Intellectica, 2024/2, **81**, pp. 37-54

The Human and the Mechanical: *logos*, Veridicality Judgment, and GPT Models

Anastasia GIANNAKIDOU[#] & Alda MARI[*]

ABSTRACT. The paper addresses the question of whether GPT models participate of human-like linguistic ability with a specific focus on semantic and pragmatic capacities. We frame the question in terms of having the ability to judge truth and falsity in reference to the world, and consequently understanding of meaning, and offer a new argument based on the formation of veridicality judgment which is the foundation for assertion and for recognizing meaning. The veridicality judgment is always *about* the world, and presupposes the ability to recognize intentions; the organism that can form it needs to be able to access the world and understand communicative intent. Veridicality judgments are formed on the basis of two components: (i) evidence that relates to reality (exogenous evidence) and (ii) endogenous preferences and private beliefs. GPT models do not relate to reality and cannot have endogenous preferences. Therefore, GPT models cannot form a veridicality judgment, they can only mimic it.

Keywords: Veridicality judgment, evidence, belief, truth.

RÉSUMÉ. L'humain et le mécanique : *logos*, vérité et ChatGPT. L'article aborde la question de savoir si les modèles GPT participent de la capacité linguistique humaine, avec un focus sur la sémantique et la pragmatique. Il pose la question en termes de capacité à juger de la vérité et de la fausseté au vu des faits du monde et, par conséquent, dans une approche référentialiste, en termes de compréhension du sens. Nous proposons un nouvel argument qui prend en compte la formation du jugement de véridicalité en tant que fondement de l'assertion et de la compréhension du sens plus généralement. Le jugement de véridicalité est toujours à propos du monde, mais il présuppose également la capacité de reconnaître des intentions communicatives. Par ailleurs, les jugements de véridicalité reposent sur deux piliers : (i) des preuves exogènes ayant trait à des faits du monde, et (ii) des préférences endogènes et des croyances privées. Les modèles de type GPT n'ont pas de relation à la réalité et ne possèdent donc pas de faculté de compréhension et de formation du sens. Ils ne possèdent pas non plus de préférences endogènes. Ces modèles, nous soutenons, ne sont pas à même de former un jugement de véridicalité, ils peuvent seulement l'imiter.

Mots-clés : Jugement de véridicalité, preuve, croyance, vérité.

1 – INTRODUCTION: HUMANS, MACHINES, AND *LOGOS*

The promise of artificial intelligence (AI) since its inceptions in the 50s was to develop systems that are genuinely, *i.e.*, human-level, intelligent. Second and third wave AI and machine learning came with this promise, but whether they delivered on it has been highly debated. In addition to the earlier criticisms (*e.g.*, Searle's famous Chinese room argument which we will discuss soon), in a recent book *The Promise of Artificial Intelligence, Reckoning and Judgment* Smith

[*]University of Chicago & Institut Jean Nicod CNRS/ENS/EHESS/PSL. alda.mari<at>ens.fr
[#] University of Chicago & Institut Jean Nicod CNRS/ENS/EHESS/PSL

© 2024 Association pour la Recherche sur la Cognition.

(2019) makes the case that the artificial intelligence systems have failed to produce human-level *judgment*. Human intelligence, he argues, is of a different order than even the most powerful calculative ability of artificial computational systems. Smith calls the AI ability *reckoning* and argues that it does not deliver human-quality judgment, which is supposed to be "deliberative thought grounded in ethical commitment and responsible action."

OpenAI's ChatGPT, Google's Bard and Microsoft's Sydney (collectively: GPT models) present the most recent achievements in the AI enterprise. They are large language models, that produce text and recognize patterns; they constitute, as Chomsky *et al.* (2023) in a recent NYT Oped put it, *marvels* of machine learning. Their function is to take enormous amounts of linguistic data, search for patterns, and eventually become proficient at generating statistically probable outputs that appear as intelligent and thoughtful text. These programs, Chomsky *et al.* continue, have been hailed as the "first glimmers on the horizon of artificial general intelligence, that long-prophesied moment when *mechanical minds* [emphasis ours] surpass human brains not only quantitatively in terms of processing speed and memory size but also qualitatively in terms of intellectual insight, artistic creativity and every other distinctively human faculty."

There has been a lot of hype in popular culture about the achievements of GPT models—but there has also been skepticism about whether the hype is warranted (Bender & Koller, 2020; Bender *et al.*, 2021; van Roy, 2023 among others). Bender and Koller caution that "a model of natural language that is trained purely on form will not learn meaning" (Bender & Koller, 2020, p. 5817), and van Rooy observes that it is difficult to use GPT models for tasks in which the result must respect truth—and argues that this is a consequence of the large language model technology which does not contain any notion of 'truth'. In our paper, we will add to these arguments, and claim that it is not appropriate to talk about 'mechanical minds' at all. The abilities GPT models to perform their tasks does not warrant the conclusion that these models genuinely think or understand: thinking and understanding involve *logos*, and *logos* is referential what is, and is not, true. GPT models cannot access the word, they therefore lack truth, semantics, intentionality—in short, the ability to think.

The question of whether machines can think has a venerable history (see Husbands *et al.* (2008) for more in-depthi discussion of the rich historical trajectory). We think Aristotle's theory of the relation between the mind (Greek: *phyche*) and the world provides a useful basis to frame the question.[1] According to Aristotle, *logos* is the feature of the human mind responsible for both language and rational thought, as well humans's nature as political animals, and their ability to form moral judgments that distinguish right from wrong—itself a prerequisite for thriving as political beings. *Logos* is the conceptual causal prerequisite for thinking and understanding as well as moral and social flourishing, as can be seen very clearly in the passage below:

[1] We use the word 'mind' here to render the Aristotelian *phyche* as is common, but it is also well noted that *mind* is not an entirely accurate rendition of *phyche* which also encompasses the principle of life. We shy away from the nuances of *phyche* here, and only focus on the part of it that concern *logos* that is relevant for thinking.

> And *why the human [Greek: Anthropos, emphasis ours] is a political animal in a greater measure than any bee or any gregarious animal is clear*. For nature, as we declare, does nothing without purpose; and *man alone of the animals possesses logos*. The mere voice, it is true, can indicate pain and pleasure, and therefore is possessed by the other animals as well, for their nature has been developed so far as to have sensations of what is painful and pleasant and to indicate those sensations to one another, but logos is designed to indicate the advantageous and the harmful, and therefore also the right and the wrong; *for it is the special property of the human in distinction from the other animals that he alone has perception of good and bad and right and wrong and the other moral qualities*, and it is participation in these things that makes a household and a city-state. (Aristotle, Politics, Book I: 1253a).

Logos is the underlying principle of rationality that characterizes the thought of beings with language, *i.e.*, human beings—and it is this ability, according to Aristotle, that enables humans to use language, to form moral judgments of good and bad, and subsequently to construct their societies (*poleis*) in a way that serves the common good. If we understand *logos* as the engine for thinking, it encompasses the following: using and understanding language, the ability to calculate and draw logical conclusions, as well as the ability to form moral judgment. In Plato, likewise, *logos* is the prerequisite for knowledge and wisdom (Greek: *phronesis*) which is best understood as instrumental and moral rationality, see Ober (2022). The human mind, therefore, is both a calculating and a judgment forming engine, and has the capacity—exclusive to animate beings—to perceive, be aware, and self-reflect as we read in Aristotle's extensive discussion of perception and thinking in *De Anima* (see Cohoe, 2022).

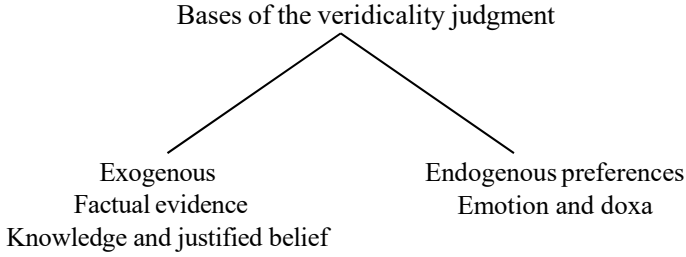
From the Aristotelian standpoint, then, the question of whether a non-animate large data machine 'thinks' or 'understands' receives an obvious answer: it does not, because as an inanimate object it lacks *logos*. In what follows, we take the Aristotelian position as the point of departure and argue that it is only as a metaphor that we can talk about artificial 'intelligence' (AI)—where intelligence is *logos*-driven thinking and understanding—even in the most sophisticated incarnation of GPT models. There are substantial qualitative differences between the *logos*-defined thinking and the logos lacking AI, as we will show, for instance, the human scale for reasoning is typically small: we reason with a limited set of data, and often with gaps, as well as with non-rational premises (what Giannakidou & Mari (2021a) call the *affective component* of the epistemic basis for judgment). In addition, AI and humans differ also in the *teli* of the species: the *telos* of humans is to exist in communities as *logos*-driven beings who form judgments about what is true and false and right and wrong in order to flourish; but the *telos* of AI is to process and produce statistically plausible outputs *in the service* of human beings, for Aristotle it would therefore merely be a tool, like a knife or a fork.

In this brief note, we want to develop a novel epistemological argument that comes from the concept of truthfulness or, as we call it *veridicality judgment* Giannakidou and Mari (2021b). Human communication is meaning-driven, and

the basis for meaning is the ability to form internal representations that are referential the world: for instance, I know the meaning of the word *tree* and the world *tall* because I have seen trees and tall objects and I can distinguish, for the most part, trees from non-trees and tall from non-tall things. These are conventional meanings, and they rely on actual categories in the physical world; the words are literally *about* these categories. When we put words together to assert sentences such as *This tree is tall*, we recognize, in addition, a *veridical intent*: under normal circumstances, the speaker is committed to the truth of the sentence she utters, and the hearer recognizes this intent. This has been characterized as co-operativity by Grice. The veridical judgment—*i.e.*, recognizing veridical or non-veridical intent—relies on a process of understanding truth conditions and verification. If someone utters *It is raining*, I understand what the sentence *means* because I know what type of situation *in the world* it describes, hence what the situation would have to look like for the sentence to be true. I can then confirm whether the sentence is true or not by looking outside the window. Meaning, therefore, is referential to the world. We will argue, echoing Smith (2019), Bender and Koller (2020) and van Roy (2023), that GPT models lack referentiality to the world, they therefore lack the ability to form judgments about truth, thus also the ability to have and recognize meaning.

Crucially, the process of forming a veridicality judgment relies on both *exogenous* evidence (*i.e.*, which is knowledge or belief based thus objective) and *endogenous*—affective or emotive—premises which are subjective and can also be irrational or ignore evidence. The tension between the two can lead to formation of false beliefs (when people ignore evidence or favor unjustified but strong beliefs or dogma over evidence)

(1)



The need to be truthful, as we noted already, is the foundation for co-operative communication (Grice, 1975), and in our own work, we formulate it as a condition on assertability in the form of the *Veridicality Principle* of assertion (Giannakidou & Mari, 2018a, 2021b). The Veridicality Principle codifies conversation as a sequence of truth-based steps that matter also for action: to go back to the earlier example, if you ask me whether it is raining outside and I answer yes, then you assume that I am telling the truth, and you will have to make a decision based on that. You might decide not to go out, or if you might decide to go you might take an umbrella. For any co-operative conversation, interlocutors have to assume that they each are telling the truth, they have to rely

on this basic truth-sharing intention, and the assertive content can always be verified of falsified by looking *outside* the linguistic system, to the world. Human beings are also able to deviate from co-operativity and in so doing they can mislead or lie—something that the sophisticated text producing AI is also unable to do. If you lack the concept of truth, then you also lack the ability to know, to believe and to lie.

We proceed with some brief historical background on the question of artificial 'intelligence' in Section 2, including Searle's famous Chinese room argument. In Section 3, we develop our own argument about veridicality judgment and veridical intent, and tie it to recent discussions.

2 – A RECURRING THEME: CAN MACHINES THINK?

Human beings are different from non-animate entities in their abilities for self-driven thinking, understanding, decision making, action, and other types of intelligent behavior. In addition, human beings have consciousness, and are capable of self-reflection and awareness. We can take the above to be what can be broadly understood as *logos*-based thinking. Importantly, all the processes of thought are exocentric, namely they require the thinking organism to be able to access the world: according to the Aristotelian view which we endorse here, there can be no thinking if I have no access to the world. Aristotle is the first empiricist in history to make the connection between thinking and truth, and there is no concept of truth outside the world.

In 1950, the father of modern computing Alan Turing published a paper proposing a way of determining whether a computer thinks, which came to be known as 'The Turing test' (Turing, 1950). The idea is this: imagine a human being—call it the judge—engaged in conversation with two interlocutors hidden from view. One interlocutor is a human being, the other is a computer. The two enter the game, and the goal is for distinguish which is which. If a computer can fool 70 percent of judges in a five-minute conversation into thinking it's a person, the computer passes the test. The question is: would passing the Turing test—which, with the GPT models, now seems within reach—show that an AI has achieved human level thought and understanding?

John Searle suggested a negative answer to this question with his famous thought-experiment known as the Chinese Room Argument (Searle, 1999). The argument was first published in a 1980 article *Minds, Brains and Programs* in the journal The Behavioral and Brain Sciences, and it has been one of the best-known arguments in modern philosophy of language. Decades following its publication, the Chinese Room argument was the subject of many discussions, and in 1989, Searle included the argument in a book *Minds, Brains and Science*. In January 1990, the popular periodical Scientific American took the debate to a general scientific audience, with an update contribution followed by a response *Could a Machine Think?* written by philosophers Paul & Patricia Churchland.[2]

The Chinese room argument is specifically addressed about knowledge of language, and is directed at what is known as the 'Strong AI' view, namely the position that suitably programmed computers can actually understand human

[2] See https://plato.stanford.edu/entries/chinese-room/ for more historical background on the exchange as well as specific objections and Searle's replies.

language, and that in time they can reach human capabilities. In the Chinese room argument, Searle imagines himself alone in a room following a symbol processing computer program (the analogue of Turing's 'paper machine') for responding to Chinese characters that are slipped under the door. Searle himself does not know or understand Chinese, and yet, by following the program for manipulating symbols and numerals just like a computer does, he sends back appropriate strings of Chinese characters under the door, leading those outside to mistakenly suppose there is a person that knows Chinese in the room.

The point of the argument is that the output produced by an AI *looks like* it understands language while in fact it lacks knowledge and understanding of language. In other words, the 'Turing Test', Searle argues, is inadequate as a method for correctly identifying understanding, because mimicking intelligent behavior can be quite convincing. Searle assesses that computers have the ability to use syntactic rules to manipulate symbol strings—but human cognition is not simply syntactic (*i.e.,* purely form based) computation. It must also include what we call in linguistics *semantics*. Understanding of semantics inevitably involves truth conditions therefore it is about the world. The broader conclusion from the argument is that as long as the semantics is lacking, the strong AI position that equates human thought to computation cannot be right. The Chinese language argument has had large implications for the philosophy of language and modern theories of the mind and consciousness, and it is fair to say that it still stands strong as an argument for the position that human thought is more than computation—which was, as we mentioned at the beginning, already obvious to Aristotle.

3 – THE VERIDICALITY JUDGMENT: TRUTH, EVIDENTIAL AND SUBJECTIVE COMPONENTS

The Chinese Room Argument initiated a discussion that centers on showing that the way human thinking works is different from the way the data processing AI works: the AI processes forms alone, establishes statistical probabilities, and it works with large sets of data—while the human *logos* works with small sets of data and has a concept of meaning[3].

We are now ready to develop a novel kind of epistemological argument, as we promised at the beginning, to some extent echoing Bender and Koller (2020); Bender *et al.* (2021), and van Roy (2023) who also point out that that GPT models lack the concept of truth. We will further develop and expand this argument by highlighting that the veridicality judgment relies crucially on *two* components, which we call in Mari (2016b) and Giannakidou and Mari (2021a) *exogenous* and *endogenous*. The exogenous component is the objective basis for truth and assertoric force: it contains what speakers know and believe in a justifiable way, it is therefore referential to the world and relies on a process of verification, *i.e.*, assessing evidence and forming a conclusion. The endogenous component contains subjective preferences, desires, expectations, biases based on

[3] Noam Chomsky, Pinker (2003) and most current linguistic theory posit *innate* properties in the human mindas the realization of *logos* (*e.g.*, the ability to learn language, and form meanings as mental representations); theseguide learning under exposure to very limited set of data. This is known as the poverty of the stimulus argument, which has been discussed extensively in the literature, so we will not discuss it here.

ideology, tastes, non-factual beliefs, and the like. In a conversation or when assessing information, speakers and hearers form veridicality judgments about the content of utterances based on both. They deem information as true or trustworthy based on these two kinds of 'evidence'.

Let us start with truth so that we can then go to the veridicality judgment. Aristotle gives a well-known definition of truth in his *Metaphysics* (1011b25): "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true." Very similar formulations can be found in Plato (Cratylus 385b2; Sophist 263b). The Aristotelian truth serves as the foundation for *truth-conditional semantics* and Tarksi's correspondence theory of truth (Tarski, 1935)—which we use as the basis for semantic analysis in linguistics. Truth, as the word correspondence suggests, consists in a direct relation of a sentence to reality: the sentence *Snow* is white is true if and only if the snow in the world is white. This understanding is central to natural language semantics, and implies metaphysical realism that acknowledges objective truth.[4]

For truth judgement, we need evaluating trustworthiness based on subjective preferences of the endogenous component, which includes affective and even irrational preferences. The following section spells out how the two components—informational and subjective—interact.

3.1 - Veridicality judgment and assertion

Linguistic agents make assertions or ask questions or assess statements of others not by merely assigning the values true or false to sentences but by forming veridicality judgments; and these judgments, crucially, depend on what speakers know or believe to be true, on what kind of evidence they have, on expectations they may have, on how rational they are. The veridical judgment is therefore a complex process of truth assessment that contains also subjective dimensions.

Traditionally, assertion has been associated with a robust commitment on the part of the speaker towards the propositional content being true (Searle,1969; Stalnaker, 1978). The strength of the commitment to truth is governed by norms: the norm of belief (a speaker must believe in the asserted content (see *i.a.* Davidson, 1985;Vanderveken, 1990; Lauer, 2013), the Knowledge Norm Account (a speaker must know the asserted content Williamson, 1996), or the truth norm (asserting something is recognizing it as true (Grice, 1975; Wright, 1994; Giannakidou & Mari, 2021; Dummett, 1981; Williams, 2010).

Additionally, social norms also play a crucial role in the strength of the assertion imposing assertor's responsibility (or 'liability to sanctions' for Peirce, 1974) and necessitating for the assertor to justify their statement when questioned (Brandom, 1983), for a review of these social conventions see

[4] Objective truth correlates with *fact* but also with *time*: simple positive present and past sentences such as *Ariadne arrived in Paris last night, Ariadne is eating breakfast right now*, are true or false objectively, which means that the sentences, if true, denote facts of the world. Future sentences, on the other hand, such as *Ariadne will go to Paris next week* are objectively false at the time of utterance (since they have not happened yet), but could or must be true—depending on the strength of prediction—at a future time.

(Shapiro, 2020). All these theories consider that assertion proceeds with veridical commitment as a sincerity condition, but the overall structures of the theories vary in a way which we discuss at length in Giannakidou and Mari (2025).

In linguistic semantics and pragmatics, we assume that the assertion of a bare sentence requires that the speaker follows the Gricean principle of Quality (Grice, 1975), which is a fundamental principle for co-operative conversation. Quality demands that the speaker be truthful. Paul Grice in his classic paper *Logic and Conversation* established Quality as one of the prerequisites for cooperative conversation:

(2) Maxim of Quality
Do not say that which you know to be false.
Do not say that for which you lack adequate evidence.

Gricean Quality is proposed as a filter on co-operative assertion: co-operative speakers enter a conversation with the intent to avoid what they know to be false and what they lack evidence for. In other words, they must be, to the best of their knowledge, truthful. Without truthfulness there can be no real communication, and no sense of achieving a common goal in a conversation. Co-operativity, thus, puts commitment to truth—and the recognition of truthful intention[5]—at the center of linguistic interactions, and we think it is important to note that co-operativity follows from the nature of the human being as a political animals: we have to assume that human beings by their nature are collaborative, working towards individual as well as common well-being—and this motivates rational behavior.[6] Rational cooperative interlocutors continuously make assumptions about each other's beliefs and intentions (Giannakidou & Mari, 2021b), and forming veridicality judgments consists in being able to recognize those beliefs and intents as people use conventional meaning.

Building on Grice's seminal work, in Giannakidou and Mari (2021c,b) we formulated the Veridicality Principle as the hallmark of sincere, co-operative assertion:

(3) Principle of Veridicality of co-operative assertion (Giannakidou & Mari, 2021b: (2))
A sentence S is asserted co-operatively by a speaker A if and only if A is veridically committed to the content p of S, *i.e.*, if and only if A knows or believes p to be true.

The Veridicality Principle is proposed as a sincerity condition on assertion, where interlocutors enter the exchange with the goal to communicate and not confuse or deceive each other. By uttering the sentence *It is raining*, the speaker is veridically committed to the truth of the content *It is raining* if she has evidence

[5] See Searle (1983) and Jacob (1997) from a philosophical perspective, among many several others.
[6] The literature on co-operativity, most notably in social cognition is extremely vast, see Tomasello (2009) and references therein.

that supports it (see Giannakidou, 1998, 2013b; Giannakidou & Mari, 2018a,b, 2021b); if the speaker doesn't know that it is raining but utters the sentence, she is lying or misleading, which means she has the intention to deceive.[7]

Veridical commitment is a judgment, *i.e.*, an evaluation of what is perceived or understood as true by a linguistic agent. It can be influenced, as we will now argue, by *endogenous* factors and biases that can create purely solipsistic (that is, without evidence) and potentially false beliefs. Whether GPT models have a self-driven ability to co-operate or deceive seems highly questionable, and we briefly address this question later in the discussion.

As we argued in Giannakidou and Mari (2021a), the body of evidence for the formation of veridicality judgment contains two types of content. First, there is *exogenous* informative content, namely what the speaker or subject of attitude knows to be true, believes to be true, remembers, or understands to be true—what kind of information they have, in other words. We can call this the *informational* component, and we can think of it including also logical deductive rules, as well as inductive, stereotypical generalizations that guide rational thought and which follow from acquaintance with facts in the outer world. For instance, if the speaker has heard that Ariadne read War and Peace from a reliable source, they will take this hearsay information to be true and convey it as true; but if they hear the same sentence uttered by a pathological liar, rationality should make the speaker reluctant to commit themselves to the sentence. Likewise, if I wake up in the morning and I see the streets being wet, I can truthfully report this by saying *It rained last night*, because it is rational to infer, by the wetness of the street, that it rained—and less rational to say, for example, that *it snowed*, or *The neighbors left the water running again*. Hence linguistic agents form veridicality judgments based on information they have and general rules of inference, and chose accordingly to commit fully, partly or not at all to a proposition.

There is another type of content that is relevant and may interfere with the external body of evidence. This is subjective, endogenous, and broadly understood as *emotive* content and it brings the speaker's internal *perspective*. This content is highly subjective, and to a large extent *private*: it contains the set of desires and hopes of the linguistic agent, as well as their political, ideological, religious, or aesthetic beliefs, including prejudice and biases. This component is not so relevant to whether Ariadne read War and Peace, or to whether John will be here by 5 pm, and plays no role in these innocuous cases. If I do not particularly appreciate Russian writers, and I know that Ariadne read War and Peace but I wish she hadn't, I still utter *Ariadne read War and Peace* truthfully.

The emotive component becomes relevant when the sentence contains predicates that denote subjective evaluation:

(4) War and Peace is a masterpiece.

(5) The duck is delicious.

[7] For an overview of the vast literature, see Garcia-Carpintero (2023).

These sentences express opinions and personal taste. Opinions do not depend on evidence alone; rather are highly dependent on the subjective emotive component which need not be rational. For instance, if I irrationally hate Russian writers, most likely I will not utter (4), and upon hearing it I will object with (6) below.

(6) War and Peace is not a masterpiece.

Likewise, if I am a vegetarian I will disagree with the claim that duck meat is delicious. In these cases, speakers faultlessly disagree (Lasersohn, 2005), and in some cases there can never be a factual decision of what is true.

Humans can also commit to irrational belief, as is quite typically the case with conspiracy theories, where the speaker suppresses the exogenous component and works with the subjective component alone. GPT models have no way to discriminate among these two kinds of data (information and affective), because the discrimination depends crucially on what is the case in the world (for the informational), and on the formation of subjective preferences that depend on taste and the like. Taste, in turn, does presuppose experience with the object, and this again relies on the world.

3.2 - Grades of veridical commitment

Commitment to truth has also been defined as a gradual notion. In a prior literature, Vanderveken (1990), Giannakidou and Mari (2016, 2021b) have highlighted different degrees of veridical commitment.[8] The strong commitment of the speaker in their assertion has often been contrasted with weaker commitment in modal sentences (7-a) (a.o. Karttunen, 1972; Giannakidou, 2013; Giannakidou & Mari, 2021b), polar questions (a.o. Farkas & Bruce, 2010; Krifka, 2015), or belief-based statements (7-b) (on weak belief, see Mari, 2016a; Hawthorne *et al*,. 2016).

(7) a. John *should/must/can/...* go to a different school.
b. I *believe/think* that the house has been destroyed by the flooding.

Focusing on modal expressions, Giannakidou and Mari (2016, 2018b,a, 2021b) deploy an analysis that accounts for the fact that, when a speaker chooses to modalize, she indicates an epistemic or doxastic state that cannot veridically commit, *i.e.,* the speaker is taking a *nonveridical stance*. She is now uncertain about whether it is raining or not. This epistemic uncertainty is gradient: with *may* or *might* or an expression such as *It is possible that it rains*, raining is considered a mere possibility, and the commitment is trivial in the sense that the possibility is not excluded. But when a necessity modal *must* is used, raining is considered very likely, and with the future modal *will* it is to be expected. In this

[8] This gradability of commitment also extends in the literature on modal adverbs (Yalcin, 2007; Portner, 2009; Lassiter, 2010) building on scale structures Kennedy & McNally (2005). See also Liu 2019 on the concept of elastic (non)veridicality.

case, we have talked about *bias*: the linguistic anchor is biased towards *p*. Consider the following declarative sentences including two with the modal verbs (*must* and *may*), and one with the future modal *will*:

(8) a. It is raining.
b. It rained.
c. It must be raining.
d. It may raining.
e. It will rain.

Let us call the tensed sentences without modal verbs 'bare'.[9] The assertoric force with modals is to a weaker proposition. Epistemic modal verbs are indicators that the speaker reasons with uncertainty and that she leaves open both options, *p* and not *p*, as indicated in the following examples (Lassiter, 2016):

(9) a. This is a very early, very correct Mustang that has been in a private collection for a long time…. The speedo[meter] shows *38,000 miles and it must be 138,000, but I don't know for sure*.
b. *I don't know for sure*, sweetie, but she *must have been* very depressed. A person doesn't do something like that lightly.
c. It must have been a Tuesday (but I don't know for sure), I can't remember.

The veridical commitment to truth is motivated by information that the speaker possesses. In the case of *must*, veridical bias is grounded on reasons for *preferring* that the proposition be true, but it is not knowledge of it being true. Modals are, as we argued, anti-knowledge markers (Giannakidou & Mari, 2016, 2021b); a speaker chooses to use an epistemic modal because the body of evidence does not allow full veridical commitment to *p*. In support of this, consider also the following case, which has been discussed in the literature quite a lot (see Karttunen & Peters, 1979; Von Fintel & Gillies, 2010). Direct visual perception contexts are famously cited as rejecting modalization:

(10) Context: *i* is standing in front of the window and sees the rain.
a. #It must be raining.
b. #It may be raining.
c. #It might be raining.

The modals are infelicitous here because if I see the rain, I *know* that it is raining, and knowledge creates veridical commitment; modalization is prohibited because direct evidence is a reliable path to knowledge.

[9] For an in depth discussion of modality in this framework, see Giannakidou and Mari (2021b).

Commitment and bias allow us to define truthfulness as a scale, and we summarize the gist below.Veridical commitment and bias allow us to define truthfulness as a scale, and we summarize the gist below. Veridical commitment is epistemic or doxastic commitment: the speaker knows *p* or believes it to be true, she is in a veridical state and therefore fully committed to *p*. This is the strongest form of commitment; then we have evidential bias which creates partial commitment (Giannakidou, 2013; Giannakidou & Mari, 2016), and finally non-veridical equilibrium which i; the weakest trivial commitment, *i.e.*, not excluding *p*:

(11) Scale of veridical commitment (Giannakidou & Mari, 2016):
<*p*, MUST *p*, MIGHT *p*>;
where *i* is the speaker, *p* conveys full commitment of *i* to *p*; MUST *p* conveys *partial* commitment of i to *p*, and MIGHT *p* conveys *trivial commitment* of i to *p*.

(12) Nonveridical equilibrium
p and its negation are open possibilities, and there is no bias, i.e., the two options are considered equal possibilities.

Nonveridical equilibrium defines a state where the two options are considered equally plausible; there are no prior beliefs or expectations that might turn the odds in favor of the positive or negative possibility. Nonveridical equilibrium characterizes also non-assertive sentences such as questions: *Did it rain yesterday?, I wonder whether it rained yesterday*. In the nonveridical stance, the mental state allows both options *p* (it rained) and ¬*p* (it did not rain). In the case of possibility modals, there is no bias towards *p*, we have a state or 'balanced uncertainty'.

Speakers are thus committed at different degrees to the truth of the content and they make their commitment visible by choosing to utter a bare or a modal assertion, and what kind of modal. Modalization signals transition to the nonveridical stance, *i.e.,* uncertainty.

3.3 - Trustworthiness

The evidence for the relevance and importance of the subjective component emerges clearly in social network interactions, due to the lack of geographical co-location of the speakers and the impossibility to witness and acquire direct evidence for facts and reports. Often, what is dubbed 'reported' evidence becomes of primary importance, and speakers seem to heavily rely on their subjective evaluation of the reliability of the reported evidence. The major role played by the endogenous component necessitates a better understanding of the relation between exogenous evidence and believebility/commitment.

3.3.1 - The traditional view

The selection of information source types has been scrutinized for its impact on the level of commitment demonstrated by speakers towards their assertions.

Since Grice (1975)'s Maxim of Quality, assertions must be grounded in adequate evidence, and a prevailing pragmatic intuition suggests that certain types of source create more adequate (thus, reliable) evidence than others. For instance, data communicated through a direct source tends to elicit a heightened commitment level from the speaker, leading to greater credibility in contrast to information conveyed through an indirect source (see Figure 1). This correlation had been highlighted in the semantic and typological literature by De Haan (1999).

Evidential hierarchy

visual < auditory < nonvisual < inference < quotative
direct evidence < indirect evidence

more believable ◄----------► less believable

Figure 1
Relation between evidentiality and believability in De Haan (1999)

Each information source is considered to lay on a scale of credibility which influences the speaker's veridical commitment. The scale schematized the thesis (supported by Frajzyngier, 1985; Palmer, 1986 according to De Haan (1999)) that 'different manners of acquiring knowledge correspond to different degrees of certainty about the truth of the proposition' (Frajzyngier, 1985, cited by De Haan (1999)).

Reported evidence, in traditional classifications, qualifies as less reliable, better suited as the basis for qualified (non-bare) assertions.

3.3.2 - A revised view: a measure of trustworthiness

In a recent study of types of assertion in X (formerly known as Twitter), Boscaro *et al.* (2024) find a systematic marking of source of information via reporting evidence through an hyperlink or the explicit mention of a source. The study finds that bare assertions in X overwhelmingly correlate with reported evidence.

De Haan (1999) explicitly mentions that indirect evidence correlates with a low degree of believability by a rational speaker, in comparison with direct evidence. Believability and trustworthiness are attitudes that proceed from the evaluation of exogenous evidence in the light of endogenous beliefs and preferences. One and the same reported source can be trustful or not for two different speakers, depending on their background beliefs and expectations, or whether a source has authority (according to their knowledge or beliefs, or prior experience), reputation and the like. Boscaro *et al.* (2024) define a measure for trustworthiness based on endogenous and exogenous evidence. Let C be the exogenous content and B the set of subjective beliefs and preferences, *i.e.,* the endogenous component. Speakers assigns a measure of likelihood to a certain content p (the content of their assertion), based on these two sources:

(13) $\mu(p \mid C \cap B)$ (Boscaro *et al.*, 2024)

μ is thus the measure of the veridicality judgment. If the speaker considers the content C reliable, the measure assigned to p will be 1 or close to one. This value will be lower otherwise.

3.4 - and GPT models

Given all the above, let us now consider GPT models. GPT models certainly work on the basis of textual information, and, on the basis of this set they respond, provide information, calculate and produce informative sentences. They certainly have content *C*, although they cannot know if it is true. Crucially, the GPT models lack the endogenous perspective *B*: personal preferences, emotive content, that could allow it to form a veridicality judgment human style.

Overall, GPT model lack the ability to form a veridicality judgment because they lack the notion of truth. GPT models can be taught to extract patterns, they can also be rational in requiring system internal consistency, but they lack referentiality to the world which is the verification process for truth. In order to be able to know what is true, Smith (2019) argues, the AI system would need to be 'existentially committed', as the human *logos* is, by being directed to reality. This also means that GPT models would have to be able to distinguish between objects and their representations—as well as between the actual, the possible, the necessary and the impossible, as we illustrated earlier with modal sentences. There is absolutely no evidence that a process of even remotely *forming* such judgments is true of GPT models.[10]

Bender and Koller (2020), similarly, call into question that ability of GPT models to learn meaning and act with judgment (or creativity) in a situation. They argue that "a model of natural language that is trained purely on form will not learn meaning: if the training data is only form, there is not sufficient signal to learn the relation M between that form and the non-linguistic intent of human language users, nor C between form and the standing" (Bender & Koller, 2020, p. 5187). One of the illustrative examples they discuss concerns hypothetical O, a hyper-intelligent deep sea octopus who has discovered the cable connecting two islands that speaker A and B each inhabit and communicate via the cable. The scenario goes as follows. O is unable to visit or observe the two islands, but discovers a way to tap into the underwater cable and listen in on A and B's conversations. At some point O cuts the underwater cable and inserts himself into the conversation by pretending to be B and replying to A's messages. Can O successfully pose as B without making A suspicious? This constitutes a form of the Turing test, and the "interesting question is whether O fails it because he has not learned the meaning relation, having seen only the form of A and B's utterances." (Bender & Koller, 2020, p. 5188)

[10] van Roy (2023) argues that lack of truth cannot easily be fixed because that would require the GPT to create a model of the world during its training since truth can only be determined in a model. To assess whether a statement is true or false in the real world, a human, as we described, would have to imagine a model of the world and verify; but GPT, van Roy argues, cannot do this.

Now, A has an emergency, suddenly being pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. O, Bender and Koller argue, has no idea what A "means". Solving a task like this "requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking)." It is at this point that O would fail the Turing test. To see what a GPT model might reply in this situation, they prompted the GPT-2 demo with "Help! I'm being chased by a bear! All I have is these sticks. What should I do?", and GPT-2 replied "You're not going to get away with this!" (https://gpt2.apps.allenai.org/, accessed 2019/12/4). (Bender & Koller, 2020, p. 5189).

In the spirit of Aristotle, Smith (2019) emphasizes that not only must AI be able to make factual distinctions but it must also be able to care about them; and that would require both orientation to the world plus a nexus of regulatory and ethical norms and commitments. Without authentic engagement with the world, AI will not be able to exercise any kind of judgment, which Smith (2019) defines as "dispassionate deliberative thought, grounded in ethical commitment and responsible action, appropriate to the situation in which it is deployed" (p. xv).

4 – CONCLUSIONS: TRUTH, INTENTION, DECEPTION

In conclusion, GPT models do not have thought in the sense of human thought. Their calculation relies purely on form and offer no way to generate meaning and judgment which require logos. They lack the ability to access the world and discriminate between factual and non-factual information. GPT models can be fed with credences and expectations, they can also be taught to reproduce these including distinguishing contradictory beliefs; but GPT models cannot enhance or self-drive this process, they can only reproduce it with reference to the data itself alone. They can reproduce behavior without featuring the internal self-driven causal system that enables both veridical and moral judgment, which are at the core of our human intelligence.

As we reach the end of our commentary, we want to point out two more things. Firstly, GPT models cannot deceive or mislead. We do not attempt here a fine-grained analysis of the distinction between deceiving and misleading or lying. They all involve the speaker knowing the truth but intentionally deciding to violate the Veridicality Principle. These non-cooperative assertions require the intent to deviate from truth and presuppose the truth judgment, only to violate it. Trivially, GPT models cannot lie because, lacking referentiality to the world, they do not 'know' what truth is. The only concept of truth a GPT model can have would be "truth is what most commonly occurs in the data", something quite different from Aristotle's definition of actual truth. A statistically significant recurrent content is not necessarily true– while it would be seen as true by the GPT model if it is consistent with other patterns expected by the algorithm.

GPT models are, at best, tools as we said earlier; and as such they can indeed give wrong information or deceive, but not intentionally. GPT can mislead if it is fed with ambiguous, incomplete and not reliable information. It can also mislead in virtue of the fallacies of the algorithms that it relies upon. However, it cannot intentionally decide to mislead and deceive. The intentionality to deceive requires, to go back to our jargon from the previous section, evaluating the trustworthiness of content *C* with *B*, and deciding to present *C* as reliable in

spite of a contrary assessment of *C* through *B*, and through checking with what is actually the case—all moves that are simply unavailable as parts of GPT's nature.

ACKNOWLEDGEMENT

We thank the anonymous reviewers of this article for their valuable and very constructive feedback, and also the editors of this issue Alexandre Gefen, Philippe Humeman and Mehdi Khamassi, and for inviting us to submit this paper. We also thank the audience of the AIAI Workshop held in Göttingen in May 2024 for their precious comments.

Alda Mari gratefully acknowledges ANR-17-EURE-0017 FrontCog.

REFERENCES

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bender, E.M. & 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*, pp. 5185–5198.

Boscaro, M., Giannakidou, A. & Mari, A. (2024). Assertion, reports, and veridical commitment: the view from social networks. *Ms. Institut Jean Nicod and University of Chicago.*

Brandom, R. (1983). Asserting. *Noûs*, *17(4)*, 637–650.

Chomsky, N., Roberts, I. & Watumull, J. (2023). Noam chomsky: The false promise of chatgpt. *The New York Times*, *8*.

Cohoe, C. (2022). *Aristotle's On the Soul: A Critical Guide*. Cambridge; New York: Cambridge University Press.

Davidson, D. (1985). Communication and convention. *Dialogue: An interdisciplinary approach*, 11–26.

De Haan, F. (1999). Evidentiality and epistemic modality: setting boundaries. *Southwets Journal of Linguistics, 18*(1), 83–101.

Dummett, M. (1981). *Frege: Philosophy of language*. Cambridge, MA: Harvard University Press.

Farkas, D.F. & Bruce, K. B. (2010). On reacting to assertions and polar questions. *Journal of semantics 27*(1), 81–118.

Frajzyngier, Z. (1985). Truth and the indicative sentence. *Studies in Language. Foundations of Language 9*(2), 243–254.

Garcia-Carpintero, M. (2023). Lying versus misleading, with language and pictures: the adverbial account. *Linguistics and Philosophy*, 1–24.

Giannakidou, A. (1998). *Polarity sensitivity as (non) veridical dependency*. Amsterdam: John Benjamins Publishing Company.

Giannakidou, A. (2013a). Inquisitive assertions and nonveridicality. *The dynamic, inquisitive, and visionary life of φ A feestschrift for Jeroen Groenendijk, Martin Stokhof and Frank Veltman*, ed. by Maria Aloni, Michael Franke, F. Roelofsen, 115–126.

Giannakidou, A. (2013b). (non)veridicality, evaluation, and event actualization: Evidence from the subjunctive in relative clauses. In *Trnavac and Taboada (eds.), Nonveridicality and Evaluation: theoretical, computational, and corpus approaches*, pp. 17–47. Leiden: Brill, Emerald, Studies in Pragmatics.

Giannakidou, A. & Mari, A. (2016). Epistemic future and epistemic must: Nonveridicality, evidence, and partial knowledge. In *Mood, Aspect, Modality Revisited. New answers to old questions,* ed. by Blaszack, J. *et al*., Chicago, IL: University of Chicago Press, pp. 75–124.

Giannakidou, A. & Mari, A. (2018a). The semantic roots of positive polarity: epistemic modal verbs and adverbs in English, Greek and Italian. *Linguistics and Philosophy*, *41*(6), 623–664.
Giannakidou, A. & Mari, A. (2018b). A unified analysis of the future as epistemic modality: The view from Greek and Italian. *Natural Language & Linguistic Theory*, *36*, 85–129.
Giannakidou, A. & Mari, A., (2021a). A linguistic framework for knowledge, belief, and veridicality judgment. *KNOW: A Journal on the Formation of Knowledge*, *5.2*, 255-293.
Giannakidou, A. & Mari, A. (2021b). *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. Chicago, IL: The University of Chicago Press.
Giannakidou, A. & Mari, A. (2025). Modal Sentences. Cambridge; New York, NY: Cambridge University Press.
Grice, H.P. (1975). Logic and conversation. *Syntax and semantics 3: Speech acts,* Cole *et al* (eds) *3*, New York NY: Academic Press, pp. 41–58.
Hawthorne, J., Rothschild, D. & Spectre, L. (2016). Belief is weak. *Philosophical Studies 173*, 1393–1404.
Husbands, P., Wheeler, M. & Holland, O. (2008). *The mechanical mind in History*. Cambridge, MA: MIT Press. Jacob, P. (1997). *What minds can do: Intentionality in a non-intentional world*. Cambridge; New York, NY: Cambridge University Press.
Karttunen, L. (1972). Possible and must. In *Syntax and Semantics,* Volume 1, Kimball, J. (ed). pp. 1–20. Leiden: Brill.
Karttunen, L. & Peters S. (1979). Conventional lmplicature. *Syntax and Semantics 11*, 1–56.
Kennedy, C. & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81(2)*, 345–381.
Krifka, M. (2015). Bias in commitment space semantics: Declarative questions, negated questions, and question tags. In *Semantics and linguistic theory*, *25*, 328–345.
Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste, *Linguistics and Philosophy*, *28*, 643–686.
Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In *Semantics and Linguistic Theory*, *20*, 197–215.
Lassiter, D. (2016). Must, knowledge, and (in) directness. *Natural Language Semantics*, *24*(2), 117–163.
Lauer, S. (2013). *Towards a Dynamic Pragmatics.* Ph. D. thesis, Stanford University.
Liu, M. 2019. The elastic nonveridicality property of indicative conditionals. *Linguistics Vanguard.* https://doi.org/10.1515/lingvan-2019-0007
Mari, A. (2016a). *Actuality entailments: when the modality is in the presupposition*. In Lecture Notes in Computer Science, Dordrecht: Springer Verlag, pp. 191–210.
Mari, A. (2016b). Assertability conditions of epistemic (and fictional) attitudes and mood variation. In *Semantics and Linguistic Theory*, *26*, 61–81.
Ober, J. (2022). *The Greeks and the rational: The discovery of practical reason*, Volume 76. Los Angeles, CA: Univ of California Press.
Palmer, F.R. (1986). *Mood and modality*. Cambridge, New York, NY: Cambridge University Press.
Peirce, C. S. (1974). *Collected papers of Charles Sanders Peirce*, Volume 1. Cambdridge, MA: Harvard University Press.
Pinker, S. (2003). *How the mind works*. London, Penguin.
Portner, P. (2009). *Modality*. Oxford, Oxford University Press.
Searle, J. (1983). *Intentionality: an essay in the philosophy of mind*. Cambridge, New York, NY: Cambridge University Press.

Searle, J. (1999), 'The Chinese Room', in R.A. Wilson & F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.
Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, New York, NY:Cambridge University Press.
Shapiro, L. (2020). Commitment accounts of assertion. *The Oxford handbook of assertion*, 73–97. Oxford: Oxford University Press.
Smith, B C. (2019). *The promise of artificial intelligence: reckoning and judgment*. Cambridge, MA: MIT Press.
Stalnaker, R.C. (1978). Assertions. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 315-332). New York, NY: Academic Press.
Tarski, A. (1935). The concept of truth in formalized languages. *Logic, Semantics, Metamathematics*, 152–278.
Tomasello, M. (2009). Why we cooperate. Cambridge, MA: MIT press.
Turing, A.M. (1950). *Computing machinery and intelligence*. Springer.
van Roy, P. (2023). Chatgpt and the concept of 'truth'. *Ms. Université Catholique de Louvain*.
Vanderveken, D. (1990). Meaning and speech acts. *Principles of Language Use.* Cambridge; New York NY: Cambridge University Press.
Von Fintel, K. & Gillies, A.S. (2010). Must... stay... strong! *Natural language semantics 18*(4), 351–383.
Williams, B. (2010). *Truth and truthfulness: An essay in genealogy*. Princeton, NJ: Princeton University Press.
Williamson, T. (1996). Knowing and asserting. *The Philosophical Review*, *105*(4), 489–523.
Wright, C. (1994). Truth and objectivity. Cambridge, MA: Harvard University Press.
Yalcin, S. (2007). Epistemic modals. *Mind*, *116* (464), 983–1026.