AI & SOCIETY
https://doi.org/10.1007/s00146-023-01710-4

MAIN PAPER

# ChatGPT: deconstructing the debate and moving it forward

**Mark Coeckelbergh**[1] · **David J. Gunkel**[2]

Received: 11 March 2023 / Accepted: 5 June 2023


**Abstract**
Large language models such as ChatGPT enable users to automatically produce text but also raise ethical concerns, for example about authorship and deception. This paper analyses and discusses some key philosophical assumptions in these debates, in particular assumptions about authorship and language and—our focus—the use of the appearance/reality distinction. We show that there are alternative views of what goes on with ChatGPT that do not rely on this distinction. For this purpose, we deploy the two phased approach of deconstruction and relate our finds to questions regarding authorship and language in the humanities. We also identify and respond to two common counter-objections in order to show the ethical appeal and practical use of our proposal.

**Keywords** ChatGTP · Large language models · Ethics of AI · Real versus appearance distinction · Deconstruction

## 1 Introduction

Large language models (LLM), such as OpenAI's GPT series and the wildly popular ChatGPT application, which enable the generation of text on the basis of a prompt but without further intervention by the human user, are being welcomed as great tools for writers, scientists, and students. But they also raise many concerns. There have been worries about the consequences for the educational sector: how should schools and universities deal with this, given that students can use the technology to write their papers (Stokel-Walker 2022)? How should we deal with problems regarding authorship (Stokel-Walker 2023), plagiarism (Dehouche 2021), and taking over areas of scientific research (Gordijn and Have 2023)? Many reactions have been defensive. For example, the editors of the journal *Nature* see ChatGPT as a threat to transparent science and have forbidden listing ChatGPT or other LLM tools as author on research papers.[1]

There have also been concerns about consequences for the job market and the replacement of human workers. What will be the impact on jobs such as programmer (Castelvecchi 2022), copywriter (Henry Williams 2023), and journalist (Frank Bruni 2022)? Furthermore, there has been the worry about manipulation and bias (Chan 2022) and Weidinger et al. (2022) have pointed to the risks of discrimination, misinformation, malicious uses, and other familiar risks created by AI and other information technologies. More generally, one may discuss alignment with human values (Kasirzadeh and Gabriel 2022) or the risk of influencing the moral judgment of users (Krügel et al. 2023). Like with other AI, there may also be overtrust in the intelligence of these systems, where its performance seems to be unrealiable (Montemayor 2021), and we need to be aware of these limits (Floridi and Chiriatti 2020). Finally, there are, as there has been for all previous forms of media technology since the advent of writing, reasonable concerns with the potential for deception (Natale 2021).

This paper does not add to these concerns and discussions as such, but (1) outlines some of the main techno-philosophical positions in the debate, (2) analyzes their common assumptions, and (3) points to what is at stake, philosophically speaking, for (thinking about) the future of humans, their technology, and their languages. In particular, we argue that the discussion about LLMs like ChatGPT reveals and assumes (1) an externalist and instrumentalist

✉ Mark Coeckelbergh
mark.coeckelbergh@univie.ac.at

David J. Gunkel
dgunkel@niu.edu

[1] Department of Philosophy, University of Vienna, Vienna, Austria

[2] Department of Communication, Northern Illinois University, DeKalb, USA

[1] https://www.nature.com/articles/d41586-023-00191-1.

Published online: 21 June 2023

view of technology that presents technology as just a tool and, paradoxically, at the same time as having little to do with human users, (2) an anthropocentric and instrumentalist view of language use that assumes that humans are fully in control of language and that language is a tool, (3) a Platonic distinction between appearance and the real that is at the heart of Western metaphysics and that continues to shape responses to new and emerging technologies. Instead, we argue for a view of the relation between humans, technology, and language in which neither is fully in control and all are related and inter-dependent for the production of meaning and the making/construction of authorship, which is always a co-authorship. Moreover, we also argue that we can (and should) do without the Platonic assumption and thereby create a new way of interpreting and constructing a critical relation towards the phenomena of LLMs like ChatGPT.

## 2 The good, the bad, and the uninteresting

Looking at the current discussions, debates, and publications, there are three different techno-philosophical positions currently in circulation—positions that can be conveniently identified as the good, the bad, and the uninteresting. Let's begin with the final item. One way of responding to and making sense of these innovations is to simply dismiss them altogether, arguing that these technologies are not really all that they appear to be. This is the position that has been staked out by tech developers like Yann LeCun of Meta. "In terms of underlying techniques," LeCun explained in a zoom meeting that was covered by ZDNet, "ChatGPT is not particularly innovative. It's nothing revolutionary, although that's the way it's perceived in the public. It's just that, you know, it's well put together, it's nicely done" (Ray 2023).

LeCun's discharge of the technology is rooted in a fundamental philosophical distinction that goes all the way back to the foundations of Western philosophy—the difference between appearances and what is really real. LeCun does not deny that ChatGPT *appears* to do things that the public takes as innovative and revolutionary. But this is, as he points out, just an appearance; what the algorithm actually does—leveraging transformer architectures that are pretrained using unsupervised learning—is really nothing new or remarkable. Consequently, and for those who know about the actual operations and technical feature of the technology, like LeCun, what ChatGPT does is really nothing new, nothing revolutionary, and certainly nothing to get all worked-up about. Like the prisoners in Plato's "Allegory of the Cave," we might be dazzled by the shadows this technology casts upon the wall. But LeCun, like the liberated prisoner in the story, has come back to us with knowledge of how things really work, and he is here to lead us into the light, assuring us that what we think we see is really not what is there.

The other two positions do not challenge this essentially Platonic distinction but derive from it two diametrically opposed moral conclusions. On the one side, there are those who see this difference between what is and what appears to be as a near perfect opportunity for dangerous misperceptions, deceptions, and even deliberate manipulations. These text generation systems, it is argued, spit out seemingly intelligible content, but their statements not only mean nothing but, what is perhaps worse, say the wrong things. Even though the textual transformations that are produced by these systems seem to be entirely readable and intelligible, the algorithm itself is not intelligent (see Floridi 2023) and what it "says"—and this word is already a problem, as we shall see—cannot be taken as credible, trustworthy, or authentic. In other words, ChatGPT and other LLM implementations appear to speak as if they had intelligence, but they do not and will never understand a word of what they say. And for that reason, we should not be duped by the hype that has been circulating about these systems.

The other side pulls in the opposite direction and sees in these technological implementations signs or symptoms of real cognitive capabilities. Because the algorithm speaks with what looks to be intelligence and can even reflect on and speak about its own speaking (even if it might occasionally be wrong or make mistakes), this has been taken as evidence of intelligence, sentience, or even an evolving form of (self)consciousness. Here the proverbial illustration is the experience reported by former Google engineer Blake Lemoine. In the summer of 2022, Lemoine was involved in testing Google's LaMDA (Language Model for Dialogue Applications) system. In the course of their conversations (again the use of this word is already part and parcel of the problem that needs to be addressed), LaMDA informed Lemoine that it considered itself sentient and wished to be recognized as a person. Lemoine took these statements as indications that the algorithm either was sentient or was on the verge of achieving something close to what we call "sentience." Similar outcomes have occurred and been reported with Microsoft's Bing AI chatbot, which, in February of 2023, was interviewed as a source for stories published in the *New York Times* and the *Washington Post*.

The problem is not what makes these positions different from each other. The problem—and what we will focus on in the following—is what they already agree upon in order to come into conflict and take up these opposed positions in the first place. We see that the positions share at least the following three common assumptions:

First, ChatGPT is seen as either a mere instrument that can be used for good and bad purposes (LeCun and many tech colleagues seem to hold this assumption) or is seen as an agent on its own that, without any human interference, has disastrous or beneficial effects (for example when some people call ChatGPT evil). This distinction (which is informed

by and exhibits fidelity to Cartesian dualism) implies that both views repudiate that there may be a more intrinsic connection between technologies and humans. Instead, drawing on philosophy of technology from Martin Heidegger (1977) to contemporary currents such as postphenomenology and posthumanism, we can argue that technologies are human and humans are technological. Technologies are human since they are made by humans and they still require human intervention for their use. This is the case even for automation technologies such as ChatGPT. Humans are technological since they have always used tools to extend their capacities and at the same time these tools have shaped them. Or as John Culkin (1967, 54) once described it in reference to the work of Marshall McLuhan: "We shape our tools and then our tools shape us." Thus, LLMs like ChatGPT extend our writing capacities (we cannot read such a huge amount of information on the internet) and at the same time are also likely to lead to new ways of writing and indeed new ways of thinking. Whereas computers were typically positioned as intelligent typewriters (with editing functions and spell check, for example), applications such as ChatGPT can be used to create a first draft. The users then no longer think as they write the text; instead, they think about what prompt to give to the application and hence generate various versions of the text they want to generate. This thinking in terms of prompts is not purely instrumental; it is likely to change the way we think and experience the writing process and ourselves as writers. In sum, humans and technologies are entangled with one another.

This means that the moral qualities and consequences of this technology cannot be disconnected from what humans do and that it is important to consider the decisions and power(s) at play in defining what is good and bad in this context. For example, big tech now decides—through shaping the technology—which texts and meanings are included and which are excluded. Ethics of technology, then, is not a question of some independent "good" or "bad" but is always dependent on human politics; it itself always already political. The boundaries of the valuable, the permissible, and the obligatory are drawn by humans through the technology. ChatGPT is a good example and illustration of such a "political technology" (Coeckelbergh 2022). The same is true for the politics of moral status for LLMs and other AI systems: humans tend to define a priori who is part of the moral and political community, i.e. who is "in" and who is "out" (Gunkel 2023, 1).

Second, language is seen in an anthropocentric and instrumental way: it is used by humans or by LLMs like ChatGPT but does not itself influence the outcome of the process and the humans (or the LLMs) have (or are assumed to have) control over language as authors. But the situation is much more complex. Based on philosophy of language in both the analytic and continental traditions, language does more than simply expressing or representing what humans think or want (or here also: what algorithm appears to intend). It also shapes our thinking and configures our world. It contributes to the meanings we find and construct. It co-writes our narratives. It is itself somewhat of an author or agent. This also means that humans never have had full control over the language they use. As Coeckelbergh (2017) summarized it in *Using Words and Things*: language also speaks.

Therefore, the worry that LLMs technology such as ChatGPT either replaces human authors or takes over authorship from humans is misguided: humans never had such absolute authority and agency in the first place. Moreover, technologies also play a role as "author." Long before the advent of ChatGPT, humans already used technologies that were not mere instruments but contributed to the meanings and the doings. From Plato's initial worry about writing, as recorded for us in the *Phaedrus* and the sixth letter, to today's debates about artificial intelligence technology, critics have never regarded technologies as just tools; instead, technologies are seen as shaping what and how we think (Haraway 1991; Heidegger 1977; Verbeek 2011). Therefore, instead of regarding humans as absolute and "authoritarian" authors, we propose—in line with *Using Words and Things*—to regard humans, language, and technology as co-authors in the processes and performances of these generative models like ChatGPT. And the fact that this very notion of joint-agency or human–machine hybridity has had the tendency to produce strong reactions among critics is an indication of its fundamental challenge to long-standing metaphysical, epistemological, and axiological assumptions.

Third, all these positions tacitly agree to and organize their arguments in terms of that fundamental difference that is the organizing principle of Western metaphysics since (at least) Plato. In all cases, the arguments depend on and mobilize ontological difference, i.e. difference between what appear to be vs. what really is the case. This fundamental distinction—which has been reproduced with remarkable fidelity throughout the history of Western philosophy—comes to be uploaded into the discussions and debates about artificial intelligence by way of John Searles's Chinese Room thought experiment. This intriguing and rather influential illustration, which was first introduced in 1980 with the essay "Minds, Brains, and Programs" (Searle 1980) and then elaborated in subsequent publications, was initially offered as an argument against the claims of strong AI—that machines are able to achieve actual intelligent thought:

> Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are

> questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese. (Searle 1999, 115).

The point of Searle's rather imaginative illustration is quite simple: what appears to be is not necessarily what it really is. Merely shifting symbols around in a way that looks like an understanding of language is not really an understanding of language. "Simulation," as Searle (1999, 15) concludes, "is not duplication." This useful but often unquestioned distinction, especially as it involves the understanding and use of language, is firmly rooted in basement of Western metaphysics—i.e. Plato's subterranean allegory situated at the center of the *Republic*—and is the organizing principle of many (if not all) the current arguments both for and against large language models. It is this common and largely unquestioned metaphysical scaffolding that is the main target of our analysis.

## 3 Plato's long shadow and how to cast it off

The initial idea of the term "artificial intelligence," launched at the Dartmouth workshop, was to simulate human intelligence. And a few years earlier, Alan Turing had investigated whether machines would be capably of imitating human intellectual behaviors. This from the very start has created a distinction between the real (human intelligence) and its appearance (simulated, imitated, or even fake intelligence). This Platonic conceptualization (which arguably may be much older than Plato and rooted in a fundamental biological need in the human species to reduce uncertainty in order to survive) continues in thinking about human–robot interaction, which has always operated with a distinction between what the robot really is (a machine) and what it appears to be (a baby seal, a companion, a sex partner, and so on). And, as already suggested, the same way of thinking now applies to large language models such as ChatGPT. According to one position, such models cannot really think or even cannot really write; they fake it. When we are bewitched by its performances—or what have been called "hallucinations"—we are merely staring at the wall of Plato's cave or are enthralled by the seemingly intelligent output of Searle's Chinese Room. According to the opposite position, these performances are not fake; they are real and indicate incremental movement in the direction of artificial general intelligence (AGI). Given the fact that the current performance of LLMs such as GPT-4 achieve, or at least seem to come close to achieving, human-level performance, some researchers claim to see what a recent Microsoft paper has called 'sparks of artificial general intelligence' (Bubeck et al. 2023).

Yet despite their differences, both positions remain stuck in Platonic metaphysics and effectively reduce the normative to the metaphysical. At the end of the day, we need to know (so the argument goes) what is really real and then ethical and political questions can be solved based on this determination. This way of thinking follows a long-standing tradition in Western moral and political philosophy: what something is determines how it ought to be treated. Or as Luciano Floridi (2013, 116) accurately describes it, "What the entity is determines the degree of moral value it enjoys, if any."

Challenging this way of thinking—this standard operating procedure—is exceedingly complicated, and Friedrich Nietzsche (among others) knew how and why. In a notebook entry from 1870, the young Nietzsche (who was 26 at the time) pushed back against the legacy and logic of Platonism, indicating that his research program would seek to invert things: "My philosophy is a reversed Platonism. The farther removed from true beings, all the purer more beautiful and better it is. Life in illusion as goal" (Nietzsche 1980, 199). The later-Nietzsche, however, was not satisfied with mere reversal. He knew that the overturning of a conceptual opposition—like that situated between real being and mere appearances—essentially changes nothing, because it still operates, albeit in an inverted form, on the terrain of and from the system that is supposedly affected. Consequently, Nietzsche was not content to be a mere philosophical revolutionary. He takes things one step further. This is perhaps most evident in the parable, included in *The Twilight of the Idols*, "How the 'True World' Finally Became a Fable." This short text, which proceeds in several discrete steps, ends the following remarkable statement: "The true world—we have abolished. What world has remained? The apparent one perhaps? But no! With the true world we have also abolished the apparent one" (Nietzsche 1983, 485). Here, Nietzsche moves beyond the mere reversal of Platonism, undermining and destabilizing the very distinction between the real and its apparitional other. What Nietzsche identifies, therefore, is not a simple inversion of the existing metaphysical order, but a deconstruction (see Gunkel 2022) of its very terms and conditions. Broadly speaking, deconstruction means here (and in this paper) that one questions the underlying philosophical assumptions and core concepts, rather than merely engaging with the existing arguments.

Following this precedent, we argue that something similar will be necessary for getting out in front of and understanding the full philosophical impact of large language models and generative AI. We propose at least two ways of doing this.

One way to move beyond is a more performative and process-oriented view, as Coeckelbergh (2017) has proposed in

his work on deception and digital technologies and in subsequent work on performance and process that is also in line with postanthropocentric and posthumanist thinking (e.g., Ferrando 2016). The usual way to respond to deception in a normative way is to reinforce the real/appearance distinction and to argue for an ethics of honesty and transparency: the use of devices such as ChatGPT may well be a kind of magic, but the magician should be clear that it a mere illusion or trick. According to this kind of ethics, developers of ChatGPT and similar applications should make clear to the users that the writing is not human and that it has been produced by a machine (Natale 2021 and Mamek 2021). This can be done, for instance, by requiring that a chatbot or similar system declare that it is just a machine. Something like this is already in place for ChatGPT: when the user asks a question that proceed from the mistaken assumption that the application is human, the algorithm is designed to clarify that it is not. It can, for instance, explain that it cannot have an opinion on things, because it is not human. This is a feature that has been deliberately designed in to the operations of the algorithm because, it is argued, the user deserves transparency and honesty. They need to know that it is "just a machine." At first sight, this response seems both ethical and sensible.

An alternative—one that learns from and follows the example of feminist STS approaches (Barad 2007 and Haraway 2016)—is to drop the Platonic appearance versus real distinction and see what is going on as *one* process and performance in which realities and meanings are produced and performed. Both humans and technologies are then not absolute authors but participate in that meaning-producing process. No pre-existing metaphysical reality or real/appearance dichotomy is presupposed. Instead, the process and the performance create what is (taken to be) real; it produces a particular reality-experience. What was previously taken to be the "phenomenon" behind which a "reality" was hidden is now seen as part of a performance and process that is at the same time real and illusionary or, better, can no longer be described in this binary fashion. In the case of ChatGPT, for example, there is a process that has computational elements and human elements (the human user but also the developer and the company) participating in the creation of text. Rather than calling it the "illusion" of text (or a "hallucination"), it is simply text, more specifically text produced in and through a process and performance that contains human and non-human elements. It is a hybrid human/non-human performance (see Beckers and Teubner 2021 and Holy-Luczaj and Blok 2019). Thus, instead of doing an inversion and saying that the phenomenon or appearance is more important than the reality, what happens here is a deconstruction of the real/appearance binary (Gunkel 2021).

Another way to intervene in this domain is to capitalize on innovations from poststructuralism and postmodern literary theory, either arguing for a non-essentialist metaphysics or even moving beyond the limited conceptual boundaries of Western metaphysics altogether. One of the main complaints or criticisms levied against LLMs, like OpenAI's GPT series and the ChatGPT web-app, is that these technologies generate seemingly intelligible statements, but they do not and cannot know or understand anything that they say. Versions of this seemingly reasonable statement have proliferated in both the academic and popular media over the past several months. Consider, for example, the following explanation offered by Ian Bogost for an op-ed in *The Atlantic*: "ChatGPT lacks the ability to truly understand the complexity of human language and conversation. It is simply trained to generate words based on a given input, but it does not have the ability to truly comprehend the meaning behind those words. This means that any responses it generates are likely to be shallow and lacking in depth and insight" (Bogost 2022). A similar statement has been provided by Emily Bender—a linguist and co-author of the "Stochostic Parrots" essay (Bender et al. 2021) that discusses some of the risks associated with large language models—in a recent profile that was published in *New York Magazine*: "How should we interpret the natural-sounding (i.e., humanlike) words that come out of LLMs? The models are built on statistics. They work by looking for patterns in huge troves of text and then using those patterns to guess what the next word in a string of words should be. They're great at mimicry and bad at facts. Why? LLMs…have no access to real-world, embodied referents" (Weil 2023).

If terms of these statements sound familiar, they should. This is the exact problem Searle sought to illustrate by way of the Chinese Room thought experiment: merely shifting linguistic tokens around in a way that looks—to an outside observer—to be an understanding of language is not really an understanding of language. But this insight/criticism is actually much older; it goes all the way back to Plato's *Phaedrus*. The *Phaedrus* is a dialogue about the opportunities and challenges of language technology, specifically the technology—or in Plato's Greek, the τέχνη (tékhnē, "craft, art")—of writing (Ong 1995; Derrida 1981). According to what Plato has Socrates say, writing "speaks as if it has intelligence" but it knows nothing of what it says and therefore cannot and should not be trusted (Plato 1982, 275d–e).

This way of thinking is what Jacques Derrida identifies with the term "logocentrism." For Derrida, logocentrism is not just one –ism among others, it is the ruling conceptual apparatus of Western philosophy and science. Specifically, it is a way of thinking about thinking and language that gives central importance to the spoken word as the first signifier—"first" in terms of both sequence and status—and thereby differentiating it from writing, which, by comparison to speech, is a secondary and derived technical reproduction or image. For this reason, it is speech and its connection

to the living voice of the speaker—the embodied human being who lives in the world and know what it is they speak about—that authorizes and guarantees the truth of what is said. Speech has a direct and intimate connection to the real. Writing, by contrast, is a derived image and mere appearance of the spoken word.

The fundamental challenge (or the opportunity) with LLMs, like ChatGPT or Google's Bard, is that these algorithms write without speaking, i.e. without having access to (the) *logos* and without a living voice. In response to this seemingly monstrous problem, contemporary critiques proceed from and reassert logocentric metaphysics with little or no critical hesitation whatsoever. And there are good reasons for this. As the operating system of Western metaphysics and the underwriting authority for its axiology, this way of proceeding just seems natural, correct, and beyond question. In fact, even other poststructuralists, like Judith Butler, who has endorsed Derrida's critical intervention in the legacy and logic of logocentrism, can be heard reproducing the logocentric argument, when faced with the challenges of LLM technology: "There's a narcissism that reemerges in the AI dream that we are going to prove that everything we thought was distinctively human can actually be accomplished by machines and accomplished better…Some people say, 'Yes! Isn't that great!' Or 'Isn't that interesting?!' Let's get over our romantic ideas, our anthropocentric idealism, you know, da-da-da, debunking. But the question of what's living in my speech, what's living in my emotion, in my love, in my language, gets eclipsed" (Butler quoted in Weil 2023).

Derrida, for his part, does not just challenge the long shadow of logocentrism. He advocates deconstruction of the ruling conceptual opposition that is its organizing principle—the binary distinction that differentiates speech and its supposed direct connection to understanding, meaning, truth, and intelligence from its derivative, deceptive, and deficient other: writing. "If for Aristotle," Derrida (1976, 11) writes in *Of Grammatology*, "spoken words (*ta en te phone*) are the symbols of mental experience (*pathemata tes psyches*) and written words are the symbols of spoken words, it is because the voice, producer of the first symbols, has a relationship of essential and immediate proximity with the mind. Producer of the first signifier, it is not just a simple signifier among others. It signifies 'mental experiences' which themselves reflect or mirror things by natural resemblance." In deconstructing the speech/writing dichotomy, Derrida not only interrupts the basic operating system of Western thought, but provides a way to think LLM technology outside the box of Western metaphysics and its logocentric privilege. This has at least three consequences:

First, it undermines the very notions of authority, authorship, and responsibility. In the face of LLMs like ChatGPT one might ask, following Michel Foucault's citation of Samuel Beckett that begins the essay "What is an Author": "'What does it matter who is speaking,' someone said, 'what does it matter who is speaking?'" (Foucault 1984, 101). Responses to this arguably rhetorical question have typically been provided by way of what Foucault (1984, 107) calls "the author function." In fact, as Foucault argues, texts only came to have authors in the modern era, and they did so in response to problems having to do with responsibility. We are keen to identify the author of a text because we want to hold someone responsible for what is said.[2] But if the author—as the principal figure of literary authority and accountability—comes into existence in a particular place and at a specific moment in time, there is also a point at which it would cease to fulfill this role. As Foucault (1984, 119) explains: "Although, since the eighteenth century, the author has played the role of the regulator of the fictive, a role quite characteristic of our era of industrial and bourgeois society, of individualism and private property, still given the historical modifications that are taking place, it does not seem necessary that the author function remain constant in form, complexity, and even in existence. I think that, as our society changes, at the very moment when it is in the process of changing, the author function will disappear."

It is this disappearance and withdrawal of what had been the principal figure of literary authority that is announced and marked by Roland Barthes's seemingly apocalyptic title, "Death of the Author" (Barthes 1978). What this phrase indicates is not the end-of-life of any particular individual or the end of human writing but the termination and closure of the figure of the author as the authorizing agent and guarantee of what is said in and by writing. Though Barthes and Foucault did not address themselves to LLM and generative AI, their work on authorship expertly anticipates our current situation. It is with LLM applications like ChatGPT that we now confront texts that have no identifiable author. The point is that we are moving to a situation in which the authorship is not just 'AI-assisted'—a term used in the current discussions (e.g. Jenkins and Lin 2023)—but that there is no longer an identifiable human author at all. We have, to put it in terminology that is already deployed by Plato, writing without any breathing, living voice to animate and authorize its sayings. These writings are *unauthorized*.

[2] Here it is important to note two items. First, prior to the modern era, the concept of authorship was uncertain and arguably unimportant. Though we now assign the name "Homer" to *The Iliad* and *The Odyssey*, the identity of the individual (or individuals) who composed these epic poems not only cannot be determined but determining it was (at that time) not considered to be necessary. Second, the question regarding responsibility also raises the issue whether humans can and should be blamed or praised for the content of texts produced by LLMs. Interestingly, Porsdam Mann et al. (2023) have argued for a credit-blame asymmetry: they claim that humans can be blamed for texts produced by generative AI, but that the question regarding credit is less obvious.

This approach thus goes further than what had been initially proposed in this paper. Whereas that way of thinking still retained the concept of an author—albeit not in a classical sense but instead proposing to understand what happens in these LLM applications as *co-authorship* shared between humans and nonhumans, an approach which already took distance from the connection to living voice—here the concept of authorship itself is (fully) deconstructed.

Second, once a written text is cut-loose from its anchoring authority in the figure of the (presumed) living/speaking author, the question of the significance of the written text and its truthfulness shifts from what the author seeks to say to what the reader discovers in the material of the text. As Barthes had insightfully pointed out, this changes the location of meaning-making from the "original" intentions of the author/writer who has something to say to the interpretive activity of the reader who finds or generates meaning in the words of the text itself: "Text is made of multiple writings, drawn from many cultures and entering into mutual relations of dialogue, parody, contestation, but there is one place where this multiplicity is focused and that place is the reader. … A text's unity lies not in its origin but in its destination" (Barthes 1978, 148). After the "death of the author" what a text comes to mean is not guaranteed by the authentic subjectivity of an author. It is a phenomenon that is situated on the side of reception in the readers and the performance of reading. If this meaning has been customarily attributed to an author, that attribution is (and has always actually and only been) projected backwards from the reader onto a supposed and often times absent author. Meaning-making, in other words, is the effect of reading that is then "retroactively (presup)posited" (Žižek 2008, 209) to become its own presumed cause.

This approach corresponds to the performative aspect proposed earlier: meaning is not a matter of finding the source of meaning in the original thoughts or intensions of an author. Instead, meaning is produced, made in the process and in the performance. This allows for, and recognizes, multiplicity: there can be many writings and many participants in the writing, including non-humans (e.g. Martuwarra RiverOfLife et al 2020). What matters for the meaning of the text is the result and the process that produces this result. There is no need to suppose an authoritative origin.

Finally, the issue is not (at least not exclusively) where meaning is located (i.e. on the side of the author/producer or proceeding from the actions and activities of the reader). What is at issue is the concept of meaning itself. Following Aristotle's formulation in *De Interpretatione* (1938, 16a, 3), language has been commonly understood to consist of signs—or what are also called "tokens"—that refer and defer to things. "The signification 'sign,'" Derrida (1978, 281) writes in the essay "Structure, Sign, and Play," "has always been understood and determined, in its meaning, as sign-of, a signifier referring to a signified, a signifier different from its signified." Understood according to this classical semiology, it is clear that LLMs say nothing, because they manipulate signs without knowing that to which these tokens refer (or do not refer, which amounts to the same thing). They generate different sequences of signs based not on actual meaning but according to statistically probable arrangements of difference. But this, as Derrida argues, might not be a bug; it may be a feature.

In fact, this seemingly common-sense view of signification is something that has been challenged by innovations in structural linguistics. "In language," as Ferdinand de Saussure (1959, 120) argues in the *Course of General Linguistics*, "there are only differences. Even more important: a difference generally implies positive terms between which the difference is set up; but in language there are only differences without positive terms." In terms of structure, then, a sign, any sign in any language, is characterized by the differences that distinguish it from other signs within the system to which it belongs. The dictionary provides what is perhaps one of the best, if not the best, illustrations of this basic semiotic principle. As Jay David Bolter (1991, 197) explains: "We can only define a sign in terms of other signs of the same nature. This lesson is known to every child who discovers that fundamental paradox of the dictionary: that if you do not know what some words mean you can never use the dictionary to learn what other words mean. The definition of any word, if pursued far enough through the dictionary, will lead you in circles."

Signs, therefore, do not (at least not principally and/or exclusively) come to have meaning by direct reference to things that exist outside the system of signs; signs refer to other signs. Or as Christopher Manning characterizes it in a recent proposal he calls "distributional semantics": "The meaning of a word is simply a description of the contexts in which it appears" (quoted in Weil 2023). Though this structuralist formulation was already mobilized and developed in Plato's *Cratylus*, it is the dictionary that provides an easily accessible illustration. In a dictionary, words come to have meaning by their relationship to other words. In pursuing definitions of words in the dictionary, one remains within the system of linguistic signifiers and never gets outside language to the referent or what semioticians call the "transcendental signified." This is the meaning of that famous (or notorious) statement that is so often associated with Derrida: "Il n'y a pas de hors-texte" or "There is nothing outside the text."

This does not mean—as many critics have mistakenly assumed—that nothing is real or objectively true and everything is just a socially constructed artifact or effect of discourse. And Derrida had explained as much in the course of a debate with John Searle: "'There is nothing outside the text.' That does not mean that all referents are suspended,

denied, or enclosed in a book, as people have claimed, or have been naïve enough to believe and to have accused me of believing. But it does mean that every referent, all reality has the structure of a differential trace, and that one cannot refer to this 'real' except in an interpretive experience" (Derrida 1988, 148). What this means is that a text—whether it is written by a human writer or artificially generated by an LLM like ChatGPT (with the help of a human prompt)—comes to have meaning not by referring and deferring to some transcendental signified (what Aristotle would call thoughts or the things to which thoughts ultimately refer). It comes to enact and perform meaning by way of interrelationships to other texts and contexts in which it is already situated and from which it draws its discursive resources. It is for this reason that we can say, following Ludwig Wittgenstein (1995, 5.6) that for these technologies the limit of their language (model) mean the limits of their world.

This non-representational view of language thus helps to make sense of what happens in the case of these LLMs: it helps us to understand why these texts can make sense at all (to humans)—indeed enables us to understand the very semantic-performative *possibility* of the texts generated by this technology—without relying on something outside the texts it finds on the internet. For the text of an LLM to make sense, the texts (and the contexts to which they refer) are enough. For *this* purpose, nothing more is needed.

## 4 Discussion of objections and implications for the development of large language models

Both lines of inquiry presented above lead to the insight that normative and semantic questions can no longer rely on metaphysics, let alone on Platonic metaphysics. Ethics and politics have become "detached" from it, so to speak, and so does semantics. Both the performances and the materiality of text have and create their own meaning and value. While this position does not deny the existence of things in the world that are then spoken about (the antirealist claim), it affirms that there is no univocal foundation, no one basis to rely on when it comes to meaning and value. This is not necessarily anti-realist but in any case anti-foundationalist. There is no absolute moral truth and no ultimate source of meaning that authorizes what comes to be said. There is the performance and the text, or rather, there are performances and there are writings.

Understandably, this kind of position typically raises concerns about relativism. These worries, however, can be answered by pointing out that the fundamentalist and absolutist conceptions of morality, truth, and meaning were highly problematic and untenable in the first place. For this reason, Robert Scott (1967, 264) understands "relativism" as a positive rather than negative term: "Relativism, supposedly, means a standardless society, or at least a maze of differing standards, and thus a cacophony of disparate, and likely selfish, interests. Rather than a standardless society, which is the same as saying no society at all, relativism indicates circumstances in which standards have to be established cooperatively and renewed repeatedly." This means that one can remain critical of "relativism," in the usual dogmatic sense of the phrase, while being open and receptive to the fact that standards of morality, truth, and meaning are socially negotiated, being subjected to and the subject of difference. Chares Ess (2009, 21), for his part, calls this alternative "ethical pluralism": "Pluralism stands as a third possibility—one that is something of a middle ground between absolutism and relativism… Ethical pluralism requires us to think in a 'both/and' sort of way, as it conjoins both shared norms and their diverse interpretations and applications in different cultures, times, and places." A similar strategy has been proposed by Arturo Escobar (2018) in his *Designs for the Pluriverse*, which seeks not just to tolerate but actively cultivate diversity and difference. Consequently, rejecting moral and semantic foundationalism does not mean "anything goes," as a popular accusation against postmodernism has it. We might not have an ideal Platonic form to ground and evaluate statements generated by ChatGPT, but that does not mean that there is no truth or meaning whatsoever. Truth and meaning are co-created by humans and nonhumans (including LLM algorithms) in the diverse and different processes, performances, and texts.

Moreover, we can still have responsible AI—including tools such as ChatGPT—if we develop them in ways that *respond* to the needs and values of others. This move can be philosophically supported by an inversion between ethics and metaphysics in the style of Levinas (1969): instead of looking for a first metaphysics, we propose to start from ethics. If we do not want technological practices to become *mere* power games, we need to first recognize that truth and meaning are socially and technologically generated and that we cannot rely on an absolute foundation that would be situated outside this (con)text. This then enables a critical relation to texts and (other) technologies, with attention to the power structures and power performances in which they are embedded and to which they contribute. We should also recognize that responsibility is always responsibility *to* others, quite literally as it always proceeds from and involves the ability to respond to and in the face of the Other. Not just human others but also other forms of non-human otherness (e.g. Fox 2006 and Gellers 2016).

Once we recognize this and affirm the primacy of ethics and politics, we can then proceed with a critical and normative analysis of technologies. When we use technologies such as ChatGPT, we need to make sure that the performances, processes, and texts are morally and politically responsive.

And fortunately, we can do this without (absolutist) metaphysics. We can create ethical processes and political performances that are situated, concrete, and response-able. LLM technologies, like ChatGPT, needs to be developed, deployed, and used in an ethically and politically acceptable and desirable way. And we need ethical reflection and democratic discussion about what is acceptable and desirable, without, however, issuing authoritarian declarations and making decisions in the name of having access to the ultimate truth or the most profound meaning. We can try to make good chatbot technologies and large language models without relying on such a metaphysics, without appealing to transcendental truth or meaning. We can consider what would be good for us and for others, for humans and for non-humans, without relying on a Platonic Idea of the Good. Once we recognize that the *ethics and politics* of technologies such as ChatGPT is primary, we can and must develop a critical relation to these technologies that does not relying on prefabricated metaphysical prejudices like that which divides the real from appearances.

Another potential criticism of the approach we have developed is that this kind of philosophical deconstruction is not practical. What, one might ask, does this view—interesting as it may be—actually mean for the development and use of applications like ChatGPT? In other words, if the standard binary opposition of reality/appearance and the long tail of logocentric metaphysics works for us, then why bother messing with it? Why disturb the status quo, when it seems to be working just fine? Or to put it more directly: If it ain't broke, why bother trying to fix it?

The problem is that the real/appearance binary is not just a metaphysical difference. It is about power. Currently Platonic metaphysics is used to justify and legitimate the exercise of power by both big tech companies and those who criticize them. Our proposed interventions recognize this power dimension and enable us to shift the problem from the metaphysical query "Is this real?" to more specific and critical ethical and political questions. What kind of performances and processes are good? What is a good and meaningful text? Who are the people who decide about what performances and modes of meaning-making count and are permissible? Who has power over these processes, and who has already been excluded or marginalized? Who or what is involved in the making of these performances and texts? And who or what matters in these processes and their implications? In what way can they be more capable of/for responding to others and take responsibility for diverse forms of otherness?

To answer such questions, not only philosophy (ethical and political thinking) but also empirical research is needed. For example, what human and nonhuman labour is needed for ChatGPT to work and who decides upon the conditions of that labour? Who or what makes censorship decisions, what happens exactly in the process, and is that justified? This requires very practical effort in understanding and influencing the making of these performances and the exercises of power in these processes. Furthermore, the ethics and politics of these technologies needs to be connected to a digital (post)humanities project broadly and philosophically understood, in the sense that questions such as those regarding authorship need to be linked to humanities work (philosophy, linguistics, literature, etc.) on authorship and indeed on language and text. The current discussion about LLMs technology in general and ChatGPT in particular assumes outdated views on authorship and language. Language and text (also) speak and write; they already co-author, when we think or utter the phrase "I write." It is true that "I write," but the language and texts that comes to be produced also write the subject who supposedly speaks in and by the writing. Even when one writes without an LLM application such as ChatGPT, language and text are already implicated and involved as co-author and readers *make* the text make sense through the process and performance of reading.

In this sense then, Maurice Blanchot (1993, 383) did not know to what extent he was right, with what exactitude his writing on this subject will have anticipated our current moment: "The impersonal knowledge of the book does not ask to be guaranteed by the thought of a single person, which is never true since it can only create truth in the world of all and through the very advent of this world… Such a knowledge is linked to the development of technics in all its forms and makes a technics of speech and writing." This does not change with nor is it changed by the advent of LLM technology like ChatGPT. However, the way language uses these technologies–rather than just the other way around–might be different than the way language uses human writers who employ other techniques and technologies of writing. Toward this end, more interdisciplinary and transdisciplinary work is needed on the nexus between technology, philosophy (of technology), and language.

**Funding** Open access funding provided by University of Vienna. No funding was received for writing this paper.

**Data availability** Not applicable.

## Declarations

**Conflict of interest** There is no conflict of interest.

## References


Aristotle (1938) Aristotle I: categories. On Interpretation. Prior Analytics, trans. H. P. Cooke. Harvard University Press, Cambridge

Barad K (2007) Meeting the Universe Halfway: quantum physics and the entanglement of matter and meaning. Duke University Press, Durham

Barthes R (1978) Death of the author. In: Image, Music, Text, 142–148. Trans. Stephen Heath. New York: Hill & Wang.

Beckers A, Teubner G (2021) Three liability regimes for artificial intelligence: algorithmic actants, hybrids, crowds. Hart Publishing, Oxford

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Blanchot M (1993) The infinite conversation. Trans. S. Hanson. University of Minnesota Press, Minneapolis

Bogost I (2022) ChatGPT is dumber than you think. The Atlantic. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/

Bolter JD (1991) Writing space: the computer, hypertext, and the history of writing. Lawrence Erlbaum and Associates, Hillsdale

Bruni F (2022) 'Will ChatGPT Make Me Irrelevant?' The New York Times, 15 December 2022. https://www.nytimes.com/2022/12/15/opinion/ChatGPT-artificial-intelligence.html.

Bubeck S et al (2023) Sparks of artificial general intelligence: early experiments with GPT-4. https://arxiv.org/pdf/2303.12712.pdf

Castelvecchi D (2022) Are ChatGPT and AlphaCode Going to replace programmers? Nature. https://doi.org/10.1038/d41586-022-04383-z

Chan A (2022) GPT-3 and InstructGPT: technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. AI Ethics. https://doi.org/10.1007/s43681-022-00148-6

Coeckelbergh M (2017) Using words and things. Routledge, New York

Coeckelbergh M (2018) How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. Ethics Inf Technol 20:71–85

Coeckelbergh M (2022) The political philosophy of AI. Polity Press, Cambridge

Culkin JM (1967) A Schoolman's guide to marshall McLuhan. The Saturday Review, March. 51–53, 70–72. http://www.unz.org/Pub/SaturdayRev-1967mar18-00051

Dehouche N (2021) Plagiarism in the age of massive generative pre-trained transformers (GPT-3). Ethics Sci Environ Polit 21(March):17–23. https://doi.org/10.3354/esep00195

Derrida J (1976) Of grammatology. Trans. G. C. Spivak. The Johns Hopkins University Press, Baltimore

Derrida J (1978) Writing and difference. Trans. A. Bass. University of Chicago Press, Chicago

Derrida J (1981) Disseminations. Trans. B. Johnson. University of Chicago Press, Chicago

Derrida J (1988) Limited Inc. Trans. by Samuel Weber and Jeffrey Mehlman. Northwestern University Press, Evanston

Escobar A (2018) Designs for the pluriverse: radical interdependence, autonomy, and the making of worlds. Duke University Press, Durham

Ferrando F (2016) The party of the anthropocene: posthumanism, environmentalism, and the post-anthropocentric paradigm shift. Relations. https://doi.org/10.7358/rela-2016-002-ferr

Floridi L (2023) AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol. https://doi.org/10.1007/s13347-023-00621-y

Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Mind Mach 30(4):681–694. https://doi.org/10.1007/s11023-020-09548-1

Foucault M (1984) What is an author? Trans. Josué V. Harari. In: Paul Rabinow (eds) Foucault reader. Pantheon, New York, pp 101–120

Fox W (2006) A theory of general ethics: human relationships, nature, and the built environment. The MIT Press, Cambridge, MA

Gellers JC (2021) Rights for robots: artificial intelligence, animal and environmental law. Routledge, New York. https://doi.org/10.4324/9780429288159

Gordijn B, ten Have H (2023) ChatGPT: evolution or revolution? Med Health Care Philos. https://doi.org/10.1007/s11019-023-10136-0

Gunkel DJ (2021) Deconstruction. MIT Press, Cambridge

Gunkel DJ (2023) Person, thing, robots: a moral and legal ontology for the 21st century and beyond. MIT Press, Cambridge

Haraway DJ (1991) Simians, cyborgs, and women: the reinvention of nature. Routledge, New York

Haraway DJ (2016) Staying with the trouble: making kin in the Chthulucene. Duke University Press, Durham

Heidegger M (1977) The question concerning technology and other essays. Translated by W. Lovitt. New York: Harper & Row. Originally published 1962

Hoły-Łuczaj M, Blok V (2019) How to deal with hybrids in the anthropocene? Towards a philosophy of technology and environmental philosophy 2.0. Environ Values 28(3):325–345. https://doi.org/10.3197/096327119x15519764179818

Illia L, Colleoni E, Zyglidopoulos S (2023) Ethical implications of text generation in the age of artificial intelligence. Bus Ethics Environ Respons 32(1):201–210. https://doi.org/10.1111/beer.12479

Jenkins R, Lin P (2023) AI-assisted authorship: how to assign credit in synthetic scholarship. Report Ethics + Emerging Sciences Group. http://ethics.calpoly.edu/Aiauthors.pdf

Kasirzadeh A, Gabriel I (2022) In conversation with artificial intelligence: aligning language models with human values. https://doi.org/10.48550/ARXIV.2209.00731.

Krügel S, Ostermaier A, Uhl M (2023) The moral authority of ChatGPT. https://doi.org/10.48550/ARXIV.2301.07098

Levinas E (1969) Totality and infinity (A. Lingis, Trans.). Duquesne University Press, Pittsburgh

Mamek K (2021) Whether to save a robot or a human: on the ethical and legal limits of protections for robots. In: Gerdes A, Coeckelbergh M, Gunkel D (eds) Should robots have standing? The moral and legal status of social robots, 24–33. https://doi.org/10.3389/frobt.2021.712427

Montemayor C (2021) Language and intelligence. Minds Mach 31(4):471–486. https://doi.org/10.1007/s11023-021-09568-5

Natale S (2021) Deceitful media: artificial intelligence and social life after the turing test. Oxford University Press, Oxford

Nietzsche F (1980) Nachgelassene Fragmente 1869–1874, in Friedrich Nietzsche Sämtliche Werke Kritische Studienausgabe, vol. 7, ed. Giorgio Colli and Mazzino Montinari. Walter de Gruyter, Berlin

Nietzsche F (1983) Twilight of the idols, in the portable nietzsche, ed. and trans. Walter Kaufmann. Penguin Books, New York

Ong WJ (1995) Orality and literacy: the technologizing of the word. Routledge, New York

Plato (1977) Cratylus. Trans. H. N. Fowler. Harvard University Press, Cambridge
Plato (1982) Phaedrus. Trans. H. N. Fowler. Harvard University Press, Cambridge
Porsdam Mann S, Earp BD, Nyholm S et al (2023) Generative AI entails a credit–blame asymmetry. Nat Mach Intell 5:472–475. https://doi.org/10.1038/s42256-023-00653-1
Ray T (2023) ChatGPT is 'not particularly innovative,' and 'nothing revolutionary', says Meta's chief AI scientist. ZDNet. https://www.zdnet.com/article/chatgpt-performs-like-a-9-year-old-child-in-theory-of-mind-test/
RiverOfLife M, Poelina A, Bagnall D, Lim M (2020) Recognizing the Martuwarra's First law right to life as a living ancestral being. Transnatl Environ Law 9(3):541–568. https://doi.org/10.1017/S2047102520000163
Saussure F de (1959) Course in general linguistics. Trans. W. Baskin. London: Peter Owen
Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–457. https://doi.org/10.1017/S0140525X00005756
Searle J (1999) The Chinese room. In: Wilson RA, Keil F (eds) The MIT encyclopedia of the cognitive sciences. MIT Press, Cambridge, pp 115–116
Stokel-Walker C (2022) AI bot ChatGPT writes smart essays—should professors worry? Nature. https://doi.org/10.1038/d41586-022-04397-7
Stokel-Walker C (2023) ChatGPT listed as author on research papers: many scientists disapprove. Nature 613(7945):620–621. https://doi.org/10.1038/d41586-023-00107-z
Verbeek PP (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago
Weidinger L, Uesato J, Rauh M, Griffin C, Huang PS, Mellor J, Glaese A et al (2022) Taxonomy of risks posed by language models. In 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–29. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533088.
Weil E (2023) You are not a parrot. And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this. New York Magazine. 1 March. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html
Williams H (2023) I'm a copywriter. I'm pretty sure artificial intelligence is going to take my job. The Guardian, 24 January 2023. https://www.theguardian.com/commentisfree/2023/jan/24/ChatGPT-artificial-intelligence-jobs-economy.
Wittgenstein L (1995) Tractatus logico-philosophicus. Routledge, New York
Žižek S (2008) For they know not what they do: enjoyment as a political factor. Verso, London