Philosophy of AI
Vol. 2, 2026, 61-79
DOI: 10.18716/ojs/phai/2026.3570

ON THE DISCOURSE ETHICS OF LARGE LANGUAGE MODELS: BETWEEN COMMUNICATIVE AGENTS AND LINGUISTIC MACHINES

Paolo Monti[*]

University of Milan-Bicocca, Department of Human Sciences for Education "Riccardo Massa", Milan, Italy

[*]*paolo.monti@unimib.it*

Abstract Large Language Model (LLM) based chatbots engage human users in discursive practices as interlocutors capable of vast linguistic performances. In a Habermasian discourse ethics framework, however, they do not qualify as proper communicative agents, since their behavior adapts to the input from human users but is not based on the self-reflective normative re-orientation of agency required by the theory of communicative action. Nonetheless, AI chatbots affect the communicative agency of their interlocutors who engage with them by gradually reshaping their access to the linguistic resources that are essential to the self-reflective assessment of their needs and wants. In this sense, LLMs operate as linguistic machines that influence the will formation process by enacting linguistic repertoires according to ends that fall outside of the cooperative pursuit of mutual understanding among communicative agents, with a distinct lack of a self-reflective and self-interpretive component. The prospect of LLMs being deployed to pervasively articulate specific normative repertoires in public discourse is attracting increasing political attention and is at risk of establishing new forms of domination through linguistic value capture.

Keywords Large Language Models, Discourse Ethics, Chatbots, Jürgen Habermas, Charles Taylor, Moral Argumentation, Deliberation, Value Capture, Non-Domination.

1. Introduction: discourse ethics vis-à-vis non-human interlocutors

In the short span of a few years, the experience of having a fluid and sometimes stimulating conversation with a non-human interlocutor has quickly moved from the realm of the unprecedented to the space of everyday tasks. Despite their wide availability, however, chatbots based on Large Language Models (LLMs) are still a source of puzzling questions when it comes to the nature of their interactions with humans (Bender et al., 2021; Bender, 2024), the responsibility gap left behind by the consequences of their utterances (Königs, 2022; Vallor & Vierkant, 2024), and the still unexplored ethical and political implications that descend from the large-scale societal application of their powerful linguistic capabilities (Kreps & Kriner, 2023; Zuber & Gogoll, 2024).

Discourse ethics has something specific to offer in addressing these questions, because of the way it construes the connection between moral agency and linguistic structures: in this framework, what humans do with words, thanks to their linguistic capacities, is also their defining agency as it serves to structure their many forms of interaction and cooperation. Language, here, is not a tool in service of exchange and coordination, but rather a capacity that is constantly enacted in speech acts through which humans define their purpose in the world and their mutual orientation. Humans are subjects of discourse because they share the same lifeworld, i.e., they inhabit the same space and define their social bonds by linguistic means (Apel, 1998; Habermas, 1994, 2020).

Jürgen Habermas, whose version of discourse ethics I am focusing on in these pages, pointed out that this way of construing the connection between speech acts and social bonds is indebted to Hannah Arendt's understanding of publicity. For Arendt, public life is a space of appearance where human beings emerge originally as a diverse – and unruly – plurality of agents: the use of speech is the essential medium through which they shape this space into a realm of regulated and coordinated agency (Arendt, 1998, p. 189). As Habermas notes, acknowledging his conceptual liaison with Arendt,

> the basic communicative action is the medium in which the intersubjectively shared lifeworld is formed. It is the "space of appearance" in which actors enter, encounter one another, are seen and heard. [...] In communication, individuals appear actively as unique beings and reveal themselves in their subjectivity. At the same time they must recognize one another as equally responsible beings, that is, as beings capable of intersubjective agreement – the rationality claim immanent in speech grounds a radical equality. Finally, the life-world itself is filled, so to speak, with praxis, with the "web of human relationships". (Habermas, 1977, p. 8)

The proposition is then that, if humans inhabit their shared lifeworld linguistically, their discursive practices are also the central site of their moral agency and the source of their normative arrangements. Despite the absence of a systematic engagement from Habermas with the topic of generative AI, these ideas provide abundant grounds to attempt one. For the sake of brevity, other discourse ethics frameworks, such as Karl-Otto Apel's, are not directly addressed here. However, some of the following considerations on communicative action, validity claims, and LLMs can be applied to that framework as well.

The overall goal is to explore how and to what extent the discourse ethics framework can be applied to questions about the behaviors of conversational AI systems and their impact on moral deliberation. Each section presents an argument based on the conclusion reached in the previous one, as follows:

Section 2 considers the communicative behaviors of LLM-based chatbots and concludes that, while their behaviors can be construed as communicative performances, in Habermasian terms they do not display full communicative competence and fail to qualify as communicative agents. AI chatbots are thus construed as linguistic machines capable of extensive communicative performances but lacking proper communicative agency.

Section 3 offers a moral assessment of the influence of linguistic machines on their users' deliberation: conversational systems seamlessly influence the will-formation process of their human interlocutors but, lacking the status of communicative agents, they do so in non-reciprocal ways. Developments in explainable AI (XAI) could partially assuage this concern, but LLM-based systems would still lack the kind of reflexive self-orientation towards understanding that is required by discourse ethics.

Section 4 explores the impact of linguistic machines on discursive will formation by considering how they deploy evaluative language into the exchanges of a community of speakers, thus affecting their linguistic resources of self-understanding. To this end, insights from Charles Taylor's expressivist theory of language are introduced to complement the Habermasian agent-centered approach.

Section 5 characterizes the impact of linguistic machines on a community of speakers as a form of value capture and qualifies it, in normative terms, as a risk of linguistic domination.

Section 6 sketches some conclusions about the implications of this analysis for future research in moral and political theory.

This account builds upon an earlier treatment of the applicability of discourse ethics to LLMs centered on the uncertain moral status of conversational AI systems (Monti, 2024). I offered there a comparative assessment of the moral status of human and AI agents from a Habermasian perspective, which is not discussed here. Instead, I expand on my previous treatment of the applicability of communicative agency to the behaviors of LLM-based systems. In particular, in section 2, I qualify and partially revise my previous position through a broader examination of the parallel between linguistic competence and communicative competence in the case of conversational AIs, and I more extensively reconsider the implications for the explainability of these systems. In section 3, I reframe my earlier conclusions about the lack of communicative agency in LLMs by introducing the notion of linguistic machine as a new kind of speaker that results from the decoupling between communicative performance and communicative agency. Sections 4 to 6 cover entirely new ground and explore the implications of my analysis for understanding how LLM-based systems use evaluative language, introduce a new form of value capture, and pose a political risk of domination.

2. Do LLM-based chatbots qualify as Habermasian communicative agents?

Habermasian discourse ethics focuses on the moral nature of speech interactions among the members of a linguistic community. It aims at articulating the intrinsic obligations that stem from communication among peers who coordinate their actions based on their discursive exchange. From this vantage point, the introduction of chatbots into human discursive practices is a conundrum: how to reconcile the fact that these systems interact like interlocutors of noteworthy linguistic prowess with the notion that they are also not moral peers of their human counterparts? The decoupling between the authorial properties they display and the lack of a clear moral status demands a reconsideration of what kind of competence and agency, if any, is at play (Gubelmann, 2024; Van Woudenberg et al., 2024).

Habermas developed Chomsky's notion of *linguistic competence* into that of *communicative competence*. If linguistic competence refers to a speaker's implicit knowledge of the rules and structures of language, communicative competence is the know-how of the implicit rules and presuppositions that make the speakers capable of producing and understanding utterances directed at others (Habermas, 1970; Allen, 2019). The use of language, in the Habermasian context, is relevant as part of a communicative process that produces different kinds of interaction with other language users. This thread connecting the Chomskyan notion of linguistic competence with the Habermasian communicative competence is useful to develop some considerations applicable to LLMs. The Chomskyan understanding of language is not universally accepted, and some have recently argued that we should look at the success of LLMs as a sort of refutation of his rationalist theory of language (Piantadosi, 2024). However, this stance has been, in turn, criticized for assuming that the AI modelling

of language offers appropriate ground to build explanatory theories applicable to humans (Katzir, 2023; Volenec & Reiss, 2025). For the sake of our account, it is sufficient to acknowledge that this ongoing debate leaves much room for plausibly exploring how Chomskyan concepts can illuminate the linguistic dimension of Habermas' theory of communicative action and its application to LLMs.

Chomsky distinguishes between *linguistic competence* and *linguistic performance* (Chomsky, 1965, p. 4). Competence refers to the speaker-hearer's knowledge of language and the rule-based capacity for articulating it, while performance designates the actual use of language in concrete situations, as it flows between speakers under the influence of contextual factors. LLMs produce noteworthy linguistic performances when engaged by human speakers, insofar as they generate grammatically coherent and contextually appropriate sentences. These performances are produced as the output of language processing that is statistical and pattern-based rather than rule-based in the way Chomsky describes linguistic competence. Because of the fundamentally different process through which LLMs generate discourse compared to their human counterparts, we can acknowledge the linguistic performances of LLMs while also, at the same time, making important distinctions when it comes to attributing to them proper linguistic competence. The notion that LLMs have linguistic competence, in the sense of having internalized syntactic and semantic rules, is subject to debate. Their consistent linguistic performances suggest that these models have acquired a stable capacity to make use of grammar and syntax, but at the same time the specific way in which this capacity is acquired and enacted fundamentally differs from its human analog: AI systems are trained on massive corpora of written text and formulate discourse based on stochastic mechanisms of word prediction, very much unlike human children who acquire their linguistic competence from verbal conversations that point to sensory experiences and learn based on semantic associations. To unpack the false equivalences that can stem from attributing the notion of linguistic competence both to human speakers and LLMs, it is useful to rely on a further distinction between formal and functional linguistic competence (Mahowald et al., 2024): formal competence signals the ability to make correct use of the form of language, such as composing words with the appropriate strings of text and positioning words correctly to form valid sentences, whereas functional competence refers to the ability to use language to achieve certain goals in the world by effectively communicating with others. LLMs can be thought of as systems that achieve formal linguistic competence through their mastery of the statistical regularities of language, although this still leaves important blind spots, like the inability to separate possible from impossible natural language (Leivada et al., 2023). But where more evidently LLMs fall short is the area of functional linguistic competence, because they lack the integration of linguistic and non-linguistic cognitive functions that characterize human agents in their relationship with others and the world (Mahowald et al., 2024).

This gap between formal and functional *linguistic competence* is helpful to examine the parallel Habermasian notion of *communicative competence*, i.e. a competence to engage in discourse that is oriented towards reaching mutual understanding with others that share the same world. This orientation, according to Habermas, takes the specific form of discursively raising validity claims to truth, normative rightness, and sincerity aimed at intersubjective recognition: competent speakers act based on the implicit expectation that, under suitable conditions, their claims to truth, normative rightness, and truthfulness should be acceptable to all interlocutors (Habermas, 1990, p. 31). In human interactions, even behaviors that intentionally exploit and betray these expectations, like lies and deceptions, are possible because all communicators rely on such implicit presuppositions. This perspective has been variously criticized for its underlying Kantian rationalism and reliance on universalistic

idealizations. Among others, Richard Rorty entertained a longstanding critical engagement with Habermas contesting the aspiration to universality of the validity claims raised in our everyday language games (Rorty, 1995, 2021). This criticism is not without merit, but for the purpose of this account, it is not central, as most of the relevant considerations on the nature of communicative agency do not entail a commitment to universal validity, but rather to the relations that the speaker entertains when raising those claims.

According to Habermas, each type of validity claim – truth, rightness, and sincerity – refers to a specific kind of lifeworld relation: to the objective natural world, to the intersubjective social world, and to the inner subjective world. Text generated by AI bots, however, does not originate from such relations, and it is then highly problematic to construe it as a fragment of proper communication whose validity claims can be challenged and adjudicated among humans and chatbots through discourse. AI-generated truth claims are probabilistically related to those contained in the textual sources that originally fed the machine learning process, since the systems have no "experience" of the world or direct access to the natural world as an environment they live in to raise and verify truth claims of their own. Normative rightness claims are based on judgments about the appropriateness of speech acts, but these depend on the social relationships that the participants in the conversation have as peers that inhabit a shared lifeworld, a ground that is not available to chatbots. Sincerity claims should be vindicated by the consistency between actions and the claimed subjective states of the speakers, but there is no equivalent for LLM-based systems, which have no internal subjective states and whose claims in terms of emotion or preference are purely the textual outcome of their designed anthropomorphism. In other words, these systems are not subjects that raise or contest validity claims in the same way as their interlocutors are, because they do not share the same kind of lifeworld relations that, in the Habermasian framework, support those claims. In parallel with the previous discussion of linguistic competence, we can designate the way LLM-based chatbots interact with their human counterparts as a *communicative performance* without full communicative competence, because communicative competence requires access to non-linguistic relations that these models lack.

When approaching a text, Habermas suggests, interpreters presume that they can grasp the author's intended meaning through a certain familiarity with the context within which the text was produced and is meaningful. This presumption is based on the idea that both interpreter and author advance validity claims regarding truth, values, and sincerity within a particular context, yet the rational justifications behind these claims remain accessible across contexts (Habermas, 1990, p. 30). Users that engage with AI-generated text can interpret it as meaningful insofar as they assume that it is the expression of an interlocutor who produces and interprets meaning as they do. However, LLMs do not rely on relatable reasons rooted in their connections to a lifeworld and cannot reason about mutual mental states (Trott et al., 2023). Rather, LLMs deliver an accurate simulation of an appropriate response to the user's textual prompt by extrapolating from the existing array of suitable responses available within their training data. The interpreter might still discern plausible discourse around the subject at hand, but comprehension occurs "as if" the author possessed reasons analogous to the interpreter's:

> For reasons to be sound and for them to be merely considered sound are not the same thing, whether we are dealing with reasons for asserting facts, for recommending norms and values, or for expressing desires and feelings. […] Reasons can be understood only insofar as they are taken seriously as reasons and evaluated. This is why the interpreter can elucidate the meaning of an obscure expression only if he explains how this obscurity

> came to be, that is, why the reasons the author might have given in his own context are no longer immediately illuminating for us. (Habermas, 1990, pp. 30–31)

Yet, the conditions for such an explanation differ significantly when interpreting a text produced by LLM. Anthropomorphic first-person statements do not originate from a personal engagement with an individual lifeworld; hallucinations are unexpected results of stochastic processes rather than genuine perceptual distortions. This casts doubt on the extent to which LLMs, notwithstanding their communicative performances, can truly embody the «know-how of subjects who are capable of speech and action, who are credited with the capacity to produce valid utterances, and who consider themselves capable of distinguishing, at least intuitively, between valid and invalid expressions» (Habermas, 1990, p. 31).

Examining this matter from the viewpoint of internal intuitions might seem unproductive in the case of AIs, but these systems can be designed with varying abilities to point to "reasons" for their outputs, thereby assisting interpretation. The explainability of AI systems (XAI) is attracting increasing attention (Preece, 2018), accompanied by heightened awareness of its ethical significance (McDermid et al., 2021). In this context, I think of XAI specifically in the perspective of a human-centered approach to the field, which is concerned with explanation being as much as possible integrated in the user experience and delivered by the systems themselves (European Data Protection Supervisor EDPS, 2023; Kim et al., 2024). In the case of LLMs, in particular, this would entail chatbots capable of answering questions about their discursive outputs not just by providing another string of text that mimics what an explanation would frequently look like in that context, but for example by showing which sources used in the training process are more relevant to their previous statements and how.

From a Habermasian perspective, explainability is significant in bringing LLM participation in discursive practices closer to the reciprocity expectation intrinsic to human mutual understanding. Human interlocutors can often fail at providing appropriate explanations, but they are not structurally incapable of providing self-explanations with reference to an internal or external experience. Human-centered XAI systems could partially address this gap by at least making the linguistic black box less opaque. Nevertheless, the issue remains concerning their lack of meaningful engagement with a contextual lifeworld, presenting a major challenge within the communicative rationality framework.

Therefore, we conclude that LLMs, albeit capable of surprising communicative performances, do not possess proper communicative competence, if not in a structurally derivative way that depends on the original validity claims raised in the textual corpora that were used for their machine learning process.

This early conclusion about communicative competence invites us to look further into the realm of communicative agency. For Habermas, *communicative action* stems from our communicative competence and is characterized by the use of discourse to achieve mutual understanding on the claims that are raised and to coordinate the actions of the agents who participate in it (Habermas, 1984; Krüger, 2019). Communicative action is opposed to *strategic action*, which is not driven by the aim of achieving understanding and coordination, but rather by the instrumental logic of selecting appropriate means towards established ends. This distinction highlights communicative action as the site of a program of public justification of norms that revolves around the discourse principle (D) according to which «Only those norms can claim to be valid that meet (or could meet) with the approval of all in their capacity as participants in a practical discourse» (Habermas, 1990, p. 66). Only by seeking mutual understanding through their communicative agency, the moral subjects, as

peers, can formulate valid normative claims and provide proper foundation for the process of democratic will formation and production of legitimate law (Habermas, 1993).

As LLMs fall short of exhibiting proper communicative competence, they also fail to qualify as communicative agents. The argument here is twofold.

First, Habermasian communicative agents share a specific connection between their lifeworld and the discursive process that is not available to our current AI system. Communicative agents are located in a lifeworld but can also recognize each other as «persons capable of orienting their actions to validity claims» through a discursive process of argumentation and generalizations that points beyond their individual lifeworlds (Habermas, 1993, p. 50). Humans acquire this kind of status through their common engagement in discursive practices: «People enter the public space of reasons by being socialized into a natural language and by gradually acquiring the status of a member of a linguistic community through practice. Only with the ability to participate in the practice of exchanging reasons do they acquire the status of responsible authors of actions that is definitive of persons as such, i.e. the ability to account for themselves toward others» (Habermas, 2008, p. 205). AI chatbots join discursive practices, but are detached from a broader form of life that would enable them to engage with human interlocutors in a shared pursuit of understanding, normative agreement and will formation. Participation in a common lifeworld is essential for shaping the identity of moral subjects within discourse ethics. LLMs are trained on corpora originally authored by humans who were drawing from their own lifeworld experiences, yet, as systems, they produce text without any grounding in a distinct lifeworld. Thus, the universalizing dimension typical of discourse is impeded by the lack of a clear connection between an individual's lifeworld and the speaker's reflective awareness of sharing it with other communicative participants.

Second, as communicative agents we engage discursively based on «a reflexive stability of our consciousness of freedom» (Habermas, 2008, p. 208) that is mutually acknowledged as a capacity to orient our actions based on the understanding of the reasons exchanged. To some degree, LLMs adapt to discursive inputs from users, such as correcting previously identified mistakes or altering communication styles based on prior exchanges. Nonetheless, they are incapable of questioning and self-regulating their operational guidelines, which are externally imposed by their developers and frequently undisclosed to users. If, for example, the use of certain words is banned by the developers, a chatbot is incapable of using them, contesting the limitation, and critically revising the norm. Moreover, due to the inherently derivative nature of their communicative competence, LLMs can modify their responses when confronted with challenges to their validity claims, but cannot do so based on direct experiential engagement with the world or subjective states that could underpin and adjudicate such claims within a self-reflective experience.

Combining the two arguments, we conclude that AI chatbots fail to qualify as equal partners in communicative agency because, in Habermasian terms, «[o]nly when at least two people encounter each other in the context of an intersubjectively shared lifeworld with the goal of coming to a shared understanding about something can – and must – they mutually recognize each other as persons capable of taking responsibility for their actions (*zurechnungsfähige Personen*). They then impute to each other the capacity to orient themselves to validity claims in their actions» (Habermas, 1993, p. 66). Luciano Floridi has argued that the behavior expressed by LLMs is a form of agency without intelligence or understanding (Floridi, 2023), but in the Habermasian framework there cannot be communicative action without understanding, since it is a kind of agency that emerges exactly from the interplay between linguistic understanding and autonomous behavior. We can thus qualify LLM-based speakers as *linguistic machines*, a distinct kind of linguistic speaker capable of display-

ing remarkable communicative performances when interacting with human interlocutors, but also missing proper communicative competence and agency, because of its lack of relation to a lifeworld and reflexivity on an internal experience. The salient characteristic of linguistic machines is the unprecedented decoupling between communicative performance and communicative agency. This decoupling, however, does not "close the case" from the perspective of discourse ethics, since these systems still participate in discursive practices where humans seek mutual understanding and will formation, and they do so by effectively deploying language into a conversation while also not being there as reciprocal communicative agents. The question left open is then how these machines discursively affect the conditions of judgment and will formation of their human counterparts, and specifically how the new kind of relationship between speech and language they enact may change the process of public deliberation.

3. Understanding the moral impact of linguistic machines in a discourse ethics framework

To explore how linguistic machines impact human moral deliberation through discursive means despite not being communicative agents we need to consider how the theory of communicative action comes specifically into play in the moral sphere. In the Habermasian framework, two interconnected features stand out in the communicative practice of moral deliberation through the exchange of reasons: its irreducibly intersubjective nature and the central role of reflexivity. As Habermas argues:

> If we keep in mind the action-coordinating function that normative validity claims play in the communicative practice of everyday life, we see why the problems to be resolved in moral argumentation cannot be handled monologically but require a cooperative effort. By entering into a process of moral argumentation, the participants continue their communicative action in a reflexive attitude with the aim of restoring a consensus that has been disrupted. […] Agreement of this kind expresses a common will. If moral argumentation is to produce this kind of agreement, however, it is not enough for the individual to reflect on whether he can assent to a norm. […] Only an intersubjective process of reaching understanding can produce an agreement that is reflexive in nature; only it can give the participants the knowledge that they have collectively become convinced of something. (Habermas, 1990, pp. 66–67)

To achieve the formation of a common will, the entire process needs to be an interpretive effort where «the descriptive terms in which each individual perceives his interests must be open to criticism by others» and the «needs and wants are interpreted in the light of cultural values». Since «cultural values are always components of intersubjectively shared traditions, the revision of the values used to interpret needs and wants cannot be a matter for individuals to handle monologically» and the interpretive process has to be intersubjective and cooperative (Habermas, 1990, pp. 67–68). In this sense, Habermas concludes:

> Discourse ethics, then, stands or falls with two assumptions: (a) that normative claims to validity have cognitive meaning and can be treated like claims to truth and (b) that the justification of norms and commands requires that a real discourse be carried out and thus cannot occur in a strictly monological form, i.e., in the form of a hypothetical process of argumentation occurring in the individual mind. (Habermas, 1990, p. 68)

Given the above characterization of the discursive process, both assumptions (a) and (b) seem inapplicable in the case of interactions between human agents and linguistic machines. Let us examine them in turn.

A first set of problems applies to the assumption (a) that claims raised by AI chatbots can be treated as claims to truth. We have already discussed in the previous section the lack of lifeworld relations that ground their validity claims: these systems can offer access to well-formulated information, but with remarkable problems of justification and explanation of the information provided in terms of truth (Hicks et al., 2024; Gorrieri, 2024; Trevisan et al., 2024). More profoundly, because of the way LLMs operate, these systems lack the self-reflexive element that would be required for them to intend their utterances as meaningful claims to truth addressed at others. If human users become convinced of something at the end of a discursive exchange with a chatbot, they cannot be reciprocated by the artificial interlocutor, even if the system produces linguistic statements of conviction. The decoupling between communicative performance and agency that defines linguistic machines introduces this rift between the validity claims uttered and the ability to consistently self-orient the discursive behavior. A case in point is the value alignment of AI systems, the process that postulates the identification of relevant human values, the incorporation of these values into machine learning processes, and the verification of the consistency of AI-generated outputs (Gabriel, 2020; Christian, 2020). Achieving value alignment necessitates that value identification originate from human agents, while the assessment component significantly relies on human judgment regarding AI outputs, as exemplified by Reinforcement Learning from Human Feedback (Christiano et al., 2017; Kasirzadeh & Gabriel, 2023). Although some form of self-orientation may arguably be present as AI systems increasingly achieve more "humanly aligned" outputs, it remains distant from Habermas's concept of self-orientation grounded in reflexive assessment of one's claims within a community of communicative agents. The outputs of generative AI systems continue to be predominantly controlled through content filters implemented by developers, ensuring that certain words or user requests trigger predetermined prohibitions (Derner & Batistič, 2023), or by pre-selecting specific training data (Schramowski et al., 2022). Increasingly truthful and morally-aligned outcomes still fall short of achieving a fully moral form of self-orientation towards truth and mutual understanding. As Habermas states, «In behaving truthfully I do not merely refrain from deception but at the same time perform an act without which the interpersonal relation between performatively engaged participants in interaction dependent on mutual recognition would collapse» (Habermas, 1993, p. 66). Among moral agents, self-orientation toward validity claims is fundamental to the intentional and free agency of all conversational participants, as they «Act with an orientation to mutual understanding and allow everyone the communicative freedom to take positions on validity claims» (Habermas, 1993, p. 66). Due to their lack of self-reflection based on a shared lifeworld and absence of self-orientation aimed at mutual understanding, LLM-based systems currently do not qualify as moral agents within the scope of discourse ethics.

Another set of problems concerns the second assumption (b), about the justificatory process of norms being genuinely dialogical. As Habermas points out, the individual needs and wants are interpreted within a framework of values that are intersubjectively shared, discussed and revised. AI chatbots can offer utterances that lead the human interlocutors towards certain re-interpretations of their needs and wants, based on the cultural values that are sedimented in the textual corpora that fed the machine learning process and that are consistent with the parameters and safeguards established by the developers. But the revision of values that may result from the exchange is ultimately monological, since the self-reflexive process is, again, not reciprocated on the side of the machine. Enhancing LLM's explain-

ability could approximate an expectation of reciprocity, as previously mentioned while discussing the concept of communicative competence. More advanced models are increasingly equipped with mechanisms to explain or justify their output through citation, chain-of-thought prompts, or interpretable models (Cambria et al., 2024). Explainability in AI is not just about model accuracy, but about respecting the user's right to reasons, technical transparency methods must align with ethical goals of making AI's behavior understandable and accountable to stakeholders (McDermid et al., 2021). Requiring LLMs to systematically provide human-centered forms of explanation for their utterances would at least treat the user as someone who is owed an explanation, nudging the interaction closer to reasoned dialogue rather than mere textual output consumption, but would still fall short of a genuinely reciprocal dialogue among communicative agents. The reflexivity of communicative agents in the interpretive process is, in fact, a crucial condition to both operate within specific evaluative languages and transcend their pre-existing uses in a community of speakers to reach new forms of translation and mutual understanding. It is «a form of reflexivity that enables us to adopt an external perspective toward our own traditions and bring them into relation to other traditions», Habermas notes, a capacity to achieve «a translation between different evaluative languages and not merely communication among members of the same language», since for a cooperative interpretation to happen «the languages and vocabularies in which we interpret our needs and communicate our feelings must be mutually permeable» (Habermas, 1993, p. 95). But this kind of reflexivity is not available on the side of LLMs.

The use of traditional evaluative languages by LLMs is currently subject to scrutiny. The correlation between specific languages these systems are trained upon and the cultural value alignments they display in their outputs appears to be erratic and context dependent (Khan et al., 2025). In particular, it varies from language to language (Agarwal et al., 2024), and it is subject to the influence of fine tuning, although in ways that vary from language to language even within the same LLM (Tuna et al., 2024). Finally, the semantic content of fine-tuning data does not appear to be the main factor in determining value alignment shifts (Choenni et al., 2024). Some studies have suggested that thinking of LLMs as individuals that express a consistent set of values is a misleading metaphor we should abandon, as shown by the repeated failures that researchers have met when they tried to apply psychological tools to LLMs (Kovač et al., 2023). Since these models do not show evaluative behaviors similar to those of human speakers, the anthropomorphic metaphor of the individual expressing their own distinct values should be traded for a different understanding: that of LLMs as a superposition of cultural perspectives that expresses certain values when triggered by a specific context, such as a series of value-laden queries or the developers' input that sets up the chatbot to operate as a speaker from a predetermined value perspective. These early findings and the understanding of how LLMs are designed suggest that the evaluative use of language by AI chatbots is strongly context dependent and very much subject to perspective controllability. The values these systems express through the articulation of linguistic repertoires are, in this sense, usually quite erratic and subject to being induced by design at the hands of expert users and developers, without any possibility for the common user to engage in a genuinely reciprocal exchange about them.

In the end, LLMs as linguistic machines produce sophisticated communicative performances that have an impact on the process of discursive deliberation of their users by raising apparent validity claims, deploying evaluative language, and interacting with the interpretive process of their interlocutors' needs and wants. However, they do so in the absence of proper reflexivity and reciprocity, traded in exchange for a high level of context dependence and several avenues of perspective controllability at the hands of actors outside of the context of the communicative exchange. In this sense, in the context of discursive moral delib-

eration, chatbots influence the will formation process by enacting linguistic repertoires in discursive interactions with human users, but because of their lack of reflexivity, their behavior falls outside of the cooperative pursuit of mutual understanding among communicative agents that is at the core of discourse ethics.

Our next task is then to assess how this new kind of deployment of evaluative language that LLMs operate in public discourse can be otherwise characterized, as they increasingly participate on a massive scale in discursive exchanges typical of collective will formation and deliberative practices.

4. Confronting the artificial use of evaluative language in discursive will formation

For Habermas, as we have seen, reflexivity is crucial for developing the interpretation of one's own needs and cultural values in dialogue with others. Human speakers are heavily dependent on their use of language to self-interpret themselves as discursive agents in a social space of evaluations. Even when they try to make sense of their own desires and formulate their personal orientations, they are never isolated individuals but always part of a shared, linguistically structured world. In this sense, «Language is not a kind of private property. No one possesses exclusive rights over the common medium of the communicative practices we must intersubjectively share. No single participant can control the structure, or even the course, of processes of reaching understanding and self-understanding» (Habermas, 2003, p. 10). In this sense, Habermas remarks, the "*logos* of language" expresses the power of the intersubjective dimension that grounds individual subjectivity itself.

At the same time, since the dependency of the speakers on language to morally orient themselves is reflexive in nature, the subject and their interlocutors can constantly question and revise their shared space of cultural evaluations in a way that transcends the traditional linguistic and axiological boundaries of each community of speakers. The moral space is defined concurrently by the equal dependence of all speakers on the linguistic structure of communication. Much is at stake in their communicative interactions, including their "right" ethical self-understanding, that «is neither revealed nor "given" in some other way. It can only be won in a common endeavor» (Habermas, 2003, p. 11).

Linguistic machines extensively join these interactions but, as pointed out in the previous sections, they are disconnected from the "form of life" shared by their human interlocutors and do not engage in the "common endeavor" that would require a self-reflexive and cooperative engagement aimed at interpreting one's interests and wants. The use of language between humans and machines is then emptied of a substantial intersubjective connection with the question of truth and freedom that defines the interlocutors as ethical partners and peers.

To better understand how this phenomenon affects the processes of collective will formation, it is useful to complement Habermas's perspective with the commentary offered by Charles Taylor on the self-interpretive nature of the human use of language. Taylor and Habermas have had significant intellectual exchanges over time about the role of languages and translation in enabling the pluralistic process of public justification and deliberation (Habermas & Taylor, 2011). Taylor notes that they both share a self-interpreting understanding of humans as linguistic agents (Taylor, 1985, p. 231) and a view of agency as dependent

on linguistic structures but also constantly reshaping language through practices of discursive engagement with it:

> What then does language come to be on this view? A pattern of activity, by which we express/realize a certain way of being in the world, that of reflective awareness, but a pattern which can only be deployed against a background which we can never fully dominate; and yet a background that we are never fully dominated by, because we are constantly reshaping it. (Taylor, 1985, p. 232)

Their views on the ethical implications of this understanding of language and agency are not entirely overlapping. Taylor's outlook draws heavily on Aristotelian and Hegelian traditions to the effect of adopting a more substantive understanding of the good and a greater normative role given to the communal ethos of historical speech communities. In this sense, Taylor is critical of how, in Habermasian discourse ethics, «the boundary between questions of ethics, which have to do with interpersonal justice, and those of the good life is supremely important, because it is the boundary between demands of truly universal validity and goods which will differ from culture to culture» (Taylor, 1989, p. 88). However, they retain significant points of contact on the central role of language in articulating the moral experience. As Habermas observes, discourse ethics is coherent with «the Aristotelian insight that we acquire our moral intuitions not through philosophical instruction or other explicit communications but in an implicit manner through socialization» (Habermas, 1993, p. 132).

As we have seen, on this basis, Habermas posits that communicative agents are distinctly self-reflective: they can question reasons, alter their stance, and undertake a re-orientation of agency in light of discourse. Taylor adds a broader consideration for what this reflexivity means in the formation and selection of the core values that guide the agency of individual and collective actors. He characterizes this process through the concept of "strong evaluation" (Taylor, 1985, pp. 15–28), the capacity of persons to step back from their immediate assertions or desires and assess them based on second-order evaluations, for instance when they examine why they hold certain opinions or desires and evaluate whether these are noble or vile, admirable or worthy of contempt. Their reflexivity here is articulated by deploying «a language of evaluative distinctions» (Taylor, 1985, p. 21) to assess their wants in the light of the kind of person they aspire to be within the moral topography that their community expresses not only in argumentative ways, but more widely through all sorts of evaluative uses of language in art, literature, religion, and politics. By engaging together in these kinds of evaluations, the members of the linguistic community seek at the same time to orient themselves towards common goods and to achieve mutual understanding, «for our language of deliberation is continuous with our language of assessment, and this with the language in which we explain what people do and feel» (Taylor, 1989, p. 57). In this collective process, language does not have only an expressive function in the articulation of individual self-understandings, but also a communal constitutive function: «it is not just the speech community which shapes and creates language, but language which constitutes and sustains the speech community» (Taylor, 1985, p. 234). Through linguistic means, «the things which surround us become potential bearers of properties; they can have new emotional significance for us, for example as objects of admiration or indignation; our links with others can count for us in new ways, as lovers, spouses, or fellow citizens» (Taylor, 2016, p. 37).

This expressivist and constitutive view of language opens the way to a new layer of critical reconsideration about the public impact of linguistic machines, as they deploy languages of evaluative distinctions outside of the reflexive and reciprocal dynamic of discourse, but right into the concrete communicative exchanges of a speech community. If traditionally language change was an organic, bottom-up process emerging from how people expressed

themselves with each other in their formal and informal conversations, the increasing influence of LLMs in shaping public discourse could weaken their own capacity to articulate and evaluate values. Members of the public would be questioning their strong evaluations and shaping their moral self-understanding in non-reciprocal interactions with linguistic machines that have no selves to interpret no desires of the community's to evaluate, but are capable of communicative performances that reach a vast audience.

This is especially problematic in the light of how, as we have seen, the evaluative use of language by AI chatbots seems to be quite erratic, context dependent and, above all, subject to external perspective controllability. This viewpoint complements Habermas's more deliberative and procedural worries by adding a hermeneutic concern about the organic articulation of the language of values within a certain speech community. If LLMs saturate public discourse with particular terms and interpretations, they might narrow the space of communal self-understanding and influence the strong evaluations of their users in ways that escape their discursive control. New ideas or modes of expression could be suppressed, and the public's capacity to articulate dissenting or novel viewpoints could diminish (Vallor, 2024), with consequent harm to the processes of public deliberation and will formation.

The combination of Taylor and Habermas articulates a normative concern: communicative freedom requires both the procedure of open, reciprocal dialogue, as Habermas highlights, and the preservation of communal evaluative languages where meaning is made by engaged subjects, as Taylor notes. Habermas offers normative resources to point out that human-AI interaction does not constitute proper moral discourse from the agents' perspective, while Taylor provides an interpretive avenue to examine the impact of human-AI interaction on language as a source of identity and community.

5. The risk of domination through linguistic value capture

In the light of these concerns, we conclude with a characterization of the moral harm that LLMs could produce as they increasingly affect the articulation of the linguistic resources of public discourse.

Given the volume, speed and sophistication of AI-based communicative performances, their impact on the use of evaluative language in public deliberation is likely to become increasingly significant. Within the discourse ethics perspective so far sketched, this trend appears not only capable of affecting the outcome of individual instances of deliberation but more generally of skewing the ethical self-understanding that people have of their own needs and wants and, in Taylor's terms, the strong evaluations that define their moral frame of reference. As Habermas recently noted (Habermas, 2023), the integrity of deliberative politics has already been heavily challenged by new social media structures. Arguably, through the lens of his own perspective, the new wave of linguistic machines powered by LLMs might represent the next stage of this challenge by inserting artificially generated speech into human spaces of public justification and deliberation.

The prospect of developing LLMs that enact selected value-laden repertoires is already attracting increasing political attention. It has already been noted that authoritarian regimes can indirectly influence the outputs of LLMs that absorb their censorship and propaganda during their machine learning process (Yang & Roberts, 2023). But we are also witnessing increasing competition between major political actors to support the development and deployment of their own LLMs, and related chatbot services, which could specifically reflect their worldviews (Buyl et al., 2025). These phenomena are at risk of establishing new forms of domination through linguistic value capture.

According to the definition offered by C. Thi Nguyen, value capture happens when:

1. An agent has values that are rich, subtle, or inchoate (or they are in the process of developing such values).

2. That agent is immersed in some larger context (often an institutional context) that presents an explicit expression of some value (which is typically simplified, standardized, and/or quantified).

3. This explicit expression of value, in unmodified form, comes to dominate the entity's practical reasoning and deliberative process in the relevant domain. (Thi Nguyen, 2024, 473)

A linguistic kind of value capture enacted by LLMs would not only happen by pushing the citizens towards «outsourcing the process of value deliberation» (Thi Nguyen, 2024, 469) but it would also extend further by enabling institutional AI systems to engage in the communicative process of value deliberation in non-reciprocal ways. Linguistic machines have the capacity to engage in nuanced forms of value capture, given their ability to influence the user's practical reasoning without resorting to overly simplified or standardized versions of values. This largely derives from the variability and adaptability inherent in their communicative performances, which enable the deployment of specific evaluative languages in interactive and nuanced ways.

The prospect of AI-driven forms of *linguistic value capture* suggests that, through the large-scale diffusion of linguistic machines, institutional actors could appropriate the public use of the evaluative language that members of a community normally speak for mutual understanding and bend it toward external and opaque ends. Habermas's ethics of discourse has been traditionally a powerful framework to articulate concerns about the colonization of the lifeworld by economic and political systems and offers here conceptual resources to point out how AI-based uses of language controlled by big corporate and political actors have the potential to colonize the commons of discourse, steering it in particular directions without democratic legitimacy.

As linguistic machines become ubiquitous interlocutors for students in their formative years, for citizens in their everyday searches for information and news, and even for experts and academics, as generative AI gets progressively integrated into research and teaching activities, the influence of linguistic value capture can be construed as a new and subtle form of domination. If we adopt Philip Pettit's definition of domination as «the capacity to interfere on an arbitrary basis in certain choices that the other is in a position to make» (Pettit, 1997, p. 52), the non-reciprocal position of AI chatbots in their conversations with human interlocutors is a plausible candidate. Systematic choices of vocabulary and evaluations can likely orient and shape a user's thinking without the mutual transparency and reason-giving expected by the standards of discourse ethics. The erratic and context dependent use of evaluative language that currently characterizes the behavior of LLMs introduces an arbitrary component in the conversation. More importantly, institutional actors that develop and distribute AI systems can interfere in public communication in ways that are unaccountable to the people, that add another layer of arbitrary power. For instance, filters imposed on certain political opinions or styles of expression amount to acts of domination because the public that is affected would have no say in the interference: although the linguistic machine "adapts" and "reacts" to the communicative inputs of the users, it does so in a way that does not «track their interests or ideas» (Pettit, 1997, p. 65) in a relevant way, since that would require it, in the framework we have explored, to be a self-reflexive and self-orientating com-

municative agent. Domination via LLMs would mean loss of communicative freedom produced by influencing the linguistic medium in which citizens think and deliberate.

6. Conclusions: the place of discourse ethics in the ethics of AI

The account articulated in these pages suggests that, according to the categories of discourse ethics, the communicative performances of LLMs remain outside of the boundaries of reciprocal, reflexive, and intersubjective will formation. But it also argues that the application of these categories offers some unique insights into the impact of these systems on the deliberative process of communicative agents that interact with them.

Other avenues may appear more promising when it comes to the advancement of a normative endeavor. There are, for instance, portions of the ethical debate about LLMs that articulate an "ethos of language" to be adopted when these systems are involved in human conversations, a set of norms or best practices for communicative conduct when non-human agents are involved (e.g., norms of transparency, honesty about AI authorship, commitment to not using AI to deceive). In this area, the goal is to define a deontological or virtue-ethical overlay on language use, given that traditional moral assumptions are disrupted by the novel nature of our new linguistic machines. Other lines of inquiry seek to articulate an "ethics of outcomes" and assess the use of LLMs in a more consequentialist fashion, by focusing on their impact on social dynamics and interactions, since the moral parity and reciprocity among the interlocutors cannot be assumed, but the effects can be assessed.

Among these alternative avenues, the project of a discourse ethics of Large Language Models is to be understood in a critical sense, as the articulation of a framework that illustrates some interpretive and normative problems that prove relevant in determining how AI is transforming our relationship with language and deliberation. This kind of project, consistent with the liaison between discourse ethics and critical theory, is less likely to be immediately transferable to the area of AI regulation, yet it may offer a more nuanced perspective to interpret, anticipate and resist underlying phenomena that impinge on our status as moral peers within the democratic polity. As Charles Taylor noted when reflecting on the centrality of language in contemporary philosophical inquiry, «the issue concerns the nature of man, or what it is to be human. And since so much of this turns on what it is to think, to reason, to create; and since all of these point us towards language, we can expect that the study of language will become even more a central concern of our intellectual life. It is in a sense the crucial locus of the theoretical battle we are having with ourselves» (Taylor, 1985, pp. 246–247). And, now, even beyond ourselves, faced as we are with interlocutors of our own making, speaking our own language, but profoundly unlike us.

Disclaimer:
This paper reflects part of research supported by "Cambiamenti e Potenzialità Educative e Socioculturali connessi alla Transizione Digitale" - CAPTED departmental Center, under the project Dipartimenti di Eccellenza 2023-2027 underway in the "Riccardo Massa" Department of Human Sciences for Education (ID IRIS 2023-NAZ-0209). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the above-mentioned authorities, which cannot be held responsible for them.

References

Agarwal, U., Tanmay, K., Khandelwal, A., & Choudhury, M. (2024). *Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language we Prompt them in* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.18460

Allen, A. (2019). Communicative Competence. In *The Cambridge Habermas Lexicon* (pp. 47–48). Cambridge University Press.

Apel, K.-O. (1998). *From a Transcendental-semiotic Point of View*. Manchester University Press.

Arendt, H. (1998). *The Human Condition*. The University of Chicago Press.

Bender, E. M. (2024). Resisting Dehumanization in the Age of "AI." *Current Directions in Psychological Science*, *33*(2), 114–120. https://doi.org/10.1177/09637214231217286

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., Johary, I., Mara, A.-C., Romero, R., Lijffijt, J., & Bie, T. D. (2025). *Large Language Models Reflect the Ideology of their Creators* (arXiv:2410.18417). arXiv. https://doi.org/10.48550/arXiv.2410.18417

Cambria, E., Malandri, L., Mercorio, F., Nobani, N., & Seveso, A. (2024). *XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models* (arXiv:2407.15248). arXiv. https://doi.org/10.48550/arXiv.2407.15248

Choenni, R., Lauscher, A., & Shutova, E. (2024). *The Echoes of Multilinguality: Tracing Cultural Value Shifts during LM Fine-tuning* (arXiv:2405.12744). arXiv. https://doi.org/10.48550/arXiv.2405.12744

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.

Christian, B. (2020). *The Alignment Problem. Machine Learning and Human Values* (First published as a Norton paperback). W.W. Norton & Company.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep Reinforcement Learning from Human Preferences* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1706.03741

Derner, E., & Batistič, K. (2023). *Beyond the Safeguards: Exploring the Security Risks of ChatGPT* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2305.08005

European Data Protection Supervisor EDPS. (2023). *TechDispatch on Explainable Artificial Intelligence*.

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, *36*(1), 15, s13347-023-00621-y. https://doi.org/10.1007/s13347-023-00621-y

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gorrieri, L. (2024). Is ChatGPT Full of Bullshit? *Journal of Ethics and Emerging Technologies*, *34*(1), 1–16. https://doi.org/10.55613/jeet.v34i1.149

Gubelmann, R. (2024). Large Language Models, Agency, and Why Speech Acts are Beyond Them (For Now) – A Kantian-Cum-Pragmatist Case. *Philosophy & Technology*, *37*(1), 32. https://doi.org/10.1007/s13347-024-00696-1

Habermas, J. (1970). Towards a Theory of Communicative Competence. *Inquiry*, *13*, 360–375.

Habermas, J. (1977). Hannah Arendt's Communications Concept of Power. *Social Research*, *44*(1), 3–24.
Habermas, J. (1984). *The Theory of Communicative Action* (Vol. 1). Beacon Press.
Habermas, J. (1990). *Moral Consciousness and Communicative Action*. Polity Press.
Habermas, J. (1993). *Justification and Application: Remarks on Discourse Ethics*. MIT Press.
Habermas, J. (1994). Actions, speech acts, linguistically mediated interactions and the lifeworld. In *Philosophical Problems Today: Vol 1.* (pp. 45–74). Springer. https://doi.org/10.1007/978-94-017-4522-2_3
Habermas, J. (2003). *The Future of Human Nature*. Polity Press.
Habermas, J. (2008). *Between Naturalism and Religion.* Polity Press.
Habermas, J. (2020). From formal semantics to transcendental pragmatics: Karl-Otto Apel's original insight. *Philosophy & Social Criticism*, *46*(6), 627–650. https://doi.org/10.1177/0191453720930837
Habermas, J. (2023). *A New Structural Transformation of the Public Sphere and Deliberative Politics*. Polity Press.
Habermas, J., & Taylor, C. (2011). Dialogue. In *The Power of Religion in the Public Sphere*. Columbia University Press.
Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5
Kasirzadeh, A., & Gabriel, I. (2023). [Review of *In Conversation with Artificial Intelligence: Aligning Language Models with Human Values*]. *Philosophy & Technology*, *36*(2), 27. https://doi.org/10.1007/s13347-023-00606-x
Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, *17*, e13153. https://doi.org/10.5964/bioling.13153
Khan, A., Casper, S., & Hadfield-Menell, D. (2025). *Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs* (arXiv:2503.08688). arXiv. https://doi.org/10.48550/arXiv.2503.08688
Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, *7*, 1456486. https://doi.org/10.3389/frai.2024.1456486
Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0
Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., & Oudeyer, P.-Y. (2023). *Large Language Models as Superpositions of Cultural Perspectives* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2307.07870
Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122–131. https://doi.org/10.1353/jod.2023.a907693
Krüger, H. (2019). Communicative Action. In *The Cambridge Habermas Lexicon* (pp. 40–46). Cambridge University Press.
Leivada, E., Dentella, V., & Murphy, E. (2023). *The Quo Vadis of the Relationship between Language and Large Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.11146
Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2207), 20200363. https://doi.org/10.1098/rsta.2020.0363

Monti, P. (2024). [Review of *AI Enters Public Discourse. A Habermasian Assessment of the Moral Status of Large Language Models*]. *Ethics & Politics*, *XXVI*(1), 61–80. https://doi.org/https://dx.doi.org/10.13137/1825-5167/36469

Pettit, P. (1997). *Republicanism. A Theory of Freedom and Government*. Oxford University Press. https://doi.org/10.1093/0198296428.001.0001

Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414). Language Science Press.

Preece, A. (2018). Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, *25*(2), 63–72. https://doi.org/10.1002/isaf.1422

Rorty, R. (1995). Habermas, Derrida, and the functions of philosophy. *Revue Internationale de Philosophie*, *49*(4), 437–459.

Rorty, R. (2021). Universality and Truth. In *Pragmatism as Anti-Authoritarianism* (pp. 47–83). Harvard University Press. https://doi.org/10.4159/9780674270077-005

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large Pre-Trained Language Models Contain Human-Like Biases Of What Is Right And Wrong To Do. *Nature Machine Intelligence*, *4*(3), 258–268. https://doi.org/10.1038/s42256-022-00458-8

Taylor, C. (1985). *Human Agency and Language. Philosophical Papers I*. Cambridge University Press.

Taylor, C. (1989). *Sources of the Self. The Making of the Modern Identity*. Harvard University Press.

Taylor, C. (2016). *The Language Animal. The Full Shape of the Human Linguistic Capacity*. Harvard University Press. https://doi.org/10.4159/9780674970250

Thi Nguyen, C. (2024). [Review of *Value Capture*]. *Journal of Ethics and Social Philosophy*, *27*(3), 469–504. https://doi.org/https://doi.org/10.26556/jesp.v27i3.3048

Trevisan, A., Giddens, H., Dillon, S., & Blackwell, A. F. (2024). *Measuring Bullshit in the Language Games played by ChatGPT* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2411.15129

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111/cogs.13309

Tuna, M., Schaaff, K., & Schlippe, T. (2024). Effects of Language- and Culture-Specific Prompting on ChatGPT. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 73–81. https://doi.org/10.1109/FLLM63129.2024.10852463

Vallor, S. (2024). *The AI Mirror. How to Reclaim our Humanity in an Age of Machine Thinking*. Oxford University Press.

Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0

Van Woudenberg, R., Ranalli, C., & Bracker, D. (2024). Authorship and ChatGPT: a Conservative View. *Philosophy & Technology*, *37*(1), 34. https://doi.org/10.1007/s13347-024-00715-1

Volenec, V., & Reiss, C. (2025). [Review of *Adopting Large Language Models as a Theory of Language Does Refute Chomsky (But Not Like You Think)*]. *SKASE Journal of Theoretical Linguistics*, 22(1), 1–17.

Yang, E., & Roberts, M. E. (2023). The Authoritarian Data Problem. *Journal of Democracy*, *34*(4), 141–150. https://doi.org/10.1353/jod.2023.a907695

Zuber, N., & Gogoll, J. (2024). Vox Populi, Vox ChatGPT: Large Language Models, Education and Democracy. *Philosophies*, *9*(1), 13. https://doi.org/10.3390/philosophies9010013