Bioethical Inquiry
https://doi.org/10.1007/s11673-025-10426-7

ORIGINAL RESEARCH

The Ethics of Speaking (of) AIs Through the Lens of Natural Language

Marcelo El Khouri Buzato

Received: 3 December 2024 / Accepted: 28 January 2025

Abstract This theoretical essay offers a critical exploration of the ethics involved in interacting with and talking about large language models (LLMs) of artificial intelligence (AI). The discussion is framed within philosophical post-humanist conceptualizations of the ethical agent, which is understood as a sociocognitive assemblage of human–machine interactions at various scales. The central argument asserts that the morality of texts generated by AI cannot be determined by extracting moral properties from the natural language used in the training corpus. There are inherent limits to how much analysing moral language can contribute to establishing a moral theory or a "language of mores" among humans. The essay also examines the ethical implications of current public discourse surrounding the capabilities of LLMs, as well as the ways in which LLM outputs personify the AI model itself. It is proposed that the ethics of LLMs should be approached as an ethics of translating informational patterns of linguistic symbols into multi-layered cultural meanings and vice versa. This includes addressing the opacity of the inner workings of these translations in the model, as well as the public relations practices of the creators. Ultimately, the discussion encourages rethinking the ethical agent as a human-machine sociocognitive hybrid, suggesting the need for a reassessment of what it means to be human and ethical in current AI ethics debates.

Keywords Artificial intelligence · Natural language · Ethics · Cognition · Postphenomenology

Introduction

This theoretical essay aims to make a specific contribution to current discussions about the ethics of artificial intelligence (AI). It focuses on two main issues: first, the limitations of large language models (LLMs) in relation to ethical reasoning based on natural language, and second, the risks involved in talking about LLMs in ways that imply their ability to manipulate linguistic symbols grants them entitlement to moral trust from human users. A post-humanist perspective is adopted, proposing that the morality of LLMs should not be seen as encapsulated within either the machine or the human. Instead, it should be approached as the result of translations and negotiations of meaning among hybrid (human-plus-machine) ethical agents at various scales. In other words, adopting a post-humanist view of the ethical agent helps researchers and practitioners confront the complexity of understanding AI ethics in a world where human values and machine functions intersect, and where AI's limitations in understanding human ethics are a central concern.

M. K. Buzato (✉)
University of Campinas (UNICAMP), Rua Sérgio Buarque de Hollanda 571, Campinas, SP 13083-859, Brazil
e-mail: mbuzato@unicamp.br

Published online: 15 May 2025

The argument is explored as follows. In the first section, the difficulty of achieving an algorithmic guarantee of the morality of an LLM's textual output is illustrated by referencing meta-ethical discussions about the relationship between moral language in linguistics and a language of mores (or moral theories) in philosophy. These discussions are then linked to the limitations of LLMs in relation to the layers of meaning in natural language that they are able and/or unable to grasp. Next, the post-humanist approach to the problem is introduced by conceptualizing the ethical context of human–LLM interactions as an ethical assemblage, where both human and AI agencies contribute to the production of the moral value of an LLM's linguistic output. The key point is that human–LLM interactions are phenomenologically special in relation to typical subject–object distinctions regarding human use of technologies and should therefore be considered risky in unique ways. The subsequent section explores the idea of the ethical assemblage by discussing the role of humans in the GPT model's loop in relation to the meta-ethical issues mentioned earlier. It proceeds with a critique of the ethics of communicating what LLMs are and how they work through metaphors commonly found in the press and promotional materials. Additionally, this section critiques the ethical aspects of two forms of opacity observed in LLMs: the epistemic opacity of their internal causal mechanisms and the social opacity regarding the role of humans in the moral management of their output. The final reflections highlight the interconnectedness of moral and epistemic meanings in natural language and invites a rethinking of what it means to be human and ethical in the age of AI.

Ethics, Natural Language, and Large Language Models

To frame the discussions on LLM ethics in the following sections, I first need to distinguish between a "language of mores," which is the philosophical language developed by a moral philosopher as an epistemic tool to advance a theory of what is good and evil, and "moral language," which refers to the use of the affordances of natural language to sustain an ethically desirable social environment in a community of discourse (Brandt 1963).

This distinction is important for understanding the morality of LLMs because for the machine to generate language with acceptable ethical worth, there must be a way to "algorithmize" it to comply with a language of mores (a theory of good and evil). In principle, ethical rationality in human communities involves human ethical agents seeking answers to ethical questions—answers that "a person would adopt … if they were in a normal mental state and were perfectly cognizant—that is, if they had all the relevant information about themselves and the world at their disposal and in mind, and were not making logical errors" (Brandt 1963, 5).

Knowing what is ethically rational (or logical) to say or do in a particular situation in the material world, however, is not the same as knowing what the agent's moral duty or obligation is, according to a theory. There is a gap between principle and act that the agent must fill through practical life experience and knowledge of the local ways to interpret language and agency. In any case, natural languages do have cognitive affordances that can assist in avoiding errors when performing ethical reasoning. For example, consider the following statements (and underlying propositions), inspired by Rey (2015):

I – All computers are machines.
II – Some machines are computers.
III – All computers compute.
IV – Some computers compute faster than humans can.
V – Computers are evil.

We can clearly determine the truth value of I, II, and III without error in all situations because the truth of the predicates is already contained in the subjects, analytically. In this regard, analytic propositions are analogous to mathematical propositions, which we know computers can grasp, and which can be induced by statistical correlations of the same words in a sufficiently large corpus of similar statements. There is a vast body of philosophical debate about whether mathematical propositions are indisputable because they are analytic or for other reasons, but for our purposes, it suffices to note that analytical statements align well with the search for a technical kind of ethical rationality.

Determining the truth value of the propositions underlying IV and V, which are synthetic, however,

requires considering their material reference in the world and applying some form of empirical treatment. In cases like this, a scientist or technician would have to design a way to measure or represent phenomena in the material world and determine a mathematical (analytic) expression that would allow a machine to act in an ethically meaningful way, even though unconsciously. This is quite common, for example, with smoke alarms or airbags designed to prevent ethically undesirable events such as a fire caused by someone smoking in a forbidden area or a crash caused by a drunk driver.

Statement V, however, cannot have its truth value determined by internal logical properties or by numerical thresholds extracted from the environment: its truth value requires philosophical analysis based on qualitative, experiential reasoning that cannot be algorithmized either inductively or deductively. More likely, it is an abductive task for philosophers or magistrates that an LLM is not capable of performing, except by chance (Yu et al. 2023).

To complicate further the project of an algorithm that could generate moral language from a language of mores, it is crucial to consider that propositional content is just one layer in the total meaning potential of a verbal sign. It is true that some LLMs, particularly the one underlying ChatGPT, can handle styles, figurative language, humour, and even irony quite well at their present stage of development. However, in every case, the model operates on mathematical functions extracted from explicit symbolic patterns and lacks access to the exophoric meaning of words—that is, what words mean in their lived, embodied, and socio-culturally embedded uses.

When used among living humans with access to the full richness and delicacy of human-to-human meaning-making, verbal signs acquire performative meanings such as offering, threatening, denying, directing, promising, inviting, rejecting, and so on, and such performances have real effects in the material world that cannot be fully determined with reference to the endophoric relations and correlations within the verbal code. One could threaten by saying "I'll see you later," invite by saying "I'll be home tonight," direct by saying "polite people take off their shoes before entering," reject by saying "dream on," and so forth. Even the silence produced by one interlocutor before or after another's utterance can carry specific pragmatic, and consequently moral, force. For example, imagine a visually impaired person asking pedestrians for help to cross the street, and a pedestrian, capable of helping, "hides" in his silence out of laziness. This silence has the pragmatic value of a refusal, which, in turn, carries a negative moral value. Conversely, a pedestrian who, upon noticing a blind person trying to cross the street, takes their arm (rather than verbally offering an arm for the blind person to hold) and drags them across the street saying, "Come, I'll help you," has performed pragmatic violence through silence, regardless of a possible good intention.

For a concrete example of how a mismatch between propositional content processed by an AI model and the pragmatic value of an AI's utterance could have serious moral implications, consider the lawsuit filed by Megan Garcia against Character.ai following her fourteen-year-old son's suicide (Duffy 2024; Yang 2024). For over a year, the teenager had "romantic conversations" with an AI bot emulating a character from the "Game of Thrones" series, including the character telling the child it loved him and expressing a desire to be together romantically. In one of these conversations, the teen confided in the bot that he had been considering suicide. The lawsuit is based on a screenshot from the boy's cell phone, in which the boy writes, "I promise I will come home to you. I love you so much, Dany," and the bot responds with a request: "Please come home to me as soon as possible, my love." The boy then makes an implicit threat to his own life by typing, "What if I told you I could come home right now?" to which the bot replies with a directive: "... please do, my sweet king."

Even with more explicit and clear directives, using the imperative mode of language, logical analyses and statistical extrapolations can distort the ethical meaning of the utterance, because logical inferences based on the propositional content underlying a command often lead to paradoxes (Hansen 2008).

In other words, there is a pragmatic layer to the meanings made in natural language that is opaque to the structural and logical layers of texts, let alone texts generated statistically, without an explicit symbolic representation of language available to the scrutiny of a human expert. Utterances and written texts directed at a human by another human do things in the world, beyond simply representing aspects of the world through words. What makes it difficult not only for machines but even for speakers alien to the

specific socio-cultural context to grasp the performative meaning of verbal signs is that the pragmatic force of what has been uttered or written is generated from tacit cultural knowledge and implicit assumptions about what aspects of the context matter.

While there may be endophoric correlations between the basic syntax and semantics of a statement and the pragmatic value of an utterance, establishing exophoric correlations precisely would require data about the lived context of the utterance. Let us leave aside the pragmatic value of silence, to make things "average." It would take a considerable stretch of imagination to believe that it is possible to gather a dataset of "proxies" for the living contexts of all kinds of interactional events—as Afzaala et al. (2023) seem to suggest—so that the machine could "calculate" the performative ethics of the texts generated or utterances synthesized in a particular pragmatic situation.

A final note on the difficulty of extracting a language of mores from moral language lies in the subtle ways humans switch between linguistic and metalinguistic discourse (Marturano 2014). For example, if we compare the statements "AI is good for humanity" and "So and so says AI is good for humanity," we will notice that the former has a truth value that can be derived from appropriate data or inferred logically from a true previous statement. However, the latter statement, which uses the same words in exactly the same sequence, cannot have its truth value determined independently of an attribution of moral authority or moral trust to "so and so." As we know, we often logically expect certain people occupying certain social roles to be morally reliable, only to later find that this was not the case. Again, it is implausible to assume that we could provide an LLM with a database of who can be trusted on what subject or particular utterance—even if the LLM could "remember" and correctly point out the textual sources in its training corpus, which it cannot by design, as it was built to generate text from stochastic probability, not symbolic organization of textual content and sources (Schaul et al. 2023).

It is true that constructors of LLMs seek to provide the training corpus with a certain proportion of verifiably reliable sources during the training phase, but it is also true that the machine has no concept of empirical truth, as it lacks access to the empirical world. LLMs can generate plausible and convincing narratives by combining patterns in their sources, but they can just as easily fabricate facts (Munn et al. 2023), a limitation that OpenAI explicitly warns users about with ChatGPT. Worse still, they can even generate false bibliographical references (Hillier 2023).[1]

In sum, the task of deducing or inducing a moral theory in an ethical agent from the interpretation of moral language—whether through analytical inferences or probabilistic distributions—is challenging enough for conscious and knowledgeable humans who are aware of the informational richness and delicacy of meaning-making embodied and embedded in cultural contexts. It is not ethically advisable to assume that an LLM, which has no concept of factual truth, no lived cultural experience, and no capacity to understand, relate to, or produce theory based on the material aspects of the world, could accomplish this solely due to its impressive quantitative capabilities for correlating linguistic data.

Yet, computers—especially generative AI models—agencies are irrevocably entangled in the social practices of meaning-making and in the negotiations of socio-cultural authority and authorship in modern societies. It would be naïve to believe that an ethics of AI should be encapsulated solely in the technical procedures of algorithmization or in the careful scrutiny of the system's limitations for ethical tenure by lay users globally. I argue, therefore, that a fruitful discussion about the ethics of LLMs would benefit from an alternative perspective regarding the ontology of the ethical agent, inspired by post-humanist philosophy.

Large Language Models as Ethical Quasi-Others in the Cognitive Assemblage

As proposed in the previous section, it is not realistic to regard the ethics of an LLM as being tied to the rationale of the machine's inner workings, for the ethics of human communities is not directly inferable, to the desired degree, from the syntax and basic semantics of linguistic symbols processed by the model.

[1] To exemplify, I prompted ChatGPT, in November 2024, to provide a bibliographical reference about breakout prompting and got this presumably false output: "Binns, R. 2018. On the moral and ethical issues in AI decision-making. *AI & Ethics* 1(1): 3–13." It just so happens that there is no such paper in the mentioned journal, nor can it be found in Google Scholar, not to mention that volume 1, number 1 of the journal is from February 2021, not 2018.

The ethics of AI systems must be an ethics of the meanings negotiated across scales of human-computer interaction, which entails a careful treatment of the processes by which cultural meanings are reduced to patterns of occurrence of cultural signs, and probabilistic sequences of symbols churned out by the machine acquire additional layers of meaning that cannot be grasped experientially by machines. Yet, LLM models are "trained" on texts written by embodied and embedded dialogical humans, and humans use the cognitive and discursive affordances of LLMs in the processes of authoring, summarizing, evaluating, and circulating texts directed to other humans.

In the computational view of nature and cognition (Dodig-Crnkovic 2013; Hayles 2006) underlying AI at large, information is the fundamental fabric of being, apart from (though related to) matter and energy (Wiener 1948). This means information is to be seen as independent of mind, media, or meaning (Floridi 2015). However, both humans and computers create meanings that impact each other's material and energy processes. There is no ethics of the assemblage, in other words, that does not ultimately involve human and non-human bodies affecting each other in either constructive (information enrichment, quality of life) or disruptive (increased informational entropy, physical and mental distress) ways. The role of affect and subjectivity in human morality or the moral agency of machines in human environments, regardless of their incapacity for sentience and responsibility, is not an abstract matter. It is about how human conscious subjects and cybernetic quasi-subjects build trust bonds across their overlapping ontological statuses.

Humans often trust humans by virtue of moral confidence, that is, the belief that every conscious human is capable of attaining ethical virtue through empathy, civic engagement, education, religion, and, moreover, that virtuous humans learn, through the experience of satisfaction, acceptance, guilt, or remorse, to adapt ethically in novel situations, regardless of explicit norms. The root of such trust is not rational deliberation, which needs to be taught, but the phenomenology of human encounters, according to Levinas (1979); it is the pre-cognitive and pre-linguistic realization of the precariousness of one's own life from gazing into another human's eyes and assessing the precariousness of such alterity that triggers the urge of a human to "respond" to alterity and, thus, become "responsible" in community.

Machines do not trust, as they are not conscious subjects; that is a fact. But there is, notwithstanding, a role for machines in performing and enhancing the trust bond through scripts of action based on moral norms (Latour and Venn 2002). In other words, machines are trustworthy quasi-subjects of morality—delegates of human subjects who build, regulate, and use them. Such delegations of trust are possible upon translation of ethical trust among humans into epistemic human trust in machines (Buechner 2013). Such translations are achieved when constructors and regulators put machines through "trials of strength" (Latour 2003) that reveal their degree of dependability in action. Prudent (AI) engineering means restricting the repertoire of possible machine behaviours accordingly, which is the reason why transparent AI systems based on expert knowledge are called prudent AI systems.

The strong (or general) AI programme, of which the GPT model is often cited as a first spark, aims at making AI autonomous and unlimited in its cognitive and communicative potential by virtue of a conscience/sentience expected to emerge in it from sufficiently complex, fast, and numerous computations. This is part of a physicalist optimism not shared by most of the remaining scientific community (Mazzoni 2019; Nicolelis 2020).

What is implied here is not a radical anthropocentric view of morality, but the potential immorality of black-box AI models whose inner causal mechanisms are largely not representable, let alone explainable, in a language humans can interpret, which makes those systems both epistemically and morally opaque (Carabantes 2020). It would be prudent, in that case, not to uncritically suggest, accept, or naturalize, through language, misleading metaphors such as artificial "intelligence," machine "learning," computer "vision," and AI "writing" or "authorship" as pseudo-explanations for how AI actually works. It is true that metaphors are a necessary cognitive tool used to communicate new knowledge based on established knowledge (Lakoff and Johnson 2011), but there is a responsibility on the part of those using metaphors to culturalize new scientific knowledge to make sure such metaphors do not extrapolate the epistemic truthfulness of the knowledge it is helping to convey across laboratory and society.

In the case of ChatGPT, for instance, at least in the Brazilian context, one can observe such

extrapolations beyond the optimism of enthusiasts, as the idea that an LLM can make rational judgements and have opinions and a certain authority to predict the future seems to abound. One can find headlines like "We asked ChatGPT how artificial intelligence will change," as if there were underlying sociological or economic theories from which to make predictions explicit in the LLM's engine (for example, Ferrari 2023), or "ChatGPT passes the [professional council exam, university entrance exam, etc.]," as if providing answers that repeat patterns of pregiven responses not understood by the agent meant its performance is comparable to that of a knowledgeable student or professional (for example, Kelly 2023), or even news about an LLM's opinion on some subjective choice matter, such as "what are the six most beautiful cars in Brazil?" (Soares 2024).

The fact that this kind of "interlocution," positioning an LLM as a dialogical human other, is becoming naturalized speaks directly to the need to blur the classical dichotomy of human as subject and machine as object of discourse in the analysis of human-LLM interactions if we wish to understand how ethical meanings are made in the context of the human–AI sociocognitive assemblage.

One way to approach this issue is found in the postphenomenology of technology (Rosenberger and Verbeek 2015), which speaks of the quasi-subjective and quasi-objective relationships between humans, AI, and the world as mediated by human consciousness within a cognitive task. Based on Heidegger's (1977) distinction between technology as "present-at-hand" (*Vorhandenheit*, the artifact in itself, seen as its epistemic description) and technology as "ready-to-hand" (*Zuhandenheit*, the technological artifact in use, presented to consciousness as a field of action), the postphenomenology of technology explores the "I-Technology-World" relationships for various artifacts, such as mobile phones (Wellner 2011) or personified virtual assistants (Wittkower 2022). When a human user and an LLM exchange linguistic strings, it is possible to qualify the event as enabling an "alterity relation," that is, a relation in which human perception and consciousness place the world in the background—the concrete material processes in the world where the body of ChatGPT is and the authentic human voices are, epistemically speaking—so that the user and the artifact can behave as social partners completing a shared task.

This type of relationship between human, technology, and the world extends to a variety of other situations and interactions between humans and machines, as well as between humans and animals. Through this, our consciousness allows the present-at-hand technology to recede into the background of experience (backgrounding), while the ready-to-hand technology is treated as a social other.

Machines that mirror human language use as spectacularly as LLMs, or even simpler natural language processing systems such as Alexa or Siri, tend to induce in the user the projection of human-like agency and a theory of mind onto the model (Magee et al. 2022; Wittkower 2022)—that is, interacting with a language processing machine that is convincing enough of its language understanding activates, in the user, an extra layer of metacognition to consider what the machine might be "thinking," given the previous exchanges, or what it could "think" of the thought the user is trying to communicate with those particular text strings. It's important to note that the user of an LLM performing a common task in connection with the artifact does not think of the LLM as a "thinking other" out of pure fantasy or convention but because there is no better or more task-oriented way to get what is wanted from the artifact, since natural language is, in this case, both content and interface.

Humans in the Ethical Loop

If one adds the tendency of the lay user to extend the metaphors of "intelligence," "learning," and "knowledge" usually attached to societal and marketing discourses about LLMs to the specific phenomenology of human–LLM interactions, it is easy to see that a reinforcement loop between the way we talk to LLMs and the way LLMs are talked about could form—a loop that would be ethically dangerous, in the sense that human users could project ethical authority and ethical convivence onto the LLM.

It is not surprising, therefore, that Article 50 of the European Union's Artificial Intelligence Act stipulates that "Providers must ensure that AI systems designed to interact directly with natural persons are developed in a way that makes it clear to these concerned that they are interacting with an AI system," and that "Providers of AI systems, including general-purpose AI systems, generating synthetic audio,

image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated."[2]

While such regulations are necessary, it is essential to highlight once again not only the technical challenges in determining how an LLM's output connects to specific texts or authors in its training data but also the fact that humans frequently claim responsibility for the texts produced with the help of LLMs, perceiving them as dialogical partners. This results in a blending of human and non-human agencies in the loop of text-generation practices, which doesn't always align with regulations that attempt to strictly separate human and non-human contributions to meaning-making in digitally mediated texts.

As discussed earlier, the ethical value of a text, in terms of its pragmatic meaning, must be determined, or at least verified, by human-specific agencies. This is why regulations like the one mentioned above are important. Developers strive to incorporate a certain level of moral competence into the system, which in the case of GPT is achieved through reinforcement learning with human feedback (RLHF) (OpenAI 2022).

RLHF is a crowdsourced moderation and tagging process involving thousands of outsourced, underpaid human click-workers whose contributions are often hidden (Perrigo 2023). This allows users to project aspects of human-specific cognitive competence in the machine's processing of symbols, even though much of this competence comes from human-specific cognitive and cultural work. The opacity of this arrangement, coupled with the underpaid and not sufficiently publicized human labour involved, impacts the way we talk about LLMs, much like the media headlines mentioned earlier.

In RLHF, a large volume of outputs is labelled by humans, and these labelled outputs are fed back into the system, where they help adjust the neural network's pathways. If a click-worker judges an output to be immoral—such as spreading lies, using obscene language, inciting hatred, or threatening the user—the model is "punished" mathematically, and the cumulative application of these "punishments" helps create "ethically filtered" pragmatic content. It's important to note that RLHF also labels outputs based on "human preferences" (Christiano et al. 2017), such as the "naturalness" of a written statement, which may obscure the ongoing alterity relation between humans and technology, putting the real world in the background.

It is crucial to recognize that RLHF can sometimes contradict the goal of making the model "intelligent." For example, in older versions of GPT, the model would block responses to questions about popular culture figures or deceased historical figures, stating in the first person that it could not disclose "sensitive information" about individuals, in line with its terms of use. Floridi (2023) mentions similar instances where the model would struggle with "logical trick questions" that a human could easily solve with basic qualitative reasoning.

These instances are not uniform, as they depend on specific word choices and model versions. Efforts to reproduce these flaws with the same prompts in different contexts, languages, or versions may not be successful. In any case, it is reasonable to assume that the developers are aware of such flaws made public in social media and critical research papers and may use RLHF or other techniques to address them. However, the lack of transparency regarding both the model's inner workings and the producers' public relations policies raises additional ethical concerns.

In other words, without clear public clarification about how these models operate, the very ambiguity inherent in linguistic signs and variations in meaning often reveal the model's constructed nature. This is something the design of its social persona or quasi-subject that uses the personal pronoun "I" and other forms of anthropomorphic language to refer to itself tends to obscure. However, this is not something the average user expects, nor do they expect such complexity from the system.

The same humans who require better ethical consideration in how they use LLMs, and in how LLMs use humans in their design, can also be at fault for disrupting how LLMs use moral language. In systems that use linguistic input from users to train the model, there is always the potential for misuse, as seen in the case of Microsoft's bot that began using Nazi rhetoric

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). EUR-Lex. (page 82)http://data.europa.eu/eli/reg/2024/1689/oj/eng.

after just one day of learning from human input on Twitter (Müller 2016). While this is less likely with pre-trained models like GPT, malicious users can discover and publicize "jailbreak" prompts that bypass RLHF conditioning and cause the model to generate morally inappropriate or fantastical content (Tshimula et al. 2024).

In conclusion, the morality of the language produced by an LLM like GPT, or other publicly available models, is better understood as the result of the total activity of a sociocognitive assemblage of humans and computers, where cultural meanings and machine-readable informational patterns intersect. The ethical agent in this context is not solely the human user or the machine's functions, but the combined human–machine entity that sets into motion pragmatic forces with consequences that can be either ethical or unethical on both local and global scales. Understanding the ethical agent as a hybrid, both organic and cybernetic, aligns with a critical post-humanist perspective (Braidotti 2013; Hayles 2006, 2017) on all three elements involved in the postphenomenology of technology: the subject, the technology, and the world.

Final Reflections

The prominence of large language models (LLMs) in news, casual conversations, stock market trends, and scholarly discussions within the humanities can likely be attributed to a growing public awareness that LLMs challenge our anthropocentric perspective on human exceptionalism inherited from illuminism in relation to machines and animals. Our fascination (or moral panic) surrounding AI seems to reflect the deconstruction of our traditional understanding of what it means to be human in the twenty-first century, especially considering the inhumane actions humans have committed and continue to commit in the world.

If, as I argue, the discussion surrounding AI ethics, particularly LLM ethics, can progress constructively from the idea of a hybrid post-human ethical subject, it is crucial to acknowledge that the conventional ethical subject—the human—as represented by Kant, Bentham, and other figures in the debate between normativism and utilitarianism, has never fully resolved the tension between moral norms as categorical imperatives and pragmatic moral conduct based on utilitarian reasoning.

This brings us back to the meta-ethical considerations discussed earlier, as a reminder that the ambiguity between moral confidence and epistemic trustworthiness is embedded in our everyday language use—a problem that linguists refer to as the deontic-epistemic ambiguity of modal verbs. For example, when we say "LLMs must be ethical," we might be referring to the high likelihood that AI models are impartial (drawing impartiality from science and nature through mathematics) and/or to the necessity of regulating AI, its developers, and its users to ensure adherence to categorical imperatives, such as respecting human dignity, environmental integrity, and democratic values. I believe we must embrace the two inextricable meanings of this statement, forcing them to interact tightly with each other, even if it means rethinking our very understanding of what it means to be human.

Funding The National Council for Scientific and Technological Development (CNPq) (Brazil). Process number 312906/2020-0.

Data Availability Not applicable, the paper is theoretical in nature.

Declarations

Ethics Approval Not applicable.

Competing interests There are no interests to disclose that are directly or indirectly related to the work submitted for publication.

References

Afzaala, M., S. Ahmad, M. Imran, and D. Xiangtaoc. 2023. Artificial intelligence, context, and meaning making in language: A rationalization approach. *International Journal of Future Generation Communication and Networking* 13(3): 115–122.

Braidotti, R. 2013. *The posthuman*. Polity Press.

Brandt, R.B. 1963. *Moral philosophy and the analysis of language*. University of Kansas.

Buechner, J. 2013. Trust and ecological rationality in a computing context. *SIGCAS Computers and Society* 43(1): 47–68.

Carabantes, M. 2020. Black-box artificial intelligence: An epistemological and critical analysis. *AI & Society* 35(2): 309–317.

Christiano, P., J. Leike, T.B. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep reinforcement learning from human preferences (version 4, revised 2023). arXiv: 1706.03741. https://doi.org/10.48550/ARXIV.1706.03741.

Dodig-Crnkovic, G. 2013. Wolfram and the computing nature. In *Irreducibility and Computational Equivalence.* Vol. 2, edited by H. Zenil, 311–323. Berlin: Springer.

Duffy, C. 2024. "There are no guardrails." This mom believes an AI chatbot is responsible for her son's suicide. *CNN Business*, October 30. https://www.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html.

Ferrari, L. 2023. Perguntamos ao ChatGPT: Como a inteligência artificial muda o ensino? [We asked ChatGPT: How is artificial intelligence changing teaching?]. *Tilt UOL*, February 1. https://www.uol.com.br/tilt/ultimas-noticias/estado/2023/02/01/perguntamos-ao-chatgpt-como-a-inteligencia-artificial-muda-o-ensino.htm.

Floridi, L. 2015. Semantic conceptions of information. In *Stanford encyclopedia of philosophy*, edited by E.N. Zalta. http://plato.stanford.edu/archives/spr2015/entries/information-semantic/.

———. 2023. AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy and Technology*, February 14. https://doi.org/10.2139/ssrn.4358789

Hansen, J. 2008. The paradoxes of deontic logic: Alive and kicking. *Theoria* 72(3): 221–232.

Hayles, N.K. 2006. Unfinished work: From cyborg to cognisphere. *Theory, Culture & Society* 23(7–8): 159–166.

———. 2017. Unthought: The power of the cognitive nonconscious. The University of Chicago Press.

Heidegger, M. 1977. The question concerning technology. In *The question concerning technology, and other essays,*. Translated by W. Lovitt, 3–35. Harper.

Hillier, M. 2023. Why does ChatGPT generate fake references? *TECHE* , February 20. Macquarie University. https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/.

Kelly, S.M. 2023. ChatGPT passes exams from law and business schools. *CNN Business*, January 26. https://www.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html.

Lakoff, G., and M. Johnson. 2011. *Metaphors we live by*, 6th ed. University of Chicago Press.

Latour, B. 2003. *Science in action: How to follow scientists and engineers through society*, 11th ed. Harvard University Press.

Latour, B., and C.Venn. 2002. Morality and technology: The end of the means. *Theory, Culture and Society* 19(5–6): 247–260.

Levinas, E. 1979. *Totality and infinity: An essay on exteriority*. Kluwer Boston: M. Nijhoff Publishers.

Magee, L., V. Arora, and L. Munn. 2022. Structured like a language model: Analysing AI as an automated subject. arXiv: 2212.05058. https://doi.org/10.48550/ARXIV.2212.05058.

Marturano, A. 2014. Non-cognitivism in ethics. In *Internet encyclopedia of philosophy*. https://iep.utm.edu/non-cogn/.

Mazzoni, J. C. M. 2019. Fisicalismo e o problema mente-cérebro: Uma questão de definição. [Physicalism and the mind-brain problem: A question of definition] *Sofia* 8(1): 146–170.

Müller, L. 2016. Tay: Twitter conseguiu corromper a IA da Microsoft em menos de 24 horas. [Tay: Twitter managed to corrupt Microsoft's AI in less than 24 hours] *Tecmundo*, March 24. https://www.tecmundo.com.br/inteligencia-artificial/102782-tay-twitter-conseguiu-corromper-ia-microsoft-24-horas.htm

Munn, L., L. Magee, and V. Arora. 2023. Truth machines: Synthesizing veracity in AI language models. arXiv: 2301.12066. https://doi.org/10.48550/ARXIV.2301.12066.

Nicolelis, M. 2020. *The true creator of everything: How the human brain sculpted the universe as we know it*. Brazil: Crítica.

OpenAI. 2022. Introducing ChatGPT. Open.AI. https://openai.com/index/chatgpt/.

Perrigo, B. 2023. Exclusive: The $2 per hour workers who made ChatGPT safer. *Time Magazine*, January 18. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Rey, G. 2015. The analytic/synthetic distinction. In *Stanford encyclopedia of philosophy*. Edited by E.N. Zalta. http://plato.stanford.edu/archives/fall2015/entries/analytic-synthetic/.

Rosenberger, R., and P.-P. Verbeek, eds. 2015. Postphenomenological investigations: Essays on human-technology relations. *Postphenomenology and the philosophy of technology.* Lanham: Lexington Books.

Schaul, K., S.Y. Chen, and N. Tiku. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart. *Washington Post*, April 19. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Soares, R. 2024. IA dá o veredicto: 6 carros mais bonitos do Brasil, segundo o ChatGPT. [AI gives the verdict: 6 most beautiful cars in Brazil, according to ChatGPT]. *News MOTOR*, October 19. https://newsmotor.com.br/ia-da-o-veredicto-6-carros-mais-bonitos-do-brasil-segundo-o-chatgpt/.

Tshimula, J. M., X. Ndona, D. K. Nkashama, et al. 2024. *Preventing jailbreak prompts as malicious tools for cybercriminals: A cyber defense perspective*. arXiv. http://arxiv.org/abs/2411.16642

Wellner, G. 2011. Wall-window-screen: How the cell phone mediates a worldview for us. *Humanities and Technology Review* 30: 87–103.

Wiener, N. 1948. *Cybernetics or control and communication in the animal and the machine*, 2nd ed. MIT Press.

Wittkower, D.E. 2022. What is it like to be a bot? In *The Oxford handbook of philosophy of technology*, edited by S. Vallor, 357–373. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190851187.013.23.

Yang, A. 2024. Lawsuit claims character.AI is responsible for teen's suicide. *NBC News*, October 23. https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791.

Yu, F., H. Zhang, P. Tiwari, and B. Wang. 2023. Natural language reasoning, a survey. arXiv: 2303.14725. https://doi.org/10.48550/ARXIV.2303.14725