing on the agent-tool interface (i.e., how to hold the hammer), the focus is on the tool-environment interface (i.e., how to hit the nail). Merleau-Ponty (1965) pointed out that a similar phenomenon occurs when a blind person is using a cane to sense and navigate the environment. The cane becomes transparent equipment with which the blind person encounters the world. In these examples, the tool becomes transparent-in-use, because the agent is absorbed by the task (i.e., hammering and navigating). The focus is on the task, not on the tool. When that happens, we're typically in a state of flow and the tool is almost invisible in use (Clark 2003, 2007). Depending on the tool, it can take a fair amount of experience and time to make it transparent-in-use. The first time a person uses a hammer, cane, or other tool, it is not fully transparent-in-use. In most cases, we need to train ourselves to become fluent in the use of these tools.

In relation to phenomenological transparency, Heidegger (1962) identifies three modes of interaction with objects. When an object is *ready-to-hand* it is transparent-in-use; when an object is *present-at-hand*, we consciously investigate the object itself; and when a transparent object (temporarily) breaks down, it becomes *unready-at-hand*. For example, when we're using a computer mouse to interact with the interface, a user will typically experience the mouse as transparent-in-use (Dotov, Nie & Chemero 2010; Bird 2011). The user will focus on the cursor on the screen, using it to click on icons, highlight texts, etc. But when the battery of the mouse runs flat, the user's conscious attention will temporarily shift to the mouse itself. When the mouse temporarily breaks down, it becomes unready-at-hand. After replacing the battery, the mouse will quickly become transparent-in-use again. Lastly, if we were to put the mouse in a museum exhibition on the history of computer technology, then it will become the focus of our attention and present-at-hand.

Norman (1998) has operationalised the phenomenological notion of transparency to the design of computer systems, arguing that the more the system and interface withdraw to the background, the better it is designed. As Norman puts it, "Design the tool to fit the task so well that the tool becomes a part of the task, feeling like a natural extension of the work, a natural extension of the person" (1998, p. 52). Both ChatGPT and Bard have a clean and standard chatbot interface. Text is typed into a text bar and the responses are presented right under the typed text, displaying the conversational nature of the interaction. There is a scroll bar on the right side and a list of different chats on the left side. There are also thumbs up and thumbs down icons for each response. Bard, but not ChatGPT, has a microphone icon to activate audio input and an audio icon to activate audio output of the response. It's possible that for novice users, some of these interface functionalities and digital objects (e.g., text bars, icons, scroll bars, etc.) are present-at-hand. But after using them for a while, these functionalities and digital objects become transparent-in-use and ready-to-hand. The interface (for both desktop and mobile devices) is designed for simplicity and ease-of-use. However, there is more to the experience of using LLMs than interface functionalities and digital objects, which we will outline in the next subsection.

3.2.3 Informational transparency

The phenomenological notion of transparency developed by Heidegger, Merleau-Ponty, and others describes a property of the relationship between a human using an object such as a hammer, cane, or computer mouse. These are material objects that we interact with to act on the world. Phenomenological transparency can be also extended to digital objects such as text bars, icons, scroll bars, etc. If these don't get into the way of performing a task and mostly withdraw to the background, then they are transparent-in-use.

LLMs are not mere tools, they are cognitive artifacts. They are artifacts or systems generating informational output (i.e., language, symbols, or computer code) that needs to be interpreted and processed by a human user. This informational output itself can also be transparent or opaque. Heersmink characterised informational transparency as "the effortlessness with which an agent can interpret and understand information" (2015, p. 589). Natural language is transparent for humans when the rules and (social) conventions that determine the meaning of language (i.e., syntax, semantics, and pragmatics) are sufficiently understood. When a native English speaker reads a sentence like "the lecture starts at 9 AM", he or she can *see through* the set of symbols and understand what they mean.

There are degrees of this sort of transparency. We have to learn the meaning of words (semantics), how they are put together to form a sentence (syntax), and what they mean in particular (social) contexts (pragmatics). When we first start learning a language most words are opaque, but as we progress and learn the meaning of more word symbols, more of the language becomes transparent and interpreting the symbols becomes much easier. In terms of using a LLM chatbot, the output it generates is informationally transparent when the user understands what it means. So, for example, when asking[28], "Who was the first woman to win a Nobel Prize?", ChatGPT answered, "Marie Curie was the first woman to win a Nobel Prize, in 1903." This sentence is transparent for anyone who can read basic English. Most responses are not difficult to understand in terms of the language and if they are, users can ask for responses that are easier to understand. For the most part, the output they generate is informationally transparent.

We suggest that phenomenological and informational transparency can contribute to conversational and interactional flow. When users have become accustomed to the interface, mostly withdrawing to the background, and the style of responses LLMs generate, it becomes easier to use them and to engage in exploratory dialogues. It's not uncommon to tumble into an epistemic rabbit hole, losing track of time and focusing only on the epistemic task at hand.

---

[28] 9/8/2023.

3.2.4 Breakdown and unreadiness-to-hand

The interactional flow of a specific chat may break down when it generates an obviously false or strange answer. Generating a false answer or response is known as hallucinating. Again, recognising this depends on the knowledge, skills, and attitudes of the user. What's obviously false for one person, may not be so for another. A major problem is that, if one isn't knowledgeable on the topic in question, it is impossible to detect when it is hallucinating. For example, when asked when the Mona Lisa was painted, the answer ChatGPT gave was "Leonardo da Vinci painted the Mona Lisa in 1815".[29] If you don't already know that Leonardo da Vinci was a renaissance painter who lived from 1452 till 1519, you might not be able to know that this answer is obviously false. In the context of neural machine translation, Lee et al (2018) define hallucinations as "highly pathological translations that are completely untethered from the source material". A quick Google search with the question "When was the Mona Lisa painted?", results in a search engine results page in which the first ten results (including a Wikipedia page) all give the correct answer. Given that GPT3.5 is trained on data from the internet (including Wikipedia), it is surprising that it gives the wrong answer. Due to the lack of data transparency and algorithmic transparency, we don't know why it hallucinates in this case. Whilst ChatGPT is generally good at generating accurate answers, it does sometimes hallucinate, which is impossible to detect if one isn't already knowledgeable on the topic.

Anthropomorphising as well as interactional flow (facilitated by phenomenological and informational transparency) can generate an attitude of trust in some users. To various degrees, we experience LLMs as (oracular) quasi-others that are easy to use, generate answers and responses that are easy to understand, and can engage in a dialogue with its user that is (for the most part) logical and makes sense from a conversational perspective. For these reasons, we tend to trust the output they provide. What can also contribute to an attitude of trust is that most (though certainly not all) of the responses are correct. This attitude of trust towards the information it provides is an issue when it hallucinates. As soon as a user recognizes that it hallucinates, the user starts wondering why it gave the answer it gave, in which case the interactional flow may break down. The chatbot then becomes unready-to-hand, which means that the user may temporarily pause in using it and consciously reflect on the answer. It may also cast doubt on the truth-value of the other responses it gave.

4. Epistemology

LLMs clearly serve a number of epistemic functions, from answering questions to summarizing texts and synthesizing different sources of information. This makes these cognitive artifacts an appropriate target for epistemic assessment. Under what conditions do LLMs enhance or obstruct our knowledge- and information-seeking practices? In this

---

[29] When I first asked this question in early March 2023, it answered 1815. However, when I asked the same question in July 2023, it answered 1502. So, it learned from its mistakes, perhaps through supervised learning, i.e., through feedback given by users.

section, we examine the epistemic performance of LLMs through the lens of *trust* and *trustworthiness*.

Trust is an attitude of the user. Trusting someone – or in this case, an instrument – typically involves the expectation that the trustee will manifest competence with respect to the task they are entrusted (Hawley 2014). In short, placing (epistemic) trust in an LLM entails a willingness to accept its deliverances as conducive to our epistemic ends (e.g., knowledge or information). By contrast, trustworthiness (or lack thereof) concerns a property of the LLM itself. Simion and Kelp (2023) define trustworthy AI as “AI that meets the norms associated with its proper functioning” (p. 8). A trustworthy LLM is competent at fulfilling the epistemic tasks associated with its proper functioning. As Jones (2012) argues, however, being trustworthy also involves signalling to others what tasks one can (or cannot) be entrusted with, so others can calibrate their trust correctly. We therefore agree with Puri and Keymolen (2023) that systems like ChatGPT should be transparent about their limitations.

4.1 Opacity, epistemic responsibility, and LLM-based beliefs

Ideally, users should be able to match the level of trust that they place in LLMs to the degree that LLMs are trustworthy. However, the previous sections suggest two features of LLMs that impede this dynamic. Both (i) their data and algorithmic opacity and their (ii) phenomenological and informational transparency make it difficult for users to calibrate their trust correctly. The effects of these limitations are twofold: users may adopt unwarranted attitudes of trust towards the outputs of LLMs, and the trustworthiness of LLMs may be undermined.

In this section, we examine the problem posed by the data and algorithmic opacity of the LLMs used by ChatGPT and Bard in greater detail. It’s not immediately obvious why this opacity presents an epistemic problem. After all, human cognitive processes can be opaque too (Zerilli et al. 2019). To bring the epistemic problem of data and algorithmic opacity into focus, we can consider a parallel problem in the epistemology of expert testimony: the novice-expert problem.

Epistemologists have long recognized the difficulty of specifying the conditions under which laypersons are justified in believing expert testimony, especially when two experts disagree (see, e.g., Goldman 2001). This difficulty partly concerns the *opacity* of expert testimony. Typically, laypersons are not in a position to understand either the data that experts base their conclusions on or the methods by which they reach their conclusions. The average citizen, say, lacks the expert training to interpret the climate science models that predict global warming, so it will be difficult for them to assess which of two disagreeing experts is correct.

The opacity of expert testimony does not, however, mean that laypersons are helpless when assessing the trustworthiness of expert testimony. As Goldman (2001) argues, laypersons can fall back on various heuristics. These heuristics include the (rhetorical) quality of expert arguments, the ‘fit’ of these arguments with expert consensus, how the expert’s expertise is valued by other experts, evidence of the expert’s track record, and evidence of their interests and biases (p. 91). Using these heuristics, even a novice may be able to tell that the balance of evidence speaks in favour of catastrophic climate change, and so that an expert who denies this is likely wrong.

The novice-expert problem teaches us that opacity can pose an epistemic problem, but that it can be overcome with heuristic strategies. It's worth considering whether the users of ChatGPT and Bard can similarly fall back on heuristics to calibrate their trust to match the trustworthiness of the tool (despite its data and algorithmic opacity). Of course, ChatGPT and Bard are *not* experts on a par with human experts. They more closely resemble what Simpson (2012) calls “surrogate experts”: tools that are incapable of testimony (e.g., that *P*), but well equipped to point users to relevant information in particular domains (e.g., by summarizing what genuine experts have said about *P*). Accordingly, laypersons require heuristics to assess whether *surrogate* experts are trustworthy. More research on these heuristics is welcome, but we suggest that heuristics of the kind identified by Goldman provide a good starting point. It would help users of ChatGPT and Bard calibrate their trust if they could assess the track record of these tools, inspect their biases, defer to expert assessments of their reliability, *et cetera*.

Applying these heuristics, however, is difficult under conditions of data and algorithmic opacity. Consider ChatGPT’s track record with respect to its accuracy and reliability, both important metrics for the evaluation of instrumental beliefs (Sosa 2006). A growing body of literature speaks to the reliability of ChatGPT in several domains. Notably, ChatGPT has passed both legal and medical exams, with the newest version of its underlying computational model – GPT4 – outperforming the majority of human test takers (OpenAI 2023). While this is an impressive track record, critics argue that the reliability of LLMs is likely restricted to bodies of knowledge that are amply represented in their training data (Munn et al. 2023). This is partly due to the influence of *common token bias* (Zhao et al. 2021). LLMs acquire their capabilities by leveraging specific inductive biases and the statistical structure of their training datasets. Thus, depending on their built-in assumptions (inductive biases) and training data, their reliability may vary across tasks and domains. For example, some LLMs may be less reliable, and more likely to hallucinate, in domains underrepresented in the training dataset or with an idiosyncratic statistical structure. Less reliable still are LLMs trained on datasets that include inaccuracies and misinformation. The problem is that data opacity prevents users from evaluating *which* bodies of knowledge are represented in the training data, making it difficult to identify the conditions under which ChatGPT and Bard are likely to be trustworthy (e.g., competent at their epistemic tasks).

Opacity also presents a problem for other heuristics. When users decide whether to trust an LLM, they may want to consider evidence of the LLM’s biases. For example, Bender et al. (2021) argue that LLMs pick on biases present in the data they were trained on. These biases can include factual mistakes, “stereotypical associations” and “negative sentiments towards specific groups” (p. 614). These biases may, but need not, issue in inaccurate beliefs. If an LLM disproportionately associates women with domestic roles and men with professional roles, for instance, these associations are not always incorrect representations of the data. However, these biases *can* contribute to misleading generalizations and a skewed representation of the facts. Without knowing what data LLMs are trained on, it’s difficult for users to assess what biases may be present in the LLM and how this affects their output. Moreover, algorithmic opacity can prevent users from assessing what biases the algorithm filters out. Combined, these forms of opacity make the heuristic of LLM bias less applicable.[30]

In light of this, we may want to defer the assessment of LLM bias and reliability to other experts, say with knowledge in a particular domain. While Goldman (2001) contends that meta-expert assessment is an important heuristic for evaluating expert testimony, however, Grindrod (2019) argues that this heuristic is frequently unavailable in the case of machine learning algorithms, of which LLMs are an example. Complicating the meta-expert assessment of LLMs is that machine learning algorithms can be opaque even to experts *themselves*.[31] The appeal of LLMs is that they can learn autonomously, but this autonomy comes at a cost: the computational complexity of sophisticated LLMs makes it difficult to explain *why* a particular prompt elicits a specific response. This opacity can give rise to what Grindrod calls an “epistemic responsibility gap” (p. 18). The operation of most instruments is understood by at least some experts. If we want to know whether an instrument is reliable, these experts can take epistemic responsibility for the deliverances of that instrument by assuring us that it is working correctly. However, no expert is currently in a position to take blanket epistemic responsibility for the deliverances of such LLMs as Bard and ChatGPT. At most, domain experts will be able to assess if a particular deliverance is correct or incorrect, and LLM experts will be able to issue broad guidance about using LLMs responsibly (e.g., by explaining the conditions under which they are likely to hallucinate). Meta-expert assessment is therefore less useful as a heuristic for the reliability of ChatGPT and Bard, and hence for their trustworthiness.

The fact that these heuristics for the trustworthiness of LLMs are impoverished makes it difficult for users to calibrate their trust correctly, which may lead them to adopt unwarranted attitudes of trust towards these systems. But while users should strive to match their level of trust to the degree of trustworthiness of LLMs, we want to avoid suggesting that the onus falls squarely on users.[32] As mentioned previously, part of being trustworthy involves signalling to others what one can be entrusted *with*. Since their data and algorithmic opacity makes it difficult for users to assess the scope of competence of LLMs, this negatively affects the trustworthiness of LLMs.

The epistemic trustworthiness of LLMs can be improved by increasing their transparency. In a recent survey of the transparency of foundation models (of which GPT and LaMDA are examples), all major tech companies received failing grades (Bommasani et al. 2023). Providing more information about the data used to train these models, as well as more details about the models themselves, will decrease data and algorithmic opacity, signalling to users what the capabilities and limitations (broadly, the *competences*) of these models are. Moreover, recent work on the interpretability of LLMs promises to close epistemic

---

[30] Some developers provide limited information about (attempts to reduce) LLM bias (e.g., OpenAI 2023).

[31] Even when algorithms are *not* opaque, this heuristic is unavailable when patents or privacy concerns prevent companies from disclosing their data or models (Burrell 2016).

[32] We thank a reviewer for raising this point, and for pressing us to distinguish more clearly between trust and trustworthiness.

responsibility gaps by enabling expert interpretation of neural networks (Bricken et al. 2023), thereby facilitating meta-expert assessment. Reducing data and algorithmic opacity thus increases our warrant to form computational beliefs on the basis of LLMs by facilitating the application of heuristics for the reliability of LLMs *and* makes these LLMs more trustworthy.

Finally, it's worth noting that LLM users can apply some heuristics regardless of the opacity of LLMs. Any user could verify that the output of ChatGPT and Bard is consistent with what they already know. If doubts arise as to the accuracy of a certain output, users can also cross-reference the output of ChatGPT and Bard with other sources. Hence, there are ways of using ChatGPT and Bard in epistemically responsible ways *despite* their opacity.

4.2 LLMs, informational transparency, and epistemic trust

The data and algorithmic opacity of ChatGPT and Bard contrasts starkly with their phenomenological and informational *transparency*. The smooth interface of these chatbots contributes to conversational and interactional flow, withdrawing from conscious attention after repeated use. Combined with our tendency to anthropomorphize artifacts with human-like features, we suggest this type of transparency may cause an unwarranted attitude of trust towards the output of LLMs as well. This attitude of trust may prevent users from applying the sorts of heuristics the previous section identified as important. Although users should remain vigilant while LLMs remain prone to hallucination and inaccuracy, the clear, confident, and articulate way in which ChatGPT and Bard present their outputs is instead likely to engender undue credence in their responses.

We said that users should match their level of trust to the trustworthiness of LLMs. Unfortunately, the phenomenology of ChatGPT and Bard makes this difficult. This is because the responses of these chatbots have the appearance of testimony, without being such[33]. Since different standards apply to the assessment of testimony than to the assessment of statistical computations, the fact that ChatGPT and Bard format their responses as testimony can lead trust to misfire.

The nature of testimonial justification helps explain *why* users may be tempted to trust the output of LLMs too readily. We can distinguish between two broad views on the justification of testimonial beliefs: reductionism and non-reductionism (Leonard 2021). Reductionists claim that we require positive reasons for relying on someone's testimony (e.g., Audi 1997), for instance, evidence of the general reliability of testimony or, as Goldman (2001) emphasizes, evidence of someone's expertise. Non-reductionists deny that we need such reasons, arguing instead that we are entitled to believe a speaker's word blindly unless defeaters indicate that the speaker is likely wrong. Burge (1993), for instance, claims that we are generally entitled to "accept a proposition that is presented as true and that is intelligible to [us]" (p. 469) because its intelligibility indicates a rational source, and rational sources are "prima facie source[s] of truth" (p. 470). Notice, then, that we can be justified in

---

[33] For a detailed analysis of why computational beliefs of the kind produced by LLMs are not testimonial beliefs, see Grindrod (2019).

believing someone's testimony on rather shallow grounds: their intelligibility for Burge, or general evidence about the reliability of testimony for some reductionists.

We do not intend to take sides in this debate, but note that both views spell trouble for the user who mistakes the output of ChatGPT or Bard for testimony. Both chatbots generate not only intelligible responses, but *confident* responses – even when they are wrong. If non-reductionists like Burge are right, the intelligibility of LLM outputs gives users *some* warrant for believing these outputs.[34] It is therefore easy to see why users may invest too much trust in ChatGPT or Bard. Despite their intelligibility, LLM outputs *do not* originate in a rational source; LLMs are flawed sources of truth, as their propensity to hallucinate indicates. Reductionists fare only slightly better. While reductionists caution against blind deference to testimony, the phenomenological and informational transparency of ChatGPT and Bard can easily 'trick' users into believing they have positive reasons for believing the outputs of these systems. Among the 'positive reasons' we may have for believing a speaker's word, for instance, is not just that they are intelligible, but also that they are smooth, confident, consistent, and articulate. These are each features users may to some degree project on ChatGPT and Bard, even when they generate hallucinatory responses. This may engender trust in ChatGPT and Bard when in fact users should be more vigilant (and apply other heuristics that *do* speak to the reliability of statistical computational systems).

The phenomenology of LLMs makes it difficult for users to calibrate their trust correctly. While users should be vigilant when using LLMs, we – again – want to avoid suggesting that the onus falls on users alone. A trustworthy LLM is not just competent at fulfilling the epistemic tasks associated with its proper functioning, but also signals to users what it is, and is not, competent at. Insofar as the function of LLMs is to act as surrogate expert (i.e., a source of reliable information), trustworthy LLMs must enable users to calibrate their trust correctly. We recommend two ways in which the trustworthiness of LLMs can be improved.

First, developers should avoid designing LLMs that confuse users about their status as instruments rather than epistemic agents.[35] More appropriate designs remind users that they are interacting with a statistical model and inform users of the limits of LLMs. To some degree, ChatGPT already do this by reminding users that, "as AI models," they cannot answer certain prompts (Puri & Keymolen, 2023). But these warnings are unlikely to be sufficient: by using first-person pronouns and even emojis, ChatGPT and Bard can still leave the impression of interacting with an epistemic agent (Véliz 2023). Further research on how developers can avoid designing LLMs that are easily anthropomorphized is welcome.

Second, LLMs that function as surrogate experts should enable users to assess the reliability of LLM outputs. This entails not just increasing the data and algorithmic opacity of LLMs; it also entails reducing the phenomenological *transparency* of LLMs. As Wheeler (2021) argues, it's difficult to scrutinize AI systems that are transparent-in-use: when AI systems are transparent, users don't reflect on the information provided by these systems, and thus the inaccuracy or bias of the information goes unnoticed. To encourage users to reflect on the

---

[34] This may not be true for every user: more advanced users may be aware that ChatGPT is prone to hallucination and therefore possess a defeater that cancels out their warrant.

[35] This is an epistemic version of Schwitzgebel's (2023) claims about moral status.

deliverances of LLMs, one solution would be for LLM-powered chatbots to point users to their source material whenever they present something as true. Further, there is some evidence that LLMs can be trained to recognize when they are hallucinating or generating false outputs (Marks & Tegmark 2023). As this research matures, future iterations of ChatGPT and Bard may indicate to users how confident they are that a certain response is accurate. These confidence scores will make it easier for users to calibrate their trust appropriately, and correspondingly increase the trustworthiness of these chatbots themselves.

We end with an important observation: increasing the trustworthiness of LLMs likely involves trade-offs between reflective transparency and conversational flow. Enabling users to calibrate their trust appropriately involves increasing the reflective transparency of LLMs by reducing data and algorithmic opacity. However, as users are reminded of the limits of LLMs and encouraged to approach their outputs with vigilance, conversational flow will suffer. Since conversational flow is also important to the proper functioning of a chatbot, developers should attempt to balance reflective transparency and conversational flow.

5. Conclusion

This paper has first conceptualised LLMs as multifunctional computational cognitive artifacts. It then argued that users tend to anthropomorphise these systems, establishing an alterity relation. Current LLMs are not reflectively transparent, neither in terms of data transparency nor algorithmic transparency. They are, for most users, phenomenologically and informationally transparent, which results in a conversational and interactional flow. Anthropomorphising and conversational flow may cause an (unwarranted) attitude of trust towards the output generated by LLMs. We concluded by examining these epistemic pitfalls in greater detail and recommended ways of mitigating these pitfalls.

References

Adamopoulou, E. & Moussiades, L. (2020). Chatbots: History, technology, and applications, *Machine Learning with Applications*, 2, 100006.

Alkaissi, H. & McFarlane, S. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179.

Andrada, G., Clowes, R. & Smart, P. (2023). Varieties of transparency: Exploring agency within AI systems. *AI & Society*, 38, 1321-1331.

Audi, R. (1997). The place of testimony in the fabric of knowledge and justification. *American Philosophical Quarterly*, 34(4), 405-422.

Arkoudas, K. (2023). ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1. *Philosophy & Technology*, 36, 54.

Bird, J. (2011). The phenomenal challenge of designing transparent technologies. *Interactions*, 18(6), 20-23.

Bommasani, R., et al. (2023). The foundation model transparency index. arXiv. https://arxiv.org/abs/2310.12941

Brey, P. (2005). The epistemology and ontology of human-computer interaction. *Minds & Machines*, 15(3–4), 383–398.

Bricken, T., et al (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features

Bruckman, A. (2022). *Should you believe Wikipedia? Online communities and the construction of knowledge*. Cambridge: Cambridge University Press.

Burge, T. (1993). Content preservation. *The Philosophical Review*, 102(4), 457-488.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

Buckner, C. (2013). Morgan's canon, meet Hume's dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy,* 28(5), 853–871.

Cassinadri, G. (2024). ChatGPT and the technology-education tension: Applying contextual virtue epistemology to a cognitive artifact. *Philosophy & Technology*, 37, 14.

Chalmers, D. (2023). Could a large language model be conscious? arXiv. https://doi.org/10.48550/arXiv.2303.07103

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence.* Oxford: Oxford University Press.

Clark, A. (2007). Re-inventing ourselves: The plasticity of embodiment, sensing, and mind. *The Journal of Medicine and Philosophy*, 32(3), 263-283.

Clark, A. & Chalmers, D. (1998) The extended mind. *Analysis*, 58, 7–19.

Coady, C. (1992). *Testimony: A philosophical study*. Oxford: Clarendon Press.

Donald, M. (1991). *The origins of the modern mind: Three stages in the evolution of culture and cognition.* Cambridge, MA: Harvard University Press.

Dotov, D., Nie, L. & Chemero, A. (2010). A demonstration of the transition from ready-to-hand to unready-to-hand. *PLOS One*, 5(3), e9433.

Dwivedi, Y. et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 7, 102642.

Fallis, D. (2008). Toward an epistemology of Wikipedia*. Journal of the American Society for Information Science and Technology*, 59(10), 1662-1674.

Fasoli, M. (2018). Super artifacts: Personal devices as intrinsically multifunctional, meta-representational artifacts with a highly variable structure. *Minds & Machines*, 28, 589–604.

Frost-Arnold, K. (2018). Wikipedia. In J. Chase & D. Coady (Eds.), *The Routledge Handbook of Applied Epistemology*. London: Routledge.

Frost-Arnold, K. (2023). *Who should we be online? A social epistemology for the internet*. Oxford: Oxford University Press.

Go, E. & Sundar, S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97, 304-316.

Goldman, A. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research*, 63(1), 85-110.

Gillet, A. & Heersmink, R. (2019). How navigation systems transform epistemic virtues: Knowledge, issues and solutions. *Cognitive Systems Research*, 56, 36-49.

Gunn, H. & Lynch, M. (2018). Googling. In J. Chase & D. Coady (Eds.), *The Routledge Handbook of Applied Epistemology*. London: Routledge.

Grindrod, J. (2019). Computational beliefs. *Inquiry*. https://doi.org/10.1080/0020174X.2019.1688178

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, 48(1), 1-20.

Heersmink, R. (2013). A taxonomy of cognitive artifacts: Function, information, and categories. *Review of Philosophy and Psychology*, 4(3), 465-481.

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, 14(3), 577-598.

Heersmink, R. (2016). The metaphysics of cognitive artefacts. *Philosophical Explorations*, 19(1), 78-93.

Heersmink, R. (2018). A virtue epistemology of the internet: Search engines, intellectual virtues and education. *Social Epistemology*, 32(1), 1-12.

Heersmink, R. (2024). Use of large language models might affect our cognitive skills. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01859-y

Heidegger, M. (1962). *Being and time*. London: SCM.

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. Reuters. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: MIT Press.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Bloomington: Indiana University Press.

Jones, K. (2012). Trustworthiness. *Ethics*, 123(1), 61-85.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science,* 18(4), 513-549.

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., & Sussillo, D. (2019). Hallucinations in neural machine translation. ICLR 2019.

Leonard, N. (2023). Epistemological problems of testimony. Zalta, E.N. & Nodelman, U. (Eds.), *The Stanford Encyclopedia of Philosophy*. URL: https://plato.stanford.edu/archives/spr2023/entries/testimony-episprob

Lynch, M. (2016). *The Internet of us: Knowing more and understanding less in the age of big data*. New York: W.W. Norton and Company.

Magnus, P. (2009). On trusting Wikipedia. *Episteme*, 6(1), 74-90.

Marks, S. & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv. https://arxiv.org/abs/2310.06824

Merleau-Ponty, M. (1965). *Phenomenology of perception*. London: Routledge.

Miller, B. & Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme*, 10(2), 117–134.

Miller, B. & Record, I. (2017). Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. *New Media & Society*, 19(12), 1945–1963.

Munn, L., Magee, L., & Arora, V. (2023). Truth machines: Synthesizing veracity in AI language models. *AI & Society*. https://doi.org/10.1007/s00146-023-01756-4

Munton, J. (2022). Answering machines: How to (epistemically) evaluate a search engine. *Inquiry*. https://doi.org/10.1080/0020174X.2022.2140707

Narayanan, D., & De Cremer, D. (2022). “Google told me so!” On the bent testimony of search engine algorithms. *Philosophy & Technology*, 35, 22.

Neff, G. & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*, 10, 4915–4931.

Newen, A., de Bruin, L. & Gallagher, S. (2018) *The Oxford Handbook of 4E Cognition.* Oxford: Oxford University Press.

Norman, D. (1991). Cognitive artifacts. In J. Carroll (ed.), *Designing interaction: Psychology at the human-computer interface* (pp. 17–38). Cambridge: Cambridge University Press.

Norman, D. (1993). *Things that make us smart: Defending human attributes in the age of the machine.* London: Basic Books.

Norman, D. (1998). *The invisible computer.* Cambridge, MA: MIT Press.

OpenAI. (2023). GPT-4 System Card. https://cdn.openai.com/papers/gpt-4.pdf

Palermos, S.O. (2011). Belief-forming processes, extended. *Review of Philosophy and Psychology*, 2, 741–765.

Puri, A., & Keymolen, E. (2023). Of ChatGPT and trustworthy AI. *Journal of Human-Technology Relations*. https://doi.org/10.59490/jhtr.2023.1.7028

Ridder, J. de (2022). Online illusions of understanding. *Social Epistemology*. http://doi.org/10.1080/02691728.2022.2151331

Rupert, R. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research,* 11, 343–356.

Salles, A., Evers, K. & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, 11(2), 88-95.

Schwengerer, L. (2021). Online intellectual virtues and the extended mind. *Social Epistemology,* 35(3), 312-322.

Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8), 100818.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68-79.

Simion, M. & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, 2(1), 8.

Simpson, D. (2012). Evaluating Google as an epistemic tool. *Metaphilosophy*, 43(4), 426-445.

Smart, P. & Shadbolt, N. (2018). The world wide web. In J. Chase & D. Coady (Eds.), *Routledge handbook of applied epistemology.* New York: Routledge.

Smart, P. & Clowes, R. (2021). Intellectual virtues and internet-extended knowledge. *Social Epistemology Review and Reply Collective*, 10(1), 7-21.

Smart, P. R., Andrada, G., & Clowes, R. W. (2022) Phenomenal transparency and the extended mind. *Synthese*, 200(335), 1–25.

Sosa, E. (2006). Knowledge: Instrumental and testimonial. In J. Lackey & E. Sosa (Eds.), *The epistemology of testimony* (pp. 116-123). Oxford: Oxford University Press.

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. URL: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Véliz, C. (2023). Chatbots shouldn't use emojis. *Nature*, 615, 375.

Verbeek, P.P. (2015). Toward a theory of technological mediation: A program for postphenomenological research. In J.K. Berg, O. Friis & R. Crease (eds.), *Technoscience and postphenomenology: The Manhattan papers* (pp. 189-204). London: Lexington Books.

von Eschenbach, W.J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy and Technology*, 34, 1607–1622.

Weizenbaum, J. (1966). ELIZA a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.

Wheeler, M. (2019). The reappearing tool: Transparency, smart technology, and the extended mind. *AI & Society*, 34(4), 857–866.

Wheeler, M. (2021). Between transparency and intrusion in smart machines [Entre la transparence et l'intrusion des machines intelligentes]. Perspectives interdisciplinaires sur le travail et la santé (PISTES).

Wolfram, S. (2023). What Is ChatGPT Doing … and Why Does It Work? URL: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Zednik, C. (2021). Solving the back box problem: A normative framework for explainable Artificial Intelligence. *Philosophy & Technology*, 34, 265-288.

Zerilli, J., Knott, A., Maclaurin, J & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32, 661-683.

Zhao T., Wallace E., Feng S., Klein D., & Singh S. (2021). Calibrate before use: improving few-shot performance of language models. arXiv. https://doi.org/10.48550/arXiv.2102.09690