Etica & Politica / Ethics & Politics, XXVI, 2024, 1, pp. 61-80
ISSN: 1825-5167

AI ENTERS PUBLIC DISCOURSE: A HABERMASIAN ASSESSMENT OF THE MORAL STATUS OF LARGE LANGUAGE MODELS

PAOLO MONTI
Università degli Studi di Milano Bicocca
Dipartimento di Scienze Umane per la Formazione "Riccardo Massa"
paolo.monti@unimib.it

ABSTRACT
Large Language Models (LLMs) are generative AI systems capable of producing original texts based on inputs about topic and style provided in the form of prompts or questions. The introduction of the outputs of these systems into human discursive practices poses unprecedented moral and political questions. The article articulates an analysis of the moral status of these systems and their interactions with human interlocutors based on the Habermasian theory of communicative action. The analysis explores, among other things, Habermas's inquiries into the analogy between human minds and computers, and into the status of atypical participants in the linguistic community such as genetically modified subjects and animals. Major conclusions are the LLMs seem to qualify as authors that originally participate in discursive practices but do display only a structurally derivative form of communicative competence and fail to meet the status of communicative agents. In this sense, while the contribution of AI writing systems in public discourse and deliberation can support the process of mutual understanding within the community of speakers, the human actors involved in the development, use, and diffusion of these systems share a collective responsibility for the disclosure of AI authorship and verification and adjudication of validity claims.

KEYWORDS
Large Language Models, Jürgen Habermas, Moral status, Responsibility, Public discourse

1. INTRODUCTION: AI HAS ENTERED THE CHAT

Generative AI systems are becoming increasingly effective at producing original texts based on inputs about topic and style provided in the form of prompts or questions. Latest AI technologies, especially Large Language Models (LLMs) like GPT-4 by OpenAI or PaLM 2 by Google, develop their capabilities through a machine learning process that feeds on fragments of public discourse as found in internet webpages, books, and articles. These systems have become increasingly successful at engaging in areas of complex and specialized writing, like poetry or academia, with results sometimes indistinguishable from those of human writers.

The diffusion of AI-generated discourse into the public sphere poses serious and unprecedented normative questions. Unlike other uses of AI technology, such as "deepfakes" – fabricated videos representing public figures in the act of saying words they never pronounced – AI writing cannot be reduced to a mere instance of forgery operated by human actors with the instrumental assistance of AI-based tools. Instead, it highlights the possibility of important pieces of the public conversation being originated by non-human authors and finding their way into moral-practical discourses about principles, norms, and policies that hold a central place in democratic deliberation among citizens.

Some political uses of LLMs are already in active development, from more experimental exercises, like the production of "toxic" models trained on data from unregulated internet message boards to more systematic analyses of the efficacy of microtargeted political messages written by LLMs and individually directed to social media users.

The problem of the moral status of LLMs as potential participants in public discourse can be fruitfully approached from different philosophical and STS perspectives (Gordon and Gunkel 2021; Sinnot-Armstrong and Conitzer 2021; Redaelli 2023). The limited scope of this article aims to highlighting which insights can be drawn from Habermasian theory and what status can be assigned to LLMs that participate in discursive practices with humans in terms of responsibility for what they generate in that context. In recent years, Jürgen Habermas has discussed some of the implications of the new technological infrastructure of communication based on the internet and social media for the public sphere and deliberative democracy (Calloni et al. 2021; Habermas 2023). He has not substantially engaged, on the other hand, with the possibility that digital technologies could soon also produce a new kind of non-human actors of public discourse and deliberation. His vast philosophical project, however, offers relevant conceptual resources to attempt this undertaking as well. This account begins by looking at two areas, mutually connected but articulated in Habermas's works at different times: first, the tension between the communicative origin of the person and the naturalistic understanding of the mind as a computer (2); second, the moral status of atypical members of the community of communicants like genetically modified individuals and animals (3). We will explore these two areas, to then attempt a characterization of the hybrid status of LLMs within our discursive practices (4) and outline a preliminary normative account of the moral responsibilities at play when fragments of discourse produced by LLMs enter public discourse and deliberation (5). The conclusions will briefly discuss how this account may fit within a larger consideration of the future impact of generative AIs on the ethics of democratic citizenship (6).

2. BETWEEN HUMAN MINDS AND MACHINE LEARNING

In *Between Naturalism and Religion*, Habermas argues that the social formation of the person through the practice of exchanging reasons with peers seems irreducible to the merely naturalistic understanding of the mind that is suggested by the frequently advanced analogy between human minds and computers. The genesis of the human mind, Habermas notes, lies in the interplay between «the perspective of an observer on what is going on in the world with the perspective of a participant in interaction» with others (Habermas 2008: 171). The "subjective mind" of the individual arises within a communicative process of understanding that constantly generates, in parallel, a linguistic "objective mind" of materially embodied symbols that is to some extent independent of its individual speakers. These subjective and objective sides of the human mind are both distinct and co-implicated. Distinct since, «On the one hand, objective mind evolved out of the interaction between the brains of intelligent animals who had already developed the capacity for reciprocal perspectivetaking [...] On the other hand, the "objective mind" claims relative independence vis-à-vis these individuals, since the universe of intersubjectively shared meanings, organized according to its own grammar, has taken on symbolic form» (Habermas 2008: 174-175). These "two minds" are, however, also tightly co-implicated, since:

> These meaning systems can, in turn, influence the brains of participants through the grammatically regulated use of symbols. The "subjective mind" of those individuated participants in shared practices develops only in the course of the socialization of their cognitive capacities. This is what we mean by the self-understanding of a subject who can step into the public space of a shared culture. As actors, they develop the awareness of being able to act one way or another because they are confronted in the public space of reasons with validity claims that challenge them to take positions.

Our self-understanding as free subjects emerges out of this interplay between the subjective and the objective mind, since «conscious participation in the symbolically structured "space of reasons" jointly inhabited by linguistically socialized minds is reflected in the accompanying performative sense of freedom» (Habermas 2008: 173). As rational subjects, our agency finds motivations «in this dimension and follows logical, linguistic, and pragmatic rules that are not reducible to natural laws» (Habermas 2008: 173). This opens up the possibility of separating the kind of causal connection that the naturalistic image of the world envisions between the individual brain and its corresponding individual mind, from a distinct form of "mental causation" that arises from the cognitive inputs that the symbolic "objective mind" feeds to the individual by stimulating judgments and considerations.

Habermas is intent in specifying that these two understandings of the life of the mind are not mutually exclusive but cannot, at the same time, be entirely reduced to the naturalistic side. They are rather the outcome of an inescapable linguistic

dualism between the perspective of the observer, as reflected in the scientific outlook, and the perspective of the participant articulated in our practical understanding of the mind. In this sense, he argues, the computer analogy that is often invoked to assimilate our thinking to the inner workings of computing machines is fundamentally flawed because it misses «the socialization of cognition that is peculiar to the human mind» (Habermas 2008: 175).

This brings us closer to our first step in the assessment of the moral status of generative AI systems, especially LLMs. Habermas is stark in remarking that the intersubjective, symbolic experience that animates the human mind is irreducible to the image of software running on computer hardware. At the same time, he does not rule out entirely the ICT analogy, as he notes:

> Talking of the mind "programming" the brain evokes metaphors from computer language. The computer analogy puts us on the wrong track insofar as it suggests the Cartesian model of isolated conscious monads [...] However, the mistaken metaphor is not "programming." Clearly, at the evolutionary level of human nature and culture, a symbolically materialized layer of intersubjectively shared, grammatically structured meanings emerges from the intensified interaction among conspecifics. Although the physiology of the brain does not permit any distinction between "software" and "hardware," the objective mind, in contrast to the subjective mind, can acquire the power to structure the individual brain. (Habermas 2008: 175).

This observation rules out a tight analogy between human minds and computers, but it also leaves the door open for a more nuanced stance when it comes to generative AI systems. LLMs escape, at least to some extent, the narrow formula of the individual hardware that runs its own pre-established software, as they are based on semi-automated learning processes fed by the same kind of socially shared "objective mind" that "programs" the individual human brain. Specifically, in the case of LLMs, the machine learning process trains the system on immense textual resources stored on the internet, on social media, and in the digital version of books and journals. Perhaps not surprisingly, then, the output of these AI systems is close to the kind of authorship and apparent creativity that we generally expect from human speakers, as they do not simply execute pre-programmed functions by rather "respond" to human prompts by articulating original pieces of writing. If we accept this distinction as consistent with the Habermasian stance on computers and programming, we can preliminarily note that, while from the "perspective of observation" humans and AIs clearly are two entirely different kinds of systems operating on their own rules and mechanics, from the "perspective of participation" the difference is much more subtle.[1]

[1] On the implications of this linguist dualism, Habermas notes: «The inescapable linguistic dualism compels us to assume that the complementarity of anthropologically deep-seated epistemic perspectives arose concurrently with the sociocultural form of life itself. The coeval emergence of the observer and participant perspectives would provide an evolutionary explanation for why the

Trained on the textual socialization of human cognition and displaying authorial properties, LLMs still lack, however, other salient traits that Habermas ascribes to humans as competent speakers in the community of communicants. AI participation in human discursive practices does not seem to lead to the formation of a sentient "subjective mind" out of their cognitive experience of socialization (Véliz 2021; Schwitzgebel 2023). This determines the novel situation of a new kind of actor that appears, in some relevant ways, capable of contributing to discursive practices as an author of discourse while, at the same time, not being fully responsible for its participation.

For Habermas, participation in discursive practices is a central aspect of becoming responsible agents: «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». This connection is rooted in the methodological primacy «enjoyed by the intersubjectively shared meanings embodied in joint practices in the sequence of explanation prior to internal states of the individuals involved» (Habermas 2008: 205). The process of becoming responsible agents is accompanied by a distinct reflexive aspect, specifically in the form of «a reflexive stability of our consciousness of freedom» (Habermas 2008: 208) rooted in the self-awareness that our convictions and our actions are grounded in meanings and reasons that inhabit ourselves and are shared, transmitted and revised within a community of communicants we belong to.

The reflexive nature of this linguistic-cultural genesis of human identities, Habermas notes, entails more than just the ability to draw from some pre-defined repertoires of signification and make use of them to articulate and justify actions. The "objective mind" is for humans a space of socialization, where the interaction with interpersonal semantic resources within shared practices is the basis for a self-conscious process of identification and projection into the future. Specifically, he argues:

> Only by growing into an intersubjectively shared universe of meanings and practices through socialization can persons develop into irreplaceable individuals. This cultural constitution of the human mind explains the enduring dependence of the individual on interpersonal relations and communication, on networks of reciprocal recognition, and on traditions. It explains why individuals can develop, revise, and maintain their self-understanding, their identity, and their individual life plans only in thick contexts of this kind (Habermas 2008: 296).

meanings that become accessible in our encounters with second persons do not admit of exhaustive objectification through the instruments of natural science» (Habermas 2008: 208).

This kind of reflexive consciousness and sense of identity is not a property that can be currently attributed to LLMs, at least based on how they operate and the kind of linguistic output they display. These preliminary considerations, then, suggest that the "perspective of participation" in practices is where the interaction between humans and AIs highlights both their common traits – as in the emergence of discursive capacities out of the learning process upon the "objective mind" of symbolic linguistic repertoires – and their differences – when it comes to the emergence of the intentional, desiring subjective mind of the human participants and the iterative, stochastic simulation that fuels the output of LLMs.[2] This, however, leaves substantially intact the problem of what kind of moral status should be attributed to this new kind of actor.

3. THE MORAL STATUS OF ATYPICAL PARTICIPANTS IN PUBLIC DISCOURSE

The question about the uncertain moral status of some specific kinds of participants in human interactions emerges within Habermas's work in at least two instances: the case of human subjects that have been genetically modified before birth and the case of animals who partake in our lives and daily practices. The two cases are obviously quite different, and they do not immediately overlap with the case of LLMs joining deliberative practices, but they are nonetheless relevant to explore the boundaries of Habermasian discourse ethics when confronted with fringe cases and atypical actors.

In *The Future of Human Nature*, Habermas points out that the moral status of genetically modified humans would be problematic insofar as the "genetic programming" (Habermas 2003: 63) artificially determines their subjectivity and their capabilities, thus putting them into a structurally unequal position within society.[3] This would in fact create an unprecedented rift in the evolution of horizontal, democratic relationships among humans: "Up to now, only persons born, not persons made, have participated in social interaction. In the biopolitical future prophesied by liberal eugenicists, this horizontal connection would be superseded by an intergenerational stream of action and communication cutting vertically across the deliberately modified genome of future generations" (Habermas 2003: 65). In other words, the moment the nature of some participants

[2] Whether one subscribes or not to the definition of LLMs as merely "stochastic parrots" (Bender et al. 2021), their inner workings are pretty commonly recognized to be a simulation of discourse achieved through a statistically based form of learning that differs substantially from the development of linguistic capacities in human subjectivities.

[3] This seems to suggest that the difference between genetic modification and education is akin to the difference between programming a computer, in the sense described by Habermas, and growing within a culture to become a free and competent participant in its conversations.

is artificially pre-determined by the intentions of others, non-peer relationships within the community of communicants also become inevitable and some members would be stuck in a structurally unequal position from which they cannot exchange roles with others.

This assessment of the relations entertained by genetically modified humans as inevitably uneven interactions offers an interesting perspective from which to look at the status of artificially made participants to discursive interactions like the LLMs. It is in fact, for Habermas, an issue that is deeply connected with the foundations of discourse ethics, insofar as it highlights that the moral space is defined concurrently by the equal form of dependence of all speakers from the linguistic structure of communication and by their active involvement in reflexively and cooperatively establishing the ethical boundaries of their process for reaching understanding and self-understanding:

> The *logos* of language escapes our control, and yet we are the ones, the subjects capable of speech and action, who reach an understanding with one another in this medium. It remains "our" language. The unconditionedness of truth and freedom is a necessary presupposition of our practices, but beyond the constituents of "our" form of life they lack any ontological guarantee. Similarly, the "right" ethical self-understanding is neither revealed nor "given" in some other way. It can only be won in a common endeavor. From this perspective, what makes our being-ourselves possible appears more as a transsubjective power than an absolute one. [...] As soon as the ethical self-understanding of language using agents is at stake *in its entirety*, philosophy can no longer avoid taking a substantive position (Habermas 2003: 11).

The special kind of "language using agents" represented by generative AI systems displays noteworthy capacities and producing outputs within the linguistic structure of communication, but disconnected from a comprehensive "form of life" shared with their human interlocutors that could provide a shared basis of engagement in a "common endeavor". The engagement in a lifeworld shared with others[4] is crucial in defining the profile of the moral subjects of discourse ethics, as they enter into a perspective of universal mutual recognition by reflecting on the normative

[4] It is interesting to notice that Habermas's articulation of the notion of lifeworld is also indebted to its Arendtian formulation. In an article published in 1977, Habermas reads Arendt through the lens of his developing theory of communicative action as follows: «the basic communicative action is the medium in which the intersubjectively shared life-world 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." This comprises the stories in which actors are involved as doers and sufferers» (Habermas 1977: 8). To our purpose, this commentary is helpful to highlight how consistently tight the link among communication, responsibility, and embodied presence in a shared life word appears in Habermas's inquiry. See also Arendt 1998: 189.

implications of the presuppositions implicit in their local experiences of communicative engagement with others:

> The ideas of justice and solidarity are already implicit in the idealizing presuppositions of communicative action, above all in the reciprocal recognition of persons capable of orienting their actions to validity claims». Of course, the normative obligations that children assume in virtue of the mere form of socializing interaction do not of themselves point beyond the limits of a concrete lifeworld (of the family, the clan, the city, or the nation). These barriers must first be breached in rational discourse. Arguments by their very nature point beyond particular individual lifeworlds; in their pragmatic presuppositions, the normative content of presuppositions of communicative action is generalized, abstracted and enlarged, and extended to an ideal communication community encompassing all subjects capable of speech and action (Habermas 1994: 50).

In the case of LLMs, the fundamental connection between arguments and "individual lifeworld" that is typical of communicative action is remarkably absent. LLMs are trained upon massive text corpora developed by countless individuals based on their own lifeworld, but as a system, they generate new fragments of discourse without being anchored to any specific lifeworld themselves. In this sense, the whole universalizing process is barred by the absence of a conspicuous link between an individual lifeworld and the speaker's reflexive consciousness of it as shared with other communicative partners. As Habermas observes, a display of cognitive and decision-making capabilities it is not sufficient to define the moral status of a person, since «[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* (*zurechnungsfiihige Personen*). They then impute to each other the capacity to orient themselves to validity claims in their actions» (Habermas 1994: 66).

The problem of "orienting their actions to validity claims" emerges, in different terms, for the designers of LLMs, in the form of what is generally designated in the literature as the value alignment problem, so as a problem intrinsic to the development of AI systems that need to identify relevant human values that are expected to guide the outcomes of the systems, implement these values into the machine learning process and assess that the output of the systems is consistent with those values (Arnold et al. 2017; Gabriel 2020; Christian 2020). But this kind of value aligning process seems quite far for the notion of self-orientation assigned by Habermas to human agents, since to achieve value alignment the identification of values needs to emerge from human actors and the assessment element is also largely dependent on human insight about AI outputs, such as in the case of Reinforcement Learning from Human Feeback (Knox and Stone 2011; Christiano et al. 2017; Kasirzadeh and Gabriel 2023). There are arguably elements of self-orientation insofar as AI systems become increasingly capable of achieving a more

"humanly aligned" orientation. Still, self-orientation as based on a reflexive assessment of their position within a community of speakers seems still definitely far from what LLMs are currently capable of expressing and the outputs of generative AIs are still largely "policed" through content filters introduced by the system developers to make sure that as certain words or requests are presented by human users, the response will be a pre-programmed no go (Derner and Batistič 2023) or by pre-filtering the training data, which in any case still does transmit human-biases into the learning process (Schramowski et al. 2022). In any case, improved AI outputs would still not meet the threshold of a fully moral form of self-orientation, since, as Habermas specifies, «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 1994: 66). Among moral persons, the orientation to validity claims is part of the intentional and free agency of all participants to the conversation, as they «Act with an orientation to mutual understanding and allow everyone the communicative freedom to take positions on validity claims» (Habermas 1994: 66). In the end, because of their distinct lack of self-reflexivity on a lifeworld and of self-orientation towards the goal of mutual understanding, LLM systems at the moment fall short of belonging to the community of speakers as peers, at least in the way humans are.

Once we acknowledge that, within the framework of discourse ethics, LLMs do not entertain the same moral status as humans, however, we are still faced with the conspicuous experience of their participation in our discursive practices. Habermas's account of the position of animals in his framework may prove useful to offer further clarification. In this regard, he notes that:

> Like moral obligations generally, our quasi-moral responsibility toward animals is related to and grounded in the potential for harm inherent in all social interactions. To the extent that creatures participate in our social interactions, we encounter them in the role of an alter ego as an other in need of protection; this grounds the expectation that we will assume a fiduciary responsibility for their claims. [...] To the extent that animals participate in our interactions, we enter into a form of contact that goes beyond one-sided or reciprocal observation because it is of the same kind as an intersubjective relation (Habermas 1994: 109-110).

These remarks open a space to consider that some non-human subjects may meaningfully participate in human interactions even though they are not peers and they are not structurally able to bear responsibility for their actions. In that context, the moral responsibilities fall on the human participants. The moral responsibilities of humans towards animals, however, according to Habermas, are limited to the scope of the specific interactions between individuals and within particular practices but do not universally bring the other species within the same moral realm (Habermas 1994: 111). In this perspective, participation in human interactions even

from a non-human status is sufficient to establish relations of responsibility, but this responsibility will be entirely up to the human participants. Compared with the case of animals, however, the peculiarity of generative AI in general, and LLMs in particular, is that their participation in our practices is performed specifically in a realm of linguistic creativity and authorship.[5]

4. LLMS AS CO-PARTICIPANTS IN DISCURSIVE PRACTICES

In light of these preliminary analyses of the status of atypical and non-human participants in human practices, I am now going to articulate more in detail how, within a Habermasian framework, AI writing systems can be acknowledged as a special kind of co-participants in human discursive practices, but not as fully communicative agents. In other words, LLMs are not moral or epistemic peers with humans but can still partake in the same public conversations as authors. Their contribution is, in this sense, not merely instrumental: they create original fragments of intelligible discourse that, when introduced into a conversation, can contribute to the process of clarification and understanding among the members of the community of communicants. In instances of public discussion and deliberation, fragments of discourse generated by AI systems can be then very well used to articulate difficult concepts, summarize different perspectives, or even introduce previously neglected ideas. Latest-generation LLMs are also able of expressing real-time interactions within the context of an online chat, which brings their contribution even closer to the same kind of back-and-forth participation typical of argumentative exchanges among peers.

As we mentioned before, it is however unprecedented that the author of a piece of contribution to a discursive engagement is not immediately recognizable also as a responsible moral agent that is, or has been, a human member of the community of speakers.[6] To understand the implications of this decoupling, it is useful to consider how Habermas characterizes, in general, the relation between authors and interpreters of a text, to then suggest a consistent characterization of AI authorship in terms of communicative competence and agency.

In the process of reaching understanding, Habermas argues, the interpreters approach a text based on the assumption that they can understand what the author

[5] This is not to deny that animals seem able to express forms of creative communication and visual performance, but not within the specific realm of human visual and written languages, as generative AIs do.

[6] Naturally, several philosophical and theological traditions have contemplated and reflected upon the possibility of engagements with spiritual and divine interlocutors through the medium of language. This present account just looks at the issue within the scope of Habermasian post-metaphysical thinking. There are, however, interesting analogies and insights that can be drawn from an engagement between AI and theological studies. See also Brittain 2020, O'Gieblyn 2021, Oviedo 2022.

is saying because of a certain grasp of the context within which the text has been conceived and makes sense. This assumption rests on the notion that both interpreter and author raise validity claims on truth, values, and sincerity within a specific context, but that the reasons why they think they can do that are rationally accessible from context to context:

> [O]nly to the extent to which the interpreter also grasps the reasons why the author's utterances seemed rational to the author himself does he understand what the author meant. The interpreter, then, understands the meaning of a text only insofar as he understands why the author felt justified in putting forth certain propositions as being true, in recognizing certain values and norms as being right, and in expressing certain experiences (or attributing them to others) as being authentic. [...] Interpreters cannot understand the semantic content of a text if they do not make themselves aware of the reasons the author could have brought forth in his own time and place if required to do so. (Habermas 1990: 30)

Based on this picture, when the user approaches an AI-generated text as the product of an author, she will still have to rely on the presupposition that at the other side of the conversation there is an interlocutor that produces and understands meaning the same way the interpreter does. LLMs, however, do not operate the same way their human readers and listeners do, based on relatable reasons that make sense within their relationships to a lifeworld and that allow for reasoning about their mutual mental states (Trott et al. 2023). They rather produce an accurate simulation of what an appropriate utterance would be in the face of the textual prompt of the user based on the elaboration of the existing repertoire of appropriate utterances available to the machine learning process. The interpreter can still find in the text some plausible discourse around the topic at stake, but the understanding will happen "as if" the author had reasons the same way the interpreter does:

> 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. That is why the interpreter cannot simply look at and understand such reasons without at least implicitly passing judgment on them *as* reasons, that is, without taking a positive or negative position on them. [...] 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: 30-31).

The conditions of this explanation, however, are different for the interpreter confronted with a text produced by a LLM, since the way obscure or dubious expressions have been generated radically differs from the kind of process that is usually found among human speakers. Anthropomorphic first-person statements do not arise from a personal connection with an individual lifeworld, hallucinations

are unforeseen outcomes of a stochastic process rather than a form of perceptual distortion (Hongbin et al. 2023). This brings into question to what extent LLMs, besides their evident authorial capacities, can be credited with 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: 31).

Whereas investigating the issue from the perspective of internal intuitions seems unfruitful, it is instead important to acknowledge that AI systems can be designed as more or less capable of providing "reasons" for their outputs when required to do so, thus supporting the interpreter's job. The problem of the explainability of AI systems is increasingly subject to scrutiny (Preece 2018), with a growing awareness of the ethical dimension of this aspect of their design (McDermid 2021).[7] What the Habermasian perspective suggests here, is that explainability is not only relevant in consequentialist terms, to improve the accuracy and readability of the outputs, but also more substantially to bring the participation of LLMs in discursive practices closer to an expectation of reciprocity that is intrinsic to the process of human understanding. However, even if advances are being made in the field of LLMs explainability, we are still left with the conundrum of their lack of vital relationship with a contextual lifeworld, which is highly problematic within the paradigm of communicative rationality. This becomes apparent by looking more closely at how the linguistic performance of LLMs can – or cannot – be characterized within the Habermasian concepts of *communicative competence* and *communicative action*.

For Habermas, who originally developed the concept inspired by Noam Chomsky's theory of linguistic competence, *communicative competence* is the implicit know-how that speakers have of the implicit rules and presuppositions that make them capable to produce and understand utterances (Habermas 1970; Allen 2019). The most fundamental presupposition, the orientation towards reaching mutual understanding, is specified into validity claims over truth, normative rightness, and sincerity. Every speaker has the implicit expectation that, under suitable conditions, their claims to truth, normative rightness, and truthfulness should be acceptable to all (Habermas 1990: 31). LLMs react to their users' utterances by simulating a human use of language that ordinarily stems out of those presuppositions; by doing so they generate understandable new pieces of discourse. In human communication, each type of validity claim rests on a kind of world relation: relations to the objective natural world, to the intersubjective social world, and to the inner subjective world. LLMs, however, do not entertain the same kind

[7] The importance of explainability to define criteria of accountability and responsibility for AIs is also signaled by the attention it is receiving from policymakers. In the 2020 Assessment List for Trustworthy AI (ALTAI), a document prepared by a group of high level experts set up by the European Commission, accountability is defined as «the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons».

of world relations as human speakers do and this affects the way they can engage beyond the level of unquestioned everyday communication into the medium of discourse where validity claims are challenged and adjudicated. In the case of current generative AI, truth claims are structurally derivative from those embedded in the textual sources that fed the machine learning process, since the systems have no "experience" of the world or direct access to the natural world to raise and verify truth claims of their own. Normative rightness claims are based on judgments about the appropriateness of speech acts, but these need to rely on the social relationships that the speakers entertain as peers that inhabit a shared lifeworld. Finally, and possibly even more problematic, sincerity claims should be vindicated by the consistency between actions and claimed subjective states of the speakers, but LLMs have no "actions" to display beyond their writing and claims about subjective states are the expression of simulated anthropomorphic approaches to user interaction rather than the reflection of any subjective state we know of. LLMs seem to possess, then, only a structurally derivative communicative competence.

Similarly limited by their lack of lifeworld relations is the ability of AI systems to express proper communicative agency through their participation in human practices. For Habermas, *communicative action* is characterized by the use of discourse to coordinate the actions of its participants (Habermas 1984; Krüger 2019). To some extent, LLMs do adapt to the kind of discursive input they receive from their users, like when correcting previous statements that have been pointed out as erroneous, or when modifying the style of communication based on previous interactions. However, these systems do not self-regulate their own guidelines, which are externally established by their developers and often not even disclosed to the users. Moreover, because of the structurally derivative nature of their communicative competence, LLMs also cannot autonomously adjust their behavior based on contestations to their validity claims, given the absence of direct experiences of the world and of recordable subjective states that can serve as a basis to support and adjudicate those claims. Ultimately, LLMs are capable of manifesting some simulation of communicative agency, but they are not autonomous communicative agents.

Luciano Floridi has influentially argued that the behavior expressed by LLMs is a form of agency without intelligence or understanding (Floridi 2023). His perspective makes sense within a general effort to downplay the kind of "intelligence" that AIs are actually capable of. However, it is noteworthy that, at least within a Habermasian framework,[8] communicative agency without understanding is not even proper agency in the first place. For Habermas, the most paradigmatic form of human agency is indeed the outcome of an interplay between linguistic understanding and autonomous behavior, where each polarity is essential in

[8] I suspect also within several non-Habermasian frameworks, but supporting this conclusion goes beyond the scope of this paper.

defining and structuring the other. LLMs at present appear to be extraordinarily prolific linguistic authors, but not full communicative agents: a crucial decoupling that brings us to the puzzling question of how responsibility for their original utterances should be assigned when they participate in our discursive practices.

5. AI AUTHORSHIP AND MORAL RESPONSIBILITY IN DISCOURSE AND DELIBERATION

The itinerary developed so far allows some tentative suggestions as to how the moral status of LLMs within our discursive practices should be assessed and to what kind of new responsibilities emerge out of the already ongoing discursive interactions between humans and AIs.

From the perspective of their participation in communicative practices, the creative capacity displayed by generative AI systems suggests that their status goes beyond that of mere technical tools in the hands of human speakers. LLMs, then, can be acknowledged as original authors even within specialized discursive practices. In these contexts, they can positively contribute with their extraordinary authorial capabilities to support the ongoing process of mutual understanding among all participants. This kind of contribution could have empowering functions, especially for human participants in public conversations who are otherwise disadvantaged by disabilities, lack of linguistic prowess, or rhetorical education (Kasneci et al. 2023; Pavlik 2023).

At the same time, the limited explanatory capacity, derivative communicative competence, and lack of proper communicative agency of these systems stand in the way of any project to construe them as a new kind of morally responsible subjects that join the community of communicants as peers with their human counterparts. AI systems participate in discourse but not in communicative action. In their discursive interactions, LLMs cannot have, at present, interchangeable roles with human counterparts, a requirement of communicative agents that is fundamental for Habermas to ensure parity among the actors of discursive and deliberative practices. However, human members of the linguistic community can integrate AI authorial contributions into their own communicative agency and vicariously provide the connection with lifeworld relations that AIs lack. Suggestions, insights, images, and information discursively organized by AI systems can resonate with the members of the community of speakers and the living relations with the world, society, and themselves. In turn, those human speakers can operate as moral proxies and stand for the validity of claims raised by AI-generated discourse.

In this perspective, we can still interact within the same discursive practice with human and non-human participants, but the process of mutual understanding needs to adapt substantially to the kind of moral status that each participant

entertains within the community of speakers. For this to happen, it is necessary that all human participants are transparently made aware if they are discursively engaging with a human or an AI and if the author of the piece of discourse they are engaging with is human or not.

On these grounds, I argue that a twofold normative stance on the participation of LLMs in discursive practices can be taken:

> (i) First, based on their authorship properties, the contribution they may bring to the articulation of public discourse, and the enhancement of otherwise discursively disadvantaged participants to the conversation, the involvement of linguistically trained AI systems in our discursive and deliberative practices is acceptable, provided that the human members of the community of speakers (a) take the necessary steps to disclose the authorship of AI contributions and the identity of those who brought them into the conversation (*responsibility as attribution*) and (b) are ready to respond to the contestation of the validity claims that are raised through those contributions, especially when it comes to claims of assertoric truth and subjective truthfulness (*responsibility as answerability*).
>
> (ii) Second, in the field of institutionalized political procedures and formal processes of argumentation and negotiation, the moral call for the disclosure of AI authorship is even more comprehensive and urgent, as the legitimacy of the deliberative procedures is based on the condition of democratic citizens as co-authors of the law, which demands a substantial correspondence between the community of speakers and the community of those who are affected by the normative outcomes of deliberation.

The ensemble of agents involved in bringing about, distributing, and re-circulating the fragments of discourse produced by LLMs collectively shares a responsibility that the AI systems cannot bear themselves, as they are not full communicative agents within the community of speakers, although through their contributions they can participate in important discursive and deliberative practices. It is important to note that the decoupling of authorship and responsibility does not allow for a one-on-one transfer of accountability from the AI system to a singular human subject. The system developers are not responsible for what a Chatbot "says" as if it they said it themselves. Similarly, anyone who brings a text drafted by a generative AI into a public debate is not solely responsible for it in the same way as if they had written it by their own hands before entering the conversation. However, this phenomenon is not the cause of a collapse of responsibility, but rather the premise of a new form of diffused responsibility between company

owners, AI system developers, service users, and social media sharers.[9] Responsibilities will be adjudicated case by case within this relational network based on the agents that played a decisive role in getting that piece of AI discourse into that specific discursive practice.[10]

Without this kind of relational moral context supplied by human speakers, we are left with fragments of "rogue discourse" generated by AIs that by entering our discursive and deliberative practices may determine the effects of communicative agency in the absence of communicative agents that are responsible for them as full members of the community of communicants that share the same world with their peers.

6. CONCLUSIONS: CITIZEN AI?

The perspective of a pervasive presence of generative AI systems within the public sphere of liberal democracies inspires motivated concerns, especially at a moment in history when the advent of social media and the rise of populist movements haven't yet exhausted their momentum and have abundantly shown how deeply technological transformations can affect the political realm (Sunstein 2017; Dijk and Hacker 2018; Urbinati 2019).

It is to be noted, in this sense, that the interpretive framework sketched here, which sees AI systems are creative participants in highly sophisticated human practices without assigning them the full status of moral agents, can be applied also to other kinds of generative AI, beyond the case of LLMs. An obvious example are visual AIs like DALL-E by OpenAI, Midjourney by Midjourney Inc., and Stable Diffusion by Stability AI. The visual creative practices where these systems express their authorial capabilities are not as central in the Habermasian account as the medium of language and discourse are. However, in the digital public sphere, the importance of the production and circulation of images and videos in shaping cultural trends and embodying political agendas can be hardly overstated (Green 2010; Bottici 2014).

In the recent past, when considering the rise of genetical engineering, Habermas had already raised significant concerns about the risk that technological innovation could reshape our moral identities and introduce new forms of political subjectivity in undesirable ways. In a sense, his core preoccupation with this process of technological transformation being appropriated by the strategic and self-serving

[9] Moving from an Aristotelian framework, Mark Coeckelbergh comes to a similar conclusion by articulating a relational account in terms of distributive and collective responsibility from individuals and organizations involved in AI development and use (Coeckelbergh 2020).

[10] Notice that the effort to disclose the AI authorship could also be expressed through the conscious adoption of technical solutions like automatic watermarks for texts produced by LLMs that can be tracked with appropriate tools. As an example, see Kirchenbauer et al. 2023.

logic of capitalism at the expense of the egalitarian and communicative nature of the democratic ethos remains largely applicable to the case of AI:

> The self-understanding of this subject [that intervenes to artificially shape future individualities] now determines how one wants to use the opportunities opened up with this new scope for decision – to proceed *autonomously* according to the standards governing the normative deliberations that enter into democratic will formation, or to proceed *arbitrarily* according to subjective preferences whose satisfaction depends on the market. In putting the question this way, I am not taking the attitude of a cultural critic opposed to welcome advances of scientific knowledge. Rather, I am simply asking whether, and if so how, the implementation of these achievements affects our self-understanding as responsible agents.
>
> Do we want to treat the categorically new possibility of intervening in the human genome as an increase in freedom that requires normative *regulation* – or rather as self-empowerment for transformations that depend simply on our preferences and do not require any *self-limitation*? (Habermas 2003: 12)

Worries about the influence of power and capital over the infrastructure of democratic societies at the expense of agency coordinated through discourse and mutual understanding have been a central focus for Habermas during his entire career. This problematic influence takes everchanging forms and it doesn't seem unthinkable that the next incarnation of "Citizen Kane", who achieves political domination through the use of media, could be in the near future a "Citizen AI" in the shape of one of the global market players that are heavily investing into the development and release to the public of services based on machine learning.

To be fair, it is unclear what the technological advancements in AI technology will produce in the near future. We could very well see soon more direct engagements of LLMs with real-world interactions of some sort, even though it is at least dubious that these interactions will count as vital relations with a lifeworld in the same way as human experience and self-awareness do. It is also possible that the explainability of future AIs will rapidly improve, thus rendering these systems more reliable and responsive partners in our own practices.

In the meantime, it is certainly up to humans to make sure that the insights that emerge out of their embodied circumstances and relational experiences, together with their self-reflexive awareness of the discursive presupposition of their mutual understanding, keep nurturing their moral insight into the responsibilities at stake in all their conversations, including those with their brand-new kind of non-human partner.

REFERENCES

Allen, A. 2019. *Communicative Competence*. In A. Allen & E. Mendieta (Eds.), *The Cambridge Habermas Lexicon*. Cambridge: Cambridge University Press, pp. 47-48.

Arendt, H. 1998. *The Human Condition*. Chicago and London: The University of Chicago Press.

Arnold, T., Kasenberg, D., Scheutz, M. 2017. *Value Alignment or Misalignment—What Will Keep Systems Accountable?*. In «Workshops at the Thirty-First AAAI Conference on Artificial Intelligence».

Bender, E.M., Gebru, T., McMillan-Major A., Shmitchell S. 2021. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. In «FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency», pp. 610–623.

Bottici. C. 2014. Imaginal Politics. Images beyond Imagination and the Imaginary. New York: Columbia University Press.

Brittain, C.C. 2020. *Artificial Intelligence: Three Challenges to Theology*. In «Toronto Journal of Theology», 36, 1, pp. 84-86.

Calloni, M., Nicoletti, M., Petrucciani, S. 2021. Filosofia, pensiero post-metafisico e sfera pubblica in cambiamento. Intervista a Jürgen Habermas. | Philosophy, Postmetaphysical Thinking, and a Changing Public Sphere. An Interview with Jürgen Habermas. In «Rivista Italiana di Filosofia Politica», 1, pp. 137-154.

Christian, B. 2020. The Alignment Problem: Machine Learning and Human Values. New York: W.W. Norton & Company.

Christiano, P.F., Leike, J., Brown, P., Martic, M., Legg, S., Amodei, D. 2017. *Deep Reinforcement Learning from Human Preferences*. In «NIPS 2017 – Advances in Neural Information Processing Systems 30», Long Beach, CA, USA.

Coeckelbergh, M. 2020. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. In «Science and Engineering Ethics», 26, pp. 2051–2068.

Derner, E., Batistič, K. 2023, *Beyond the Safeguards: Exploring the Security Risks of ChatGPT*. Pre-print In: «arXiv:2305.08005».

Dijk, J.A.G.M. van, Hacker, K.L. 2018. *Internet and Democracy in the Network Society*. London and New York: Routledge.

Floridi, L. 2023. AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. In «Philosophy & Technology», 36, art. 15.

Gabriel, I. 2020. *Artificial Intelligence, Values, and Alignment*. In «Minds & Machines», 30, pp. 411–437

Gordon, J.-S. and Gunkel, D.J. 2021. *Moral Status and Intelligent Robots*. In «The Southern Journal of Philosophy», 60, 1, pp. 88-117.

Green, J.E. 2010. The Eyes of the People. Democracy in an Age of Spectatorship. Oxford: Oxford University Press.

Habermas, J. 1970. *Towards a Theory of Communicative Competence*, In «Inquiry», 13, pp. 360–75.

Habermas, J. 1977. *Hannah Arendt's Communications Concept of Power*. In «Social Research», 44, 1, pp. 3-24.

Habermas, J. 1984. *The Theory of Communicative Action*, vol. 1, Boston, MA: Beacon Press.

Habermas, J. 1990. Moral Consciousness and Communicative Action. Cambridge: Polity.

Habermas, J. 1994. Justification and Application. Remarks on Discourse Ethics. Cambridge MA and London: MIT Press.

Habermas, J. 2003. *The Future of Human Nature*. Cambridge: Polity.

Habermas, J. 2008. *Between Naturalism and Religion*. Cambridge: Polity.

Habermas, J. 2023. A New Structural Transformation of the Public Sphere and Deliberative Politics. Cambridge: Polity.

Hongbin, Y., Tong, L. Aijia, Z., Wei, H., Weiqiang, J. 2023. *Cognitive Mirage: A Review of Hallucinations in Large Language Models*. Pre-print In: «arXiv:2309.06794».

Kasirzadeh, A., Gabriel, I. 2023. *In Conversation with Artificial Intelligence: Aligning language Models with Human Values*. In «Philosophy & Technology», 36, 27.

Kasneci, E. et al. 2023. *ChatGPT For Good? On Opportunities And Challenges Of Large Language Models For Education*. In: «Learning and Individual Differences», 103

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T. 2023. *A Watermark for Large Language Models*. Pre-print In: «arXiv:2301.10226v3».

Knox, W.B., Stone, P. 2011. *Augmenting Reinforcement Learning with Human Feedback*. In «Proceedings of the ICML Workshop on New Developments in Imitation Learning», Bellevue, WA, USA,

Krüger, H. 2019. *Communicative Action*. In A. Allen & E. Mendieta (Eds.), *The Cambridge Habermas Lexicon*. Cambridge: Cambridge University Press, pp. 40-46.

McDermid, J.A., Jia, Y., Porter, Z., Habli I. 2021. *Artificial Intelligence Explainability: the Technical and Ethical Dimensions*. In «Philosophical Transactions of the Royal Society A», 379, 2207.

O'Gieblyn, M. 2021. *God, Human, Animal, Machine. Technology, Metaphor, and the Search for Meaning*. New York: Doubleday.

Oviedo, L. 2022. *Artificial Intelligence And Theology: Looking For A Positive – But Not Uncritical –Reception*. In« Zygon», 57, 4, pp. 938-952.

Pavlik, J.V. 2023. *Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education*. In « Journalism & Mass Communication Educator», 78, 1.

Preece, A. 2018. *Asking 'Why' in AI: Explainability of Intelligent Systems – Perspectives and Challenges*. In «Intelligent Systems in Accounting, Finance and Management», 25, 2, pp. 63-72.

Redaelli, R. 2023. *Different Approaches To The Moral Status Of AI: A Comparative Analysis Of Paradigmatic Trends In Science And Technology Studies*. In «Discover Artificial Intelligence», 3, 25.

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*. In «Nature Machine Intelligence», 4, pp. 258-268.

Schwitzgebel, E. 2023. *AI Systems Must Not Confuse Users About Their Sentience Or Moral Status*. In «Patterns», 4, 8.

Sinnott-Armstrong, W., Conitzer, V. 2021. *How Much Moral Status Could Artificial Intelligence Ever Achieve?* In: S. Clarke, H. Zohny, J. Savulescu (Eds.) *Rethinking Moral Status*. Oxford: Oxford University Press, pp. 269-289.

Sunstein, C.R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.

Trott, S., Jones, C., Chang, T., Michaelov, J., Bergen, B. 2023. *Do Large Language Models Know What Humans Know?* In «Cognitive Science», 47.

Urbinati, N. 2019. *Me the People. How Populism Transforms Democracy*. Cambridge, MA: Harvard University Press, 2019.

Véliz, C. 2021. *Moral Zombies: Why Algorithms Are Not Moral Agents*. In «AI & Society», 36, pp. 487–497.