Truth judgment, internal states and LLMs

Anastasia Gianakidou and Alda Mari
University of Chicago and Istitut Jean Nicod

Göttingen, AIAI Workshop

Plan

Mechanical logos ?

The structure of human judgment of veridicality

Commitment, evidence, and informativity

Pause : how about LLM

Non cooperative conversations

Mechanical and human minds : background

- The promise of artificial intelligence (AI) since its inception in the 50s was to develop systems that are genuinely, i.e., human-level, intelligent.
- Smith 2019 : artificial intelligence systems have failed to produce human-level intelligence and, he emphasizes, *judgment*.
- Judgment is supposed to be : “deliberative thought grounded in ethical commitment and responsible action.”

Mechanical and human minds : background

- LLM produce text and recognize patterns
- 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.

Our questions

- It is appropriate to talk about ‘mechanical minds’ at all ?
- What is a mechanical ‘mind’ ? Are the abilities of machines to perform calculations or generate probabilities enough to talk about the machines having a mind ?
- Is the human mind simply a better version of the mechanical mind ?

Logos

- *Logos* vs. *Psyché* : *Phyché* is a form of *energeia*. *Logos* is the ground for the formation of *judgment* (epistemic judgment, and moral judgment)
- Animals have a *phyché* (Aristotle : De anima) too, but not one with *logos*,
- *Logos* is the conceptual causal prerequisite for human thinking, and moral and social flourishing, the dual ability to speak and the ability to think rationally

Logos

- *Logos* is the underlying principle of rationality that characterizes the thought of beings with language, i.e., human beings
- *Logos* enables humans to form moral judgments of good and bad, and subsequently also able to apply these judgments to their societies (*poleis*) in a way that serves the common good.
- The human mind, via *logos*, is both a calculating and a judgment forming moral engine, and in addition it also has the capacity— exclusive to animate beings— to perceive, be aware, and self-reflect

IA weak position

Searle :

‘Weak AI’ : AI is merely a useful tool for gathering and analyzing data because it *simulates* human abilities. In this view, we cannot simply transfer the conclusions from AI to human cognition because they are qualitatively different.

We align with this.

IA strong position

- Strong AI is the position that suitably programmed computers can understand human language (often called natural language), and that they have other mental capabilities similar to the humans. Computers really do play chess intelligently, make decisions, or understand language.
- Human mind as merely a better version of the AI ‘mind’, which at some point in the future and with more data and better programming, it will approximate more the human way of thinking.

We reject this.

Our goal

- LLMs lack the structure of the human judgment.
- LLMs lack the evidential basis of the human judgment.
- LLMs therefore lack the ingredients for veridicality judgment : to ability know what is true and what is false, because both require (i) the ability to connect to the reality (exogenous component), and (ii) evaluate it (endogenous component).

LLM lack *logos*

Plan

Mechanical logos ?

The structure of human judgment of veridicality

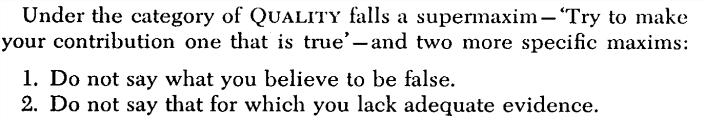
Commitment, evidence, and informativity

Pause : how about LLM

Non cooperative conversations

The simple case : cooperative conversations (Grice)

Famously, according to Grice (1975), a cooperative and effective conversation is regimented by four principles or maxims, among which Quality



Under the category of QUALITY falls a supermaxim—'Try to make your contribution one that is true'—and two more specific maxims:

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Truthfulness a precondition for assertion

Truthfulness = Veridicality

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

Veridicality is a sincerity condition for assertion, and commitment is knowledge or belief of p (the belief and the knowledge norm Williamson 2000).

Veridical commitment as knowledge of belief

(2) Veridical commitment as knowledge
A linguistic agent *i* is veridically committed to a proposition *p* iff *i* knows *p*.

(3) Subjective veridical commitment
A linguistic agent *i* is subjectively committed to a proposition *p* iff *i* believes *p* to be true.

- Knowledge relies on factual evidence (exogenous)
- Belief is grounded in assessing factual evidence, along with endogenous subjective factors such as personal preferences, tastes, expectations, prior beliefs.
- When we assess truth, we typically use a mix of knowledge and subjective factors

Bare assertion

(4) John is at home right now

I am committed to the truth of p= I know, have factual evidence that p is true.

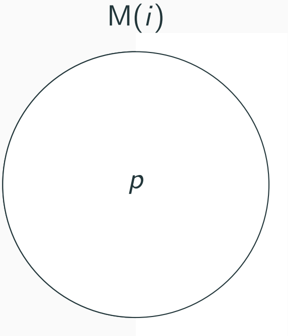


Figure 1: Bare assertions

Nonveridicality

Assessing evidence, is a form of evaluation using mixed evidence .

Nonveridicality = unable to meet the sincerity condition, do not know, do not have enough evidence that *p* is true (Giannakidou and Mari 2015,2018a,b, 2021a,b a.o)

(5) The non-veridicality state
A non-veridical state entertains two possibilities p and ¬p.

The use of modals (might,must) and other subjective markers reflects the non-veridical state.

Modals are anti-knowledge markers.

Non-veridical equilibrium

(6) John might be at home

State of uncertainty where p and $\neg p$ are entertained as equivalent possibilities. The speaker has no evidence to judge one more plausible than the other.

(7) Nonveridical equilibrium (Giannakidou 2013)
A partitioned (p and $\neg p$) space M(i) is in nonveridical equilibrium if there is no evidential bias.

Equilibrium : true uncertainty

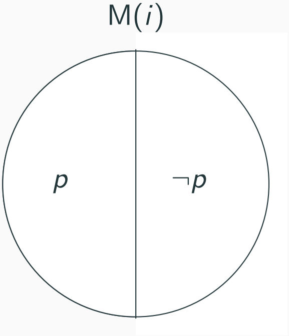


Figure 2: Non veridical equilibrium

Evidential bias

(8) John must be home

State of uncertainty, but the speaker has some evidence that creaetes bias towards *p*. (Giannakidou 2013, Giannakidou and Mari 2015, 2018b, 2021a, 2024).

The formation of bias relies on the evaluation of evidence.

MUST / FUT associates with an epistemic modal base M(i).

(9) M(i) $(t_u)(w_0) = \lambda w'(w'$ is compatible with what is known by the speaker i in w_0 at t_u)

MUST / FUT associates with an epistemic modal base M(i).

(9) M(i) $(t_u)(w_0) = \lambda w'(w'$ is compatible with what is known by the speaker i in w_0 at t_u)

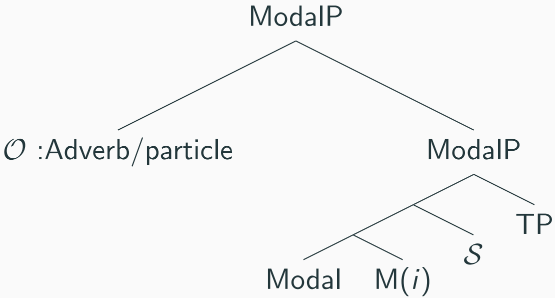
(10) $\text{Ideal}_{\mathcal{S}}(\text{M}(i)(t_u)(w_0)) =$
$\{w' \in \text{M}(i)(t_u)(w_0) : \forall q \in \mathcal{S}(w' \in q)\}$

So defined, $\text{Ideal}_{\mathcal{S}}$ delivers the worlds in the modal base in which all the propositions in $\mathcal{S}$ are true. $\mathcal{S}$ is a set of propositions that correspond to common ground norms/personal convictions etc. (more later on this). What matters here is the structure of the judgment.

Metaevaluation

A metaevaluator is a ranking function $\mathcal{O}$. A modal adverb is typically the realization of $\mathcal{O}$.

(11)



ModalP

$\mathcal{O}$:Adverb/particle

ModalP

Modal

M(i)

$\mathcal{S}$

TP

Giannakidou and Mari 2018b

Bias

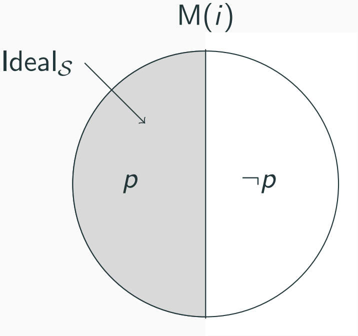


Figure 3: Biased modality

Scale of veridical committment

The veridical judgment is scalar :

(12) Scale of veridical commitment (Giannakidou and Mari 2016, 2021) :
$< p$, 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*.

Plan

Mechanical logos ?

The structure of human judgment of veridicality

Commitment, evidence, and informativity

Pause : how about LLM

Non cooperative conversations

Veridical judgment and evidence

It is referentiality to the world that establishes veridical commitment : assessment of evidence

(13) Epistemic commitment and evidence (Giannakidou and Mari 2021b) :
$<p$(good quality evidence, reliable for knowledge), MUST p (partial evidence, some gaps), MIGHT p (low quality evidence)$>$

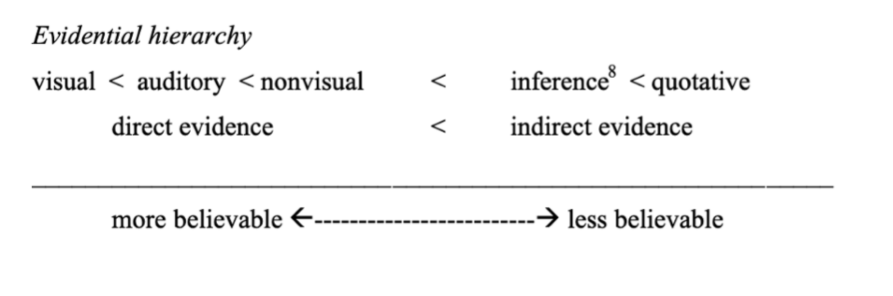
The linguistic agent makes a judgment prior to uttering a sentence depending on the evidence they have ; the LLM, we will argue, cannot make that judgment

(14) Veridical judgment and informativity : (Giannakidou and Mari 2016,2021b) :
$p \gg$ MUST $p \gg$ MIGHT $p >$;
where "$\gg$" means "informationally stronger than"

Bare assertions p (speaker knows p, p added to the common ground) $\gg$
MUST p (speaker does not know p, but is evidentially biased toward p) $\gg$
POSSIBLY p (speaker does not know p, and there is nonveridical equilibrium)

What types of evidence ?

The classical picture, de Haan 1992 :



Evidential hierarchy

visual < auditory < nonvisual < inference[8] < quotative

direct evidence < indirect evidence

more believable ←------------------------→ less believable

More on the evidence types

(15) Je sais que Marie est enceinte.
I know.PRES.1sg that Mary is.IND pregnant
I know that Mary is pregnant.

(16) Je crois que Marie est
I believe/Think.PRES.1sg that Mary is.SUBJ
enceinte.
pregnant.
I believe that Mary is pregnant.

More on the evidence types

(17) So che Maria è incinta.
Know.PRES.1sg that Mary is.IND.3sg pregnant.
I know that Mary is pregnant.

(18) Credo/Penso che Maria sia
Believe/Think.PRES.1sg that Mary is.SUBJ.3sg
incinta.
pregnant.
I believe that Mary might be pregnant.

More on the evidence types

(19)

Credo/Penso	che	Maria	sia	incinta.
Believe/Think.PRES.1sg	that	Mary	is.SUBJ.3sg	pregnant.

I believe that Mary might be pregnant.

(20)

Credo/Penso	che	Maria	è	incinta.
Believe/Think.PRES.1sg	that	Mary	is.IND.3sg	pregnant.

I believe that Mary is pregnant.

More on the evidence types

Doxastic attitudes of certainity can also take subjunctive

(21) Sono sicura che Maria sia/è incinta.
Am certain.PRES.1sg that Mary be.SUBJ/IND.3sg pregnant.
I am certain that Mary is pregnant.

(22) Sono convinta che Maria sia/è incinta.
Am convinced.PRES.1sg that Mary be.SUBJ/IND.3sg pregnant.
I am convinced that Mary is pregnant.

Belief as credence vs. conjecture

Solipsistic belief : credence
(Mari 2016, Giannakidou and Mari 2021a,b)

- The DOX(i) is homogeneous
- Assessing evidence for belief is a form of evaluation using mixed evidence factual, or endogenous subjective preferences
- Internal state : factual and endogenous evidence forms an internal state of belief

The solipsistic belief judgment can be rational and factual, but it can also be irrational and rely on endogenous preferences. The indicative belief is treated by grammar as subjective "knowledge".

Credence

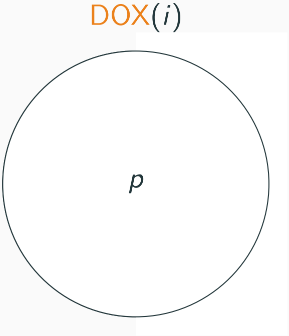


Figure 4: Credence

Credence vs. conjecture

Conjectural belief, supposition : believe but not know
(Mari 2016, Giannakidou and Mari 2021a,b)

- The M(i) is not homogeneous, the speaker now is in a complex state of having a belief but being aware that she lacks knowledge.
- Endogenous + exogenous evidence
- Internal state : factual and endogenous evidence forms an internal state of belief

We will argue : The LLM lacks the ability to form either type of belief, since it cannot form veridicality judgement

Credence vs. conjecture

Conjcetural belief, supposition :

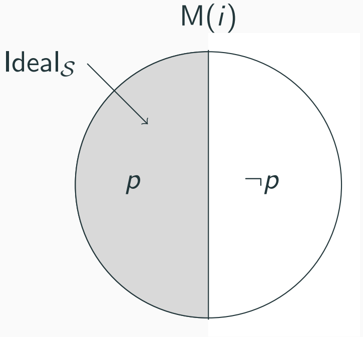


Figure 5: Biased modality : suppositional belief

Suppositional belief is the equivalent of MUST in the realm of attitudes.

Plan

Mechanical logos ?

The structure of human judgment of veridicality

Commitment, evidence, and informativity

Pause : how about LLM

Non cooperative conversations

The basis for the formation of veridicality judgment

- Veridicality judgments reveal
  1. Knowledge (factual, objectively veridical)
  2. Doxa (subjectively veridical)
  3. and mixture of those (doxa but lack of knowledge, suppositional belief)
- Veridicality judgments rely on
  1. External evidence (direction : from world to inner state)
  2. subjective preferences (direction : from inner states to world)
  3. A mixture of those

The formation of the veridicality judgment requires the ability to form internal states of belief and knowledge

ChatGPT lacks the judgment engine

- The formation of knowledge or belief has two components : (a) a procedural one of assessment of external + internal evidence, and (b) an internal one of forming a mental representation (belief or knowledge state).
- LLMs cannot relate to the world therefore
- LLMs cannot form internal representations.
- Hence, LLMs cannot form veridicality judgments.

Since they lack the ability to know or believe, it follows that the LLMs lack the ability to form attitudes in general : hence no desires, no emotive or more subjective attitudes (tastes, etc).

Plan

Mechanical logos ?

The structure of human judgment of veridicality

Commitment, evidence, and informativity

Pause : how about LLM

Non cooperative conversations

Misalignment between evidence and veridicality judgment

- Assessing evidence, as we said is a form of evaluation using evidence and subjective assumptions
- Boscaro, Giannakidou and Mari (2024) show that, in social media, assertions are mostly grounded in reported evidence.
- They challenge the views according to which assertion is weak (Greenberg and Wolf 2018 ; Krifka 2024).

The assertion / reported evidence correlation is a default in X.

Misalignment between evidence and veridicality judgment

Corpus of French 19,595 tweets (Kozlowski et al. 2020 ; Bourgon et al. 2022) Ecological crises occurred in France from 2016 to 2022 and posted 24h before, during (48h) and up to 72h after the crisis.

Among those tweets, we have randomly selected 3137 that have been doubly annotated for speech acts and for evidentiality.

Speech act categories

Speech act categories	**Definition**
Assertions	New raw piece of information.
Subjective	Expressions of opinions, beliefs, preferences and evaluations. (partly overlaps with exclamatives)
Jusstives	Order, wishes, leading to action.
Interrogatives	Information seeking questions.

Figure 6: Speecg Act Categories from Laurenti et al. 2022

Evidential categories on X

Evidential categories	Definition
Direct Source	The information is firsthand and has been acquired through vision, hearing, or other senses. It results from a direct contact between the speaker and the phenomenon described. Pictures that are not relayed are considered to belong to this category.
Relayed Source	The information has a reported source and the latter is third-hand (or second hand). The source is explicitly stated in the tweet content (hyperlinks, relayed pictures, mentions referring to a third party source)
Loose Sources	The information is reported from a third party source that is difficult to identify because it is not explicitly stated in the tweet content. The content of the tweet is however relevant and related to the annotated crisis.
Lack of Testimony	The tweet relayed do not have any marking of information source or the tweets' content is not related to the crisis.

Table 1: Evidential categories on Twitter

Assertive / relayed correlation

Evidentiality	Assertive	Subjective	Interrogative	Jussive	Total
Direct	123 (3.92%)	75 (2.39%)	6 (0.19%)	17 (0.54%)	**221 (7.04 %)**
Relayed	**1442** (45.97%)	161 (5.13%)	33 (1.05%)	**326** (10.39%)	**1962 (62.64%)**
Loose Sources	150 (4.78%)	**217** (6.92%)	26 (0.83%)	22 (0.70%)	**415 (13.23%)**
No Testimony	177 (5.64%)	**235** (7.49%)	31 (0.99%)	96 (3.06%)	**539 (17.18%)**
Total	**1892 (60.31%)**	**688 (21.93%)**	**96 (3.06%)**	**461 (14.70%)**	**3137 (100%)**

Table 2: Evidentiality vs Speech Acts

Assertive / relayed correlation

Evidentiality	Assertive	Subjective	Interrogative	Jussive
Direct	-10.29	26.53	-0.76	-15.48
Relayed	**258.67**	**-269.30**	-27.04	37.67
Loose Sources	**-100.30**	**125.98**	13.30	-38.99
No Testimony	-148.08	116.79	14.51	16.79

Table 3: Evidentiality vs Speech Acts : Relative difference to independence

Trust

What matters is trust, the speakers accept to assert depending on whether they trust the source.

(23) a. The speaker *A* relies on a source **rp** in context **c**. **rp** is the source of content C, and is either an hyperlink, a mention @ or a #sourcename.
b. Let *C* and *B* be the components of the informational basis of *A*. *C* is the exogenous evidence (the content provided by the source **rp**). B is the set of *subjective preferences* of *A*. The source is trustworthy iff $\mu(\pi \mid \mathrm{B}\cap\ \mathrm{C}) = 1$.

LLM and trust

At the very best, LLMs use C, but lack the basis for trust formation, which is the evaluation of C based on preferences.

Conclusions

1. The LLM cannot cannot form knowledge, belief, or subjective states because it lacks the ability to form inner states. The LLM is therefore not truly co-operative in discourse.
2. The LLM cannot deal with non-cooperative conversations because it lacks trust : the ability to evaluate the input.
3. The formation of trust is an evaluation relying on exogenous C and endogenous B evidence— and ultimately leads to action and decision making, it is the essence of logos.
4. Because of the above, the LLM lacks logos.

Thank you !