If God Looked Into AIs, Would He Be Able To See There Whom They Are Speaking Of?

Mirco Sambrotta
Institute of Philosophy, Slovak Academy of Sciences, v.v.i.

SAMBROTTA, M.: If God Looked Into AIs, Would He Be Able To See There Whom They Are Speaking Of?
Philosophica Critica, vol. 9, 2023, no. 2, ISSN 1339-8970, pp.42-54.

Can Large Language Models (LLMs), such as ChatGPT, be considered genuine language users without being held responible for their language production? Affirmative answers hinge on recognizing them as capable of mastering the use of words and sentences through adherence to inferential rules. However, the ability to follow such rules can only be acquired through training that transcends mere formalism.Yet, LLMs can be trained in this way only to the extent that they are held accountable for their outputs and results, that is, for their language production.

Keywords: AI – Large Language Models – Responsibility – Inferentialims – Rule-following.

Introduction

A notable feature of the current generative AI boom is the machine processing of natural language by Large Language Models (LLMs). LLMs are neural-networks, or better, current transformer-based neural natural

language processing systems[1] (such as GPTs and ChatGPTs),[2] which made available for the first time an AI with human-level performance on a wide range of cognitive tasks. One of such achievements obviously is (or seems to be) the ability of generating language. These models can generate text by processing prompts (e.g., a short passage of text, often a single sentence), and autonomously produce coherent continuations, potentially achieving flawless performance across various conversational topics based solely on linguistic input. Unlike previous text generators like SCIgen, LLMs seems able to generate well-formed and meaningful texts as outputs often indistinguishable from that by humans. There is indeed compelling evidence indicating that they can even outperform many human writers.Ultimately, language processing in LLMs by generative AI easily goes undetectable by human scrutiny, passing in this way the well-known Turing Test (Turing 1950) for determining whether a machine can demonstrate human intelligence.[3]

Even more interesting and surprising is the fact that their applications extend to the production of academic research, as evidenced by numerous papers accepted for publication after undergoing peer review, where AIs are credited as co-author. [4] This prompts the question of whether LLMs could be eligible for authorship. The Committee on Publication Ethics (COPE), the World Association of Medical Editors (WAME), and the prestigious journal *Nature* (vol. 613, no. 7945, p. 612) have recently called for banning ChatGPT's authorship on the basis that, even if AIs genuinely can make scientific contributions, they "cannot take responsibility" for their output (Miller 2023, p. 7).

However, if LLMs such as ChatGPT cannot be held accountable in a normative sense, and accordingly cannot count as an author, the key

[1] The terms 'transformer' or 'transformer-based' refer to a new generation of natural language processing (NLP) architectures pioneered by Vaswani et al. (2017).

[2] GPT (Generative Pre-trained Transformer) uses the decoder part and BERT (Bidirectional Encoder Representations from Transformers) uses the encoder part. In both cases, the resultant language models can then be used for various NLP tasks (Natural Language Processing).

[3]See Elkins and Chun (2020).

[4] See C. Stokel-Walker (2023).

question becomes whether they can still fulfil the requirements (meet the standards) for making scientific contributions. In the first place, do they really count as language users?[5] The question is whether artificial neural nets can deal with and process conceptual contents. If not, should LLMs be treated (like automatic grammar and spell-checkers or translators) as producing written output without rising to the level of language user? If so, what else would they need to rise to the level of language user? The difference, no doubt, is the respective presence or absence of understanding. Can then a natural language processing (NLP) machine understand the meanings of natural language? Are there LLMs that grasp and understand concepts? Or should regard machine language performance as mere simulation of understanding (e.g., mere syntactic manipulation)? But what does *understanding* amount to?

1Intentionality

Linguistic understanding is intimately connected to the directedness distinctive of (at least some of) our linguistic utterances (and psychological states). Why? Because language cannot be made sense of without appeal at the same time to the idea of a kind of contentfulness that is distinctive of at least some of our linguistic utterances (and psychological states). We can then pretheoretically specify the content of a linguistic utterance by saying, for instance, that it is of or about something, represents something in the world, has a peculiar kind of link to the world, and so on. Therefore, language cannot be made sense of without appeal to intentionality. There cannot be language in absence of intentionality. Accordingly, understanding hinges on intentionality.

But what exactly is intentionality? Brentano defined intentionality as "reference to a content, a direction upon an object" (1970). Similarly, Searle emphasizes that: "...if a state S is intentional then there must be an answer to such questions as: What is S about? What is S of? What is it an S that?" (1983). More recently, Bender and Koller take (linguistic) meaning as '…the

[5]On the other hand, if LLMs are not language users, and accordingly fail to make a scientific contribution, then they obviously cannot meet standard for authorship. Naturally no person or entity which fails to make a scientific contribution should be listed as an author. Surely this state of affairs would justify Nature's insistence that LLMs cannot be credited as authors.

relation between a linguistic form and communicative intent' (2020, p. 5185). That is, according to them, the meaning of a word is its communicative intent.[6] In language processing, we are guided by the pursuit of certain intentions, which we express in linguistic expressions. Understanding meaning is then the ability to map an expression onto its intent.

In a nutshell, discursive intentionality is what is exhibited by language users, as concept users, who can say and understand that things are thus and so**,** who can make and understand claims or judgments that are about something, about objects in the world, and so forth [7]

But how does intentionality arise? How this kind of relationship between a linguistic item (i.e., a sign) and things outside of it come about? By virtue of what, by what means, does its rapport with their environment arise? This is hardly a causal connection.[8] At the same time, it is hard to appeal to some specific subjective experiences. It is hard to maintain that specific subjective experiences are necessary for linguistic understanding and hence for intentionality. To understand language does not require any subtle ingredient such as qualia (or mysterious 'inner light').

2Inferential Rules

[6] This view is somehow remineshent of Grice's (1957), which understands linguistic meaning in terms of speaker's meaning, and speaker's meaning in terms of the intention of a speaker to induce a belief in the audience by an utterance accompanied by the audience's recognition that the utterance was produced with that very intention.

[7] It is implausible that the normal subject does not have the term "intentionality," but she has a battery of other terms— most notably, "about" and "directed"—with which she can express her concept(ion) of intentionality. Without such a conception, she would be unable, for example, to understand simple traffic signs: 'Traffic signs are about something: a sign on Interstate 95 that says "New York, next exit" is *about* the sorts of actions that have to be taken in order to reach New York from where the reader is at' (Kriegel 2011, p. 55).

[8] The question is not a causal one either (i.e., how linguistic understanding is causally produced), but rather a conceptual one: what criteria do we apply to decide whether some being understands?

If intentionality, and accordingly the meaning of any concept (even concepts close to sensory experience such as “pain” and “red”) neither derive from a causal relationship to the environment nor from some sensory experience resulting from interactions with the environment, then how does meaning get its connection with the world? Sellars offers the following answer:

“…it is by virtue of the fact that we draw inferences that meaning gets its connection with the world” (Sellars 1962, p. 246).

“It is only because the expressions in terms of which we describe objects… locate these objects in a space of implications, that they describe at all, rather than merely label” (Sellars 1958, pp. 306-307).

Thereby, the meaning of an expression is not something that lies behind the expression itself (or to which the expression refers). Nor does it exist without its relations to other expressions, whose meanings are themselves determined only by their relations to other expressions. Rather, conceptual content consists in the inferences to other concepts. In particular, something qualifies as a conceptual content just insofar as it stands in relations of *material consequence* and *incompatibility* with other such contents. Material consequence and incompatibility relations, by contrast to formal logical ones, thus articulate the contents of non-logical concepts. In turn, propositional contents (which are a principal species of conceptual content) are what can perform the *office* both of premise and of conclusion in inferences. These complex relations of the individual parts of the corpus to all other parts of the corpus then ground the meaning of the isolated parts in a given context.

This view clearly stems from the work of Wittgenstein, according to which the meaning of an expression is its role within our *language games*, "its use in the language" (Wittgenstein, 1953, § 43):

‘Compare the meaning of a word with the "function" of an official. And "different meanings" with "different functions"’ (Wittgenstein, 1969, p. 69).

The inferentialist semantic claim is that what distinguishes specifically discursive (paradigmatically, but not exclusively, propositional) contents is the *roles* they play in material consequence and incompatibility relations (Brandom 1994; Peregrin 2014). By playing the role they do in a network of such relations that expressions acquire the propositional content that

makes possible the discursive that consists in explicitly claiming or judging that things are thus-and-so[9]:

'Any web of relationships among linguistic items equips the items with more or less complex roles, which may be considered as their "meanings"' (Peregrin 2021, p. 314).

The meaning of an expression is thus its role (i.e., its place) in the complexity of relations to all other expressions. To grasp this meaning, hence to understand the expression, is nothing but to grasp and understand its inferential role.

But roles are conferred by *rules*! Wittgenstein's insight of language games as rule-governed identifies meanings with the roles conferred on the linguistic items by the rule governing their use. Therefore, meanings are the roles conferred on the linguistic items by the norms that govern their application, paradigmatically in judgment– and the deployment of judgeable, that is, propositional contents. Following Brandom's "inferentialism" (1994), an expression's contentfulness consists in its use or occurrence being governed by *inferential rules*. Contentful items incorporate norms of inference (i.e., material consequence and incompatibility relations), which they are subject to. Grasping the meaning of a word, understanding it, is thus mastering this bundle of rules. As a result, to count as a language user, one must *know how* to make inferences and so draw conclusions from his premises. One must know how to distinguish what is evidence for and against that claim, and what else is ruled out as incompatible. If so, mastering a language turns out to be primarily a skill, and understanding an expression a knowing how to skillfully employ it.

Do computers possess this know-how, this skill, this practical ability? Surely, we can say that computers make inferences, that they can draw

[9] Brandom's claim that "intentional states and attitudes have the contents they do in virtue of the role they play in the behavioral economy of those to whom they are attributed" (1994, p. 134). Brandom's language here matches that used by Wilfrid Sellars. When Sellars says that something's meaning what it does is constituted by the "role" it plays in a speaker's "behavioral economy" (1957). According to Sellars, though, the function of semantic statements is not however to describe such roles (1957, pp. 527-532). Similarly, Brandom never regards the function of meaning-ascriptions as that of describing community attitudes.

conclusions from premises, decide whether one sentence implies or contradicts another sentence, determine the steps which are reachable and the steps of the derivation that are precluded, and so on and so forth. So, there is a sense in which they could individuate information by its place in an inferential network.

Should then conclude that computers are language user and understand language? Of better, should we see this as sufficient to conclude that computers are language use and understand language? People who think so are not thinking in normative terms. They are thinking about dispositions to make inferences. But inferential roles, hence meanings, are not a matter of our dispositions, they are a matter of the norms that we bind ourselves by. Then the proper question to ask should be: Do computers really follow rules instead of merely exhibiting certain dispositions and regularities in their behavior?

3Rule-Following & Machines

Intellectualism (a contemporary version of *Platonism*) sees every practice as underwritten by a rule or principle: something that is or could be made discursively explicit. For instance, the *computational theory of the mind* endorses the possibility of explicitly stating in rules all the implicit practical background skills (i.e., practices) necessary to institute those rules. This view is shared by the program of *symbolic artificial intelligence*, which endorses the possibility of explicitly codifying in programmable rules all the implicit practical background skills necessary to institute those very rules.

Is this actually possible? Can the practical capacities to follow rules be cashed out in terms of computational operations? Computational operations (or processes) are what we can call computer's internal states (or software states). They are programmed instructions (i.e., instantiations of codes, algorithms, software, etc.). Computer's internal states are merely the state resulting from its programming (i.e., from the programmed instructions). Software states can be understood as a device's internal representations or "interpretations." Then the question we face is the following: Can we attribute to the machine the ability to follow rules (such as the inferential rules governing language use) in light of such interpretations or internal representations?

Wittgenstein's analysis of rule-following nicely fits here: we cannot, and need not, know which internal representation the pupil relies on when correctly continuing a series of numbers (Wittgenstein 1953, §143-§185). He goes on to suggest that the rules governing human activity in general (and linguistic practice in particular), cannot be explicated by the Platonic tradition of reference to ineffable objects, nor by a subjective 'interpretation' at the moment of each instantiation of the rule.[10] Otherwise, we would encounter the dilemma of an *infinite regress* (Wittgenstein 1953, §201). If performing an action correctly requires consulting the rule or proposition that guides it, then the act of consulting that rule itself requires, in turn, consulting the rule that guides the act of consulting the previously consulted rule and so forth.[11] Thereby, rules neither determine actions nor determine or stablish meaning.

However, the language of thought view proposed by Fodor (1975; 1987) assumes that computers use language in a way that involves internal mental life. According to the Wittgensteinian perspective advocated here, language use is a matter of following certain (social) rules governing the employment of expressions.[12] But rule-following is not a matter of what is in the mind:

"If God had looked into our minds he would not have been able to see there whom we were speaking of" (Wittgenstein 1953, p. 217).

Rule-following (and accordingly language use) is not a matter of having inner representations of any kind.[13] And therefore, it cannot boil down to

[10] "However, that neither a mental image nor the referent should be identified with the meaning of an expression should be clear at least since Frege (1892, 1918)" (Peregrin 2021, p. 315).

[11] In the same way, if a device's output inherited its meaning by the device's internal state, then the internal state could not itself consist in representation (and hence being meaningful) without presupposing further representation; we would thus again be faced with an infinite regress.

[12] Recall that, following Inferentialist semantic, such rules are inferential rules.

[13] Wittgenstein (1953, §158) also explicitly ponders the idea of assessing whether somebody has mastered reading by further investigating neural processes in the brain. Wittgenstein rejects the idea by emphasizing our knowledge of such matters, that is, the way we commonly assess whether somebody knows how to read. Surely, these would be the criteria to identify any neural processes jointly necessary and sufficient for mastery of reading, and not vice versa.

mere computation either. To qualify as language users, computers should be able to follow the rules governing language usage indipendently on their internal states and mechanisms (e.g., computational processes, code implementations, algorithms, etc.).

Conclusion

According to the Committee on Publication Ethics (COPE), the World Association of Medical Editors (WAME), and the prestigious journal *Nature,* LLMs can genuinely make relevant scientific contributions, and therefore master language use, but they "cannot take responsibility" for their output (Miller 2023, p. 7). The view presented here, though, suggests that exercising authority and taking responsibility is exactly what one must do in order to count as grasping and understanding what, in social practice, shows up as conceptual and propositional contents, and thus to count as a competent language user.

But why can't LLMs such as ChatGPT be held accountable in a normative sense for their output? Why can't we adopt normative attitudes towards them?[14]

What is needed for us to attribute them responsibility is to be in-principle possible for us to assess their behaviour (i.e., their attempts at rule-following) as right or wrong. In turn, what is needed for LLMs to count as responsible is to be in-principle possible for them to learn from their mistakes, to improve in response to (output) failures. But this means that it is to be in-principle possible for us to teach them by communicating wrongdoing, or better, train them by 'sanctioning' their wrongdoing, even punishing. Social sanction and reward, praise and blame, enable us to influence and regulate one another's behaviour: they get us to act, or get us to refrain from acting, so as to adjust our various actions in relation to one another. For example, we may get upset and blame our roommate for eating our leftovers without asking. Likewise, we jail tax evaders and drug dealers, and so on. In both cases, we aim to deter their behavior, and

[14] Taking a certain normative attitude is what Dennett (1987) calls "the intentional stance" and what Sellars characterizes as adopting a certain intention towards the beings in question, insofar as admitting a being bound by certain rights and obligations "is not to classify or explain, but to rehearse an intention" (Sellars, 1962, section VII).

prevent similar occurrens from occurring in the future. The same should be possible with LLMs.[15] Like our accountability practices toward fellow humans, we should be able to hold AI to account by rewarding or imposing sanctions. And to locate and attribute accountability, we should be able to demand answers and understand the commitments undertook by technology itself.[16]

Ultimately, we should see them as persons, where persons are agents individuated by their position in the normative relations into which they enter through adopting normative attitudes to one another: namely, the normative web of rights and obligations.

However, since the majority of the praise or blame for the actions of such machines should belong to the machines themselves not (or not only) to the programmers and designers, the problem lies in our inability to exact, say, punishment or reward upon machines. We cannot hold machines themselves responsible, given that they cannot appreciate such responses as punishment or reward. Indeed, even the most techno-optimists are concerned that machines' inability to suffer (and enjoy) thwarts our ability to directly hold them accountable—this is often why responsibility is ascribed indirectly, via the machines' associates.[17] Thus, it is certain that if we pursue this technology, then, in the future, highly complex interactive AI systems could perhaps be agents with corresponding rights and responsibilities. Yet the AIs of today still cannot be seen to be agents in this sense.[18]

Acknowledgements: This work was supported by the Slovak Research and Development Agency under the Contract No. APVV-22-0323

[15] For a similar view regarding AIs in general, see Allen and Wallach (2009) and Hellström (2013).

[16] David Shoemaker (2011) argues that our actual moral responsibility practice embodies three distinct conceptions: attributability, answerability and accountability.

[17] For example, Nyholm (2018, 2020) accepts that robots cannot suffer.

[18] Some recent work, though, supports the idea that AI systems will become increasingly able to recognize and learn from our morally significant reactions. See, e.g., Ren (2009) and Knight (2016).

(project 'Philosophical and methodological challenges of intelligent technologies'), Templeton project Ref. R64145/CN053 'The Necessity of God's Existence: Divine Names and Modality,' and Schwarz grant 2021/OV3/017.

Bibliography

ALLEN, C., WALLACH, W. (2009): *Moral machines: Teaching robots right from wrong*. Oxford University Press.
BENDER, E., KOLLER, A. (2020): Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185-5198.
BRANDOM, R. (1994): *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, Mass.: Harvard University Press.
BRENTANO, F. (1970): Psychology from the Empirical Standpoint. In: Morick, H. (ed.): *Introduction to the Philosophy of Mind: Readings from Descartes to Strawson. Scott.* Foresman: Glenview, Ill.
DENNETT, D. C. (1987): *The intentional stance*. Cambridge (Massachusetts): MIT Press.
ELKINS, K., CHUN, J. (2020): Can GPT-3 pass a writer's Turing Test?. In: *Journal of Cultural Analytics,* 2371, 4549.
FODOR, J. (1975): *The Language of Thought*. Harvard University Press.
FODOR, J. (1987): *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.
FREGE, G. (1892): Über Sinn und Bedeutung. *Zeitschrift Für Philosophie Und Philosophische Kritik,* 100, 25–50.
FREGE, G. (1918): Der Gedanke. *Beiträge zur Philosophie des deutschen Idealismus,* 2, 58–77.
GRICE, H. P. (1957): Meaning. In: *Philosophical Review,* 66 (3), 377-388.
HELLSTRÖM, T. (2013): On the moral responsibility of military robots. In: *Ethics and Information Technology*, 15 (2), 99–107.
KNIGHT, W. (2016): *Amazon working on making Alexa recognize your emotions*. MIT Technology Review.
KRIEGEL, U. (2011): *The Sources of Intentionality.* New York: Oxford University Press.
MILLER, R. (2023): Holding Large Language Models to Account. In: Müller, B. (ed.): *Proceedings of the AISB Convention*. Swansea: Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 7-14.

NYHOLM, S. (2018): Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. In: *Science and Engineering Ethics*, 24(4), 1201–1219.
NYHOLM, S. (2020): *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.
PEREGRIN, J. (2014): *Inferentialism: Why Rules Matter*. New York: Palgrave Macmillan.
PEREGRIN, J. (2021): Do Computers "Have Syntax, But No Semantics"?. In: *Minds and Machines*, 31, 305–321.
REN, F. (2009): Affective information processing and recognizing human emotion. *Electronic Notes in Theoretical Computer Science*, 225, 39–50.
SEARLE, J. (1983): *Intentionality*. Cambridge University Press.
SELLARS, W. (1957): Intentionality and the Mental. In: Herbert Feigl, Michael Scriven, and Grover Maxwell (eds.): *Minnesota Studies in the Philosophy of Science, Volume II: Concepts, Theories, and the Mind-Body Problem*. Minneapolis: University of Minnesota Press.
SELLARS, W. (1958): Counterfactuals, Dispositions, and Causal Modalities. In: Herbert Feigl, Michael Scriven, and Grover Maxwell (eds.): *Minnesota Studies in the Philosophy of Science, Volume II: Concepts, Theories, and the Mind-Body Problem*. Minneapolis: University of Minnesota Press, 225-308.
SELLARS, W. (1962): Naming and saying. In: *Philosophy of Science,* 29 (1), 7-26.
SHOEMAKER, D. (2011): Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121, 602-632.
STOKEL-WALKER, C. (2023): ChatGPT listed as author on research papers: many scientists disapprove. *Nat News*, 613 (7945), 620–621.
TURING, A. M. (1950): Computing Machinery and Intelligence. *Mind*, *New Series*, 59 (236), 433-460.
VASWANI, A. et al. (2017): Attention is All you Need. In: Advances in Neural Information Processing Systems, 5998-6008.
WITTGENSTEIN, L. (1953): *Philosophical Investigations*. Oxford: Blackwell.
WITTGENSTEIN, L. (1969): *On Certainty*. Oxford: Blackwell.

Mirco Sambrotta, PhD.
Institute of Philosophy, Slovak Academy of Sciences, v.v.i.
Klemensova 19
811 09 Bratislava 1
filomisa@savba.sk