*Interdisciplinary Description of Complex Systems* 24(3), 208-221, 2026

# USE THEORY OF MEANING AND GENERATIVE AI: CAN CHATBOTS BE RULE FOLLOWERS?

**Tomislav Janović[1, *] and Barbara Babič[2]**

[1]University of Zagreb, Faculty of Croatian Studies
Zagreb, Croatia

[2]c/o University of Zagreb, Faculty of Croatian Studies
Zagreb, Croatia

DOI: 10.7906/indecs.24.3.1
Regular article

*Received*: 22 February 2026.
*Accepted*: 28 March 2026.

## ABSTRACT

Given the unprecedent advancements in the field of simulation of human linguistic behavior by generative AI models, one might expect – notwithstanding the obvious limitations of such models – that this development will be theoretically advantageous to the deflationist and the reductive theories of meaning. To test this hypothesis, we first explicate what we take to be the Use Theory of Meaning. Roughly, it is neither the language users' representational states (or the neural correlates thereof) nor the reference relation to extramental reality ("fixation of reference"), but the regular use of linguistic expressions in various contexts that accounts for the phenomenon of meaning. We first narrow down this rather vague claim that has its roots in in Wittgenstein's Philosophical Investigations. We do that (1) by explicating two crucial notions from Sellars' early conception of language games, and (2) by showing how these notions are complemented and further developed by Horwich in his non-normative, deflationist and reductive account of meaning. We then examine the potential of generative AI models to generate, in a conversational manner, longer portions of text which not only conform to the syntactic rules of natural language, but which also seem to satisfy, and to a surprising degree, its semantic and pragmatic constraints. We consider the similarities and differences between an artificial and a human language user, with special regard to the issues of rule following and normativity.

## KEY WORDS

use theory of meaning, generative AI models, use-property, rule following, normativity

## CLASSIFICATION

JEL: H83

*Corresponding author, $\eta$: tomislav.janovic@gmail.com; +385 1 245 7660;
University Campus Borongaj, Borongajska cesta 83d, HR – 10 000 Zagreb, Croatia

## INTRODUCTION

The puzzle of meaning, either of linguistic expressions or of its mental counterparts, is one of the central topics of contemporary theoretical philosophy. It is the puzzle that has motivated the work of the very founders of the analytic movement – Frege, Russell and Wittgenstein – as well as their successors and critics in the second half of the 20th century – from Quine, Sellars and Strawson to Davidson and Kripke. Currently, the debate seems to have settled around two interdependent questions: 1) the question of the semantic content of words, sentences and linguistic expressions in general, i.e. what exactly this content consists in and 2) the question of its foundation, i.e. what exactly are the facts, properties or processes that determine and justify our meaning attributions [1]. Although individual contributions to the debate do not always reflect this division of philosophical labor, it is often useful to keep the two groups of issues and theories – the semantic and the metasemantic ones – separate. As our own contribution is concerned, we see ourselves as saying something about the foundational issue – what is it that makes linguistic expressions meaningful. More specifically, the question we want to examine is whether there is anything about the recent developments in language modeling AI that give us reason to abandon the traditional, intuitive view – generally known as *representational theory of meaning* – and endorse the neo-Wittgensteinian notion that what makes our linguistic tokens meaningful are their communally established and context sensitive use-properties.

What exactly use-properties are and how they determine the semantic contents of words and sentences, depends on the type of Use Theory of Meaning (UTM) one adheres to. There are different versions of this theory – different attempts at specifying the meaning-conferring properties of language use[1]. The earliest version can be traced to Sellars' work on language games which was underway even before the appearance of Wittgenstein's *Philosophical Investigations* [2]. Of the later versions, the two most elaborate ones are Brandon's and Horwich's, both from the early 1990s. For reasons we will explain in due course, we take Horwich's version, complemented by some insights from Sellars, as representative of the UTM. The Ordinary Language School initiated by Austin and further developed by Grice, Strawson, Searle and others, although also inspired by Wittgenstein's seminal concept, we do not consider as a UTM in the sense we are interested in. Its focus is on how words and other linguistic expression work in everyday life and not on the sources of their meanings.

Our point of departure is the following hypothesis: Given the unprecedent advancements in the field of simulation of human linguistic behavior by generative AI models, notwithstanding the obvious limitations of such models one might expect that this development will be theoretically advantageous to the deflationist and reductive theories of meaning. To test this hypothesis, we first explicate what we take to be the core claims of such a theory – a UTM in its most explicit and most elaborated form. The key formulations are found in the works of Sellars and Horwich who, despite belonging to two very different generations of philosophers, both try to elaborate on Wittgenstein's seminal concept of language games. We then examine the potential of Large Language Models (LLMs) to generate, in a conversational manner, longer portions of text which not only conform to the syntactic rules of natural language, but which also seem to satisfy, and to a surprising degree, its semantic and pragmatic constraints. We consider the similarities and differences between an artificial and a human language user, with special regard to the issues of rule following and normativity. Finally, we draw some conclusions regarding our initial hypothesis and the conditions under which one would be willing to attribute linguistic understanding to an artificial language user.

## USE THEORY OF MEANING

As with many other philosophical theories, the UTM is far from being an easily identifiable set of claims, much less a consistent one. There are important differences between conceptions of authors that are typically cited as its main proponents. Nevertheless, there are several ideas over which many authors would agree, the most important one being captured by the Wittgensteinian slogan "meaning is use" [2; §43]. Roughly, it is neither the language users' representational states (or the neural correlates thereof) nor the reference relation to extramental reality ("fixation of reference"), but the *regular use* of linguistic expressions through various contexts that accounts for the phenomenon of meaning. In this section, we narrow down this – as it stands – rather vague claim by 1) explicating two crucial notions from Sellars' early conception of language games, and 2) showing how these notions are complemented and further developed by Horwich in his non-normative, deflationist and reductive account of meaning.

### SELLARS' INFERENTIALIST ACCOUNT OF LANGUAGE USE

There are several original ideas by which Wilfrid Sellars has contributed to contemporary philosophy of language generally and to the UTM specifically. Regarding our main purpose, two notions are crucial. The first derives from the distinction between the rule-conforming and the rule-obeying behavior, while the second has to do with the role of inferences in learning and "playing" various "language games" – the view generally known as "inferentialism".

In several of his works, most explicitly in "Some Reflections on Language Games" [4]. Sellars makes the point – a philosophical commonplace in the meanwhile – that despite appearing as a social and norm-governed activity *par excellence*, many cases of our use of linguistic devices do not look as instances of rule obeying behavior, i.e., as *intentionally* "fulfilling the demands of an envisaged system of rules" [4; p.32]. Rather, they seem much more as instances of merely *rule-conform* behavior, as if they "just happen to contribute to the realization of a complex pattern" [4; p.32] – in the manner of a bee dance, for instance. Instead of treating the problem as an either-or-issue, Sellars approaches it in a much more subtle way. Firstly, he anticipates a potentially fatal objection to his functionalist and pragmatist view: mastering the rules of language logically presupposes the knowledge of another language – a metalanguage in which these rules would be represented. Secondly, to avoid this problem, i.e., to show how rule-obeying behavior might come about without the deployment of a meta-language, he introduces an intermediary concept – "pattern governed behavior". It is a type of behavior that becomes entrenched by selective reinforcement: those community members who have already mastered the language rules shape the behavior of those who have not by making the latter "respond to a pattern of one kind by forming another pattern related to it" [4; p.34]. These transitions from one pattern to another are determined by "transformation rules", like the rules of inference of the type 'p, q; therefore p and q' or 'if p than q, p; therefore q' etc. Such rules can be explicitly represented in the cognitive system of the experienced (self-conscious) language user. Within that system, they have the status of "rules of criticism" or "ought-to-bes" – they function *as* norms, i.e., as pieces of knowledge about one's social obligations (as participants of a language game). However, in that, explicit form the rules typically *cannot* be conveyed to another player/speaker, especially not to a one whose cognitive system does not have the capacity to represent them as such – as explicit obligations or "ought to bes". This is why the language rules itself are typically conveyed by conditioning – language learners are trained, by selective reinforcement, to apply them as "rules of action" or "ought to dos". In other words, language users begin to conform to these rules not because they recognize their normative force, but simply because they have mastered a type of pattern governed behavior – in specific, the inferential rules like the ones mentioned above. In Sellars's technical jargon, the rules are

learned and applied, at least initially, not as explicit rules of criticism but as implicit rules of action ("ought to dos").

What exactly are these rules? In case of language games, the rules to be mastered and applied are the ones determining both the language syntax and its meanings. Regarding the latter, it is crucial to appreciate Sellars' insistence on the importance of what he calls "material" inference in pattern governed behavior (of which natural language is the most sophisticated kind). Unlike "formal" inferences, which reflect the syntactic (grammatical) structure of language, "material" inferences are exemplified by transitions from linguistic patterns like 'Whales are mammals' to patterns like 'Whales do not lay eggs' or from 'It has been raining' to 'The streets are wet'. Such inferences are critical for the language learner's ability to fully participate in a language game. Namely, it is the very disposition for drawing relevant inferences in respective contexts that makes language users successful and their utterances meaningful. Simplified, the meaning of language items is constituted by their context sensitive and pattern governed use.

Obviously, inferential rules, as *inter*linguistic transitions or verbal responses to others' utterances – are not the only type of transitions in a language game. As the proponents of the traditional theories of meaning emphasize, language items must somehow be linked to the world – to their extralinguistic referents. In Sellar's meta-jargon, this takes place either by "language-entry" or by "language-exit" transitions. Our linguistic reaction to a corresponding visual sensation (stimulus) – e.g., uttering 'cable car' when seeing one – is an example of the former, i.e., a transition from perception of a thing, type of thing or a situation to a linguistic expression (sentence). Conversely, words that we utter typically elicit rule governed reactions by our interlocutors – for instance, moving from the tracks upon hearing 'Beware, cable car!'. Such actions are governed by language-exit rules. Notwithstanding their importance, these two types of transitions, in Sellars' conception, are not "moves" *in* a language game; they are rather positions "involving" such a move on the one side and an extralinguistic entity (e.g., state of affairs, proposition etc.) on the other. In his own words,

> [t]o occupy a position in a language is to think, judge, assert that so-and-so; to make a move in a language is to infer from so-and-so, that so-and-so. And although sensations do have status in the English language game, their role in bringing about the occupation of an observation sentence position is not that of a thought serving as a premise in an inference [4; p.36].

This is why language-entry and language-exit transitions are not so much in the focus of the UTM, especially not in the focus of its inferentialist version inaugurated by Sellars. What proponents of this kind of theory are actually interested in are the interlinguistic transitions – their lawfulness, their social origins and their normative import. The author who has dealed with these issues in considerable length and made inferentialism a philosophically prominent view is Robert Brandom. However, Brandom's [5] version of inferentialism is not appealing to our project for several reasons, the most important being that in his conception the language rules are *intrinsically* normative – they tell us how we ought to use language items and how this determines their communal meanings. What we are interested in is how this very normativity came about and how its origins are related to the phenomenon or meaning. Horwich's version of UTM, to which we now turn, addresses these latter issues.

## HORWICH'S NON-NORMATIVE VERSION OF UTM

Since its initial formulation, Horwich's theory has been gradually refined against the background of criticisms that it has been or could be exposed to. At its core lies an apparently simple idea derived from several influential remarks by late Wittgenstein and taken up by

Sellars – that the way language items are publicly used in respective contexts determines the meaning of their constituents. In his three books, published after the first formulation of the theory, Horwich elaborates on this idea [6-8]. His general approach is a radical one – he treats the legacy of the post-Wittgensteinian philosophy of language, mind and logic much more as a burden than as a starting point for a successful theory of meaning. To remove this burden, he tries to show that most of the traditional constraints imposed on the notion of meaning are unjustified; that they are products of either misunderstanding or of uncritical acceptance of certain presuppositions about how language works – how its basic elements are related to the elements of thought on the one, and the elements of the world on the other side [6; pp.2-3, 13, 14]. Once these confusions and unjustified assumptions are cleared out of the way, it should, as he hopes, become obvious that a reductive and a deflationist account provides a more plausible answer to the foundational question of meaning than its competitors.

Horwich's catalogue of "pseudo-constraints on an adequate account of meaning" [6; ch.1], comprises a host of issues or points of disagreement with the received (representationalist) view. With these issues in mind, he gives "a fairly complete list of objections that have been made against use theories of meaning" [6; p.54] by some of the most prominent philosophers of language and mind of the 20$^{th}$ century. In his later works, he updates this list by taking a stand towards (then) newer developments in the field and their implications for the UTM. It would, of course, be inappropriate (not only for spatial reasons) to give even an abridged overview of these objections and rejoinders – of how Horwich's version of the UTM purports to explain, or rather explain away, various aspects of truth, reference, representation, intentionality, intensionality, compositionality, apriority, normativity, holism, internalism, vagueness and other issues born out of the long-standing discussion on meaning. In what follows, we therefore address only those aspects of his theory that we deem relevant for the assessment of our initial hypothesis, assuming that its other claims are defensible, or at least not obviously refutable. These core claims are: (1) the acceptance claim, (2) the non-normativity claim, (3) the inferentialism claim, and (4) the modest holism claim.

We start with Horwich's central definition that purports to satisfy the aforementioned constraints, or (as he would have it) apparent constraints, on a plausible account of meaning:

> The meaning of a word, w, is engendered by the non-semantic feature of w that explains w's overall deployment. And this will be an acceptance-property of the following form: 'that such-and such w-sentences are regularly accepted in such-and-such circumstances' is the idealized law governing w's use is (by the relevant 'experts', given certain meanings attached to various other words) [7; p.28].

What this definition states is that a word w acquires its meaning from the fact that the sentences in which it occurs (w-sentences) tend to be accepted – held true in a law-like manner – with respect to the circumstances of their utterance. The law governing the use of w is "idealized" in the sense that, like other statistical laws (e.g., the ideal gas laws in physics), it purports to express the cumulative effect of myriads of micro-level events which have contributed to the current use of w. As such, it is a *ceteris paribus* law: it is "subject to variety of factors that cause deviations from the behavior that would 'ideally' occur" [8; p.119]. This law – the "rule" the language users are "implicitly following" – cannot be spelled out: with each public occurrence of w within a w-sentence its meaning changes and solidifies, dependent on other words in the sentence and the specific conditions of *their* utterance and acceptance. Theoretically, all previous uses of w "feed into" its present use and thus contribute to the meaning as use ($meaning_u$) of w in the individual speaker's idiolect. If it were possible to precisely spell out the implicit rule ("ideal law") governing this use, we would have a perfect explanation of not only why w applies to things that it does apply (with occasional errors), but

also why we think that it *should* apply to these things. It would capture both the non-semantic feature constituting w's meaning and its (apparent) normative import.

In the late 1990s and the early 2000s, this view has raised many eyebrows and incited many critiques. Before we turn to some of the most controversial issues (from our standpoint), an important disclaimer is in order. The cited definition applies primarily to "literal, semantic meaning of an expression type", i.e., to "that which is expressed independently of the speaker's intentions, beliefs or context, and is known by anyone who understands the language" [6; p.3]. That is, the non-literal types of meaning and akin linguistic phenomena – a metaphoric meaning, propositional meaning, implicature, irony etc. – are *not* covered by the definition. However, Horwich assumes that they are *derivable* from an adequate account of literal or basic meaning. And the explanation of solidification of the non-literal $\text{meaning}_u$ is the same as the one applying to literal $\text{meaning}_u$ – regularity of use. In other words, these sophisticated types of meaning present only an apparent threat for his version of UTM. For, once the theory is properly worked out, it should, Horwich hopes, be able to cover more subtle aspects of language and remove the threat. With that important constraint in mind, we now turn to other clarifications relevant for our main purpose of this article.

A rather obvious difficulty that threatens to compromise Horwich's project is the very notion of *acceptance* which plays the key role in the aforementioned definition. Namely, if accepting a sentence containing w reduces to *holding it true*, it is hard to see, at least *prima facie*, how this can be squared with the reductive, non-semantic characterization of meaning advertised by Horwich. For, obviously, being a semantic notion *truth* should not feature in a (purportedly) non-semantic definition. To neutralize this objection, Horwich: 1) endorses a deflationary theory of truth[2] and 2) tries to show how "acceptance", as a relation between a sentence and a mental state, "can be explicated in non-semantic terms", that is how it can be re-described "in purely physical, behavioral and psychological terms" [6; p.95]. Leaving out the details of his explication, the upshot is that if a language user S reasons and behaves *as if* he/she holds a given sentence true, this can be taken as his/her acceptance of its propositional content. In Horwich's own words, "what S accepts may be inferred, given principles of inference and decision theory, on the basis of what he utters and what he does" [3] [6; p.96].

Accordingly, the "right" use of a word by a language user does not require his/her *understanding* of its meaning – at least not in the form of a specific and occurrent mental state or process. What is required is an ability to use the word in a way approximating its communally established use – a solution similar to the one proposed by Sellars (as laid out in the previous section). So, if what constitutes its use in the user's idiolect sufficiently overlaps with what constitutes its use in the communal language, this fact alone is sufficient to make w meaningful, irrespective of the alleged meaning-giving or reference-fixing role of mental states. Thus, no "right" inner state (belief, feeling, meaning intention, trace in mental lexicon etc.) – or, for that matter, *any* inner state or process accompanying an utterance – seems to be necessary to turn a string of sounds (or graphic marks etc.) into a meaningful token of a language type. For Horwich, the familiar feeling "that there is something we know when we understand a word" [6; pp.16-17] does not ground its meaning – at least not independently of the way it is used, provided that the linguistic and extralinguistic context of this use is suitably considered. In fact, it is the other way around: the more an individual use of a word converges to its communally established "basic use", the more one's "understanding" of it – one's belief or impression (if any) that the sentences in which it occurs are true – is appropriate. Accordingly, since the communal use of a word is grounded in its individual use-property $u(x)$, it is the latter, rather than any posit from the repertoire of the traditional metasemantic theories, which is "explanatory basic". For instance,

> … the fact that the expert and hence communal deployment of "dog" is the result of the word's having use property '*u*(*x*)' constitutes the fact that it means what it does – i.e. that it means DOG – in the communal language. If a community member's deployment of the word results from the same property '*u*(*x*)', then the meaning of "dog" in his idiolect will be the same as its meaning in the communal language. He will then qualify as knowing *implicitly* what the word means, and thereby as understanding it [6; p.17].

That is, the implicit knowledge of a word's meaning – its "understanding" – depends on its convergence to the statistically prevalent use. At least in typical cases, its individual use will not be governed by any process that can be characterized as *explicit* rule following (in the sense of Wittgenstein, Sellars, Kripke and many other authors engaged in solving, or merely stating, the so-called paradox of meaning). What individual use is determined by is a communally established rule of use – a type of pattern governed behavior in Sellars' sense (as explained above). Whether this amounts to *implicit rule following* – this depends (as we will see) on how one understands this notion[4].

But how does the use of a language type in a language community – a specific acceptance practice – *normatively* affect its individual uses? How do individual speakers *get to follow* the "idealized" law of use when producing a token of a language type? Namely, there is an apparent tension between the purely descriptive characterization of meaning and its evaluative import – the obvious fact that language speakers (as manifested in their reactions) distinguish "correct" from "incorrect" uses of words and sentences. Horwich takes great pains to resolve this tension by claiming that meaning – in the sense of meaning$_u$ – is intrinsically *non*-normative despite having normative implications [6; pp.37-39, 7; pp.12-14, 8; pp.135-139]. It is the communal or lawlike character of meaning that accounts for its apparent evaluative import – for the intuition that we *ought to* use w in accordance with its regular, communally established use. Or to put things the other way around, the regular, communally established use of w is the source of the apparently normative feature of its individual uses. In Horwich's words [6; p.38], it is fully "consistent with the *non*-normative nature of meaning that predicates should be applied only to things of which they are true".

To substantiate this claim Horwich draws a parallel with other "fact-to-value principles" – the ones well-known from ethics and aesthetics – and concludes that in neither of these cases is one committed to the view that the "antecedent circumstances" determining the correctness of such principles are intrinsically normative. That is, one can characterize these circumstances in purely *descriptive* terms and still maintain that people feel they "ought", or see themselves as "obliged", to behave in certain ways and not in others. Obviously, those language users that conform to the principles will (*ceteris paribus*) be better off than those who do not. But that does not make the principles intrinsically normative. In the case of language, we (tacitly) know that we will be more successful in achieving our communicative goals if we elicit certain beliefs in our interlocutors – "in so far as having certain beliefs is correlated with assenting to certain sentences (in virtue of their meanings)" [6; p.38].

As we understand Horwich's version of UTM, there is no reason to treat 'assenting to' (or 'accepting of') an individual use of a language item as an exclusively psychological notion[5]. This concept can be construed as a *behavioral disposition* to appropriately react to such a use. Of course, there can be many – perhaps too many – possible reactions. And in individual cases it might be difficult to distinguish the genuine acceptance of a sentence token from an apparent one, or even from its non-acceptance. For instance, it might be hard to distinguish someone's *intentional* non-compliance with a verbal request to pass the salt from non-acceptance of the sentence expressing that request – due to a supposed deviation of the individual use of the verbal

phrase ('Please, pass me the salt') from its communal use. But this does not seem to endanger the general plausibility of the deflationist, non-semantic view of meaning as advanced by Horwich.

One's behavioral reactions to one's interlocutor's use of words, like in the just cited example, is but one type of lawlike transitions constitutive of our use of language. In Sellars' terminology (as explicated in the previous section), such reactions are norm-conform in the sense that they are determined by "language-exit rules" – rules describing behavioral reactions to linguistic prompts. (For instance, "If one's interlocutor, under such-and-such circumstances, utters 'Please, pass me the salt', one passes the salt.") On the other side, there are the "language-entry rules" determining the reverse type of reactions – how a language user verbally responds to extralinguistic goings on. Horwich's example (taken from Sellars) is our lawlike tendency to utter "red" (either in public or in "inner" speech) when perceiving a surface which reflects light of the respective wavelength. This example, together with other color words, approaches an ideal case of conformity to the communal language-entry rules. Namely, there is no reason (*ceteris paribus*) to expect individual divergences in verbal reactions to such (unambiguous) environmental cues. Less ideal examples are species names, numerical predicates, evaluative words, mental terms etc. For, according to Horwich [7; p.20], "as we move from one such type to another there is likely to be considerable divergence of structure". Nevertheless, the overall ("ideal") lawfulness of use of all such terms in a language community will be preserved – the sentences in which tokens of the respective expression types occur will be accepted, under respective circumstances, to a fairly high degree. And this holds regardless of possible individual deviations from the basic regularities or occasional disagreements between laypeople and experts over their "correct" use.

Finally, the third type of rules – the ones especially relevant for the assessment of our initial hypothesis (due to the nature of AI agents under consideration) – are the ones governing a language user's inferences from what has been said to what is implied (and can be further said). These are the interlinguistic rules of which the most general ones determine the use of logical connectives ('or', 'and', 'entails' etc.), reasoning schemes (e.g., *modus ponens*), and various epistemic norms (including the use of the predicate 'true'). What Sellars calls "material inferences" (see in previous text) belongs to a less basic albeit not less important type of inference rules. Given the respective environmental conditions, these reasoning patterns establish how we infer one type of semantic contents from another, i.e., how, after accepting some basic sentence in which a word occurs, we come to accept other, less basic but epistemically related sentences in which the same word occurs.

The inference rules, complemented by contextual and other factors, are thus the most general and the most plausible model for explaining the overall use of a word w by a speaker – it is $\text{meaning}_u$. Apart from the environmental context in which a sentence token S containing w is used (e.g., the presence of a red surface when uttering 'red'), the most significant contextual factor is the role played by the co-occurring words in S. For, obviously, the law-like acceptance of S containing w can provide S with a definite meaning only relative to the law-like acceptance of other words occurring in S. Yet, if the latter, in so far as they are meaningful, also have their acceptance conditions, does not that lead to a kind of infinite regress or a vicious type of holism? Not according to Horwich. In his conception, such an unwanted consequence can be avoided by assuming that there is "a limited stock of interrelated basic meanings on which all other asymmetrically depend" [7; p.22]. By showing what these basic meanings (or "core uses") are and how they influence the meanings of other terms [7; pp.167-173], he sees himself in a good position to conclude that "meaning interdependence (aka 'holism') *per se* does not prevent each meaning-property from being constituted in a distinctive non-semantic way" [7; p.54]. The only provision, of course, is that each co-occurring word is used in line with its communally established use (as reflected in their acceptance conditions).

Now it is time to see how this theory, that we took as representative of a reductive, non-normative approach to meanings, applies to the linguistic (or quasi-linguistic) behavior of an artificial language user – one whose existence could not have been envisaged at the time in which the normative version of UTM has been proposed.

## LARGE LANGUAGE MODEL CHATBOT AS A LANGUAGE USER

So, is there anything regarding the philosophical issue of meaning, especially in view of the way this issue has been treated by the UTM, that we can learn from contemporary LLM-type of chatbots? According to the UTM, as we have shown, the use of language within a language community is a pattern-governed social practice in Sellars' (and Wittgenstein's) sense; or, in Horwich's terms, it is an instance of implicit rule following. In case of natural language, the rule "followed" is an idealized, statistical, *ceteris paribus* law determining the use of a word across various linguistic and extralinguistic contexts (based on its "core use"). The former type of context is constituted by the co-occurring words in the sentence, the latter by the environmental conditions of its utterance. Obviously, and for principal reasons, this latter type of context cannot be represented, still less accommodated, by a purely language processing device like an LLM-type of chatbot.

Nevertheless, being generated by selective reinforcement and maintained in a law-like way, the quasi-linguistic behavior of a chatbot hardly requires a philosophical argument to be seen as pattern-governed. However, unlike the case of a human language learner who is nudged into conforming to the pattern by the linguistic community, in the case of an LLM-type of chatbot this role – the role of instituting the rule – is performed by a machine-learning algorithm (created and administered by a human language user). Equally obviously, the "rules" embedded in the algorithm and "followed" by the model are not explicit. In fact, it is doubtful whether the chatbot's behavior can be viewed even as implicit rule following. This (as we will shortly see) depends on how one interprets the processes of learning and applying rules. In any case – again in stark contrast to a human language learner who is gradually exposed to a limited, typically preselected set of words, phrases and sentences (some of which explicitly, e.g., ostensively, linked to extramental entities) – an LLM "learns" by being *un*selectively exposed to an extremely huge set of textual data from sources as various as possible. The only way these "raw" data are preadapted is by "tokenization" – by dividing the strings of symbols into smaller, syntactically defined chunks of text which, for our present purpose, can be treated as elementary semantic units (roughly corresponding to "words" in Horwich's definition)[6].

The goal of the "learning" process is to "teach" the model to "predict" the next word in a sentence based on its "experience" with previous sentences containing the same word. In the initial phase of this process, this is achieved by "unsupervised learning" – the model is exposed to data which are "unlabeled" in the sense that its inputs do not have predetermined outputs. Each initial input (the datum entered) is supplied with an appropriate value (weight) indicating the "relevance" of a given input for a given output. These values are initially set randomly and then updated through several training cycles. The greater the number of parameters (weights) used, the greater the precision of the output values. The goal of this training phase is to let the "hidden patterns" between data gradually emerge so that the data can be grouped according to the degree of their mutual connectedness ("similarity").

In the next phase, the model undergoes fine-tuning on more specific datasets. This is achieved by "supervised learning" – the model receives inputs paired with "correct" outputs, which improves its "contextual understanding" and thus its ability to generate expected responses. For models like ChatGPT, this includes reinforcement with human feedback. When generating responses, the model uses the "autoregressive process": it predicts one token at a time based on the previously generated tokens and the verbal context. The technical details aside, what is

crucial is that for each step the model calculates the probabilities of the succeeding tokens and among them picks the "correct" one. Principally, this "choice" is based on the probability calculation, although some additional, more "creative" criteria can also be applied[7].

As a result of this training process, seemingly meaningful verbal outputs – actually, strings of symbols – are produced as reactions to verbal prompts. These reactions not only manifestly "observe" the syntactic rules of language (in Sellars' terminology: formal inference rules), but also, and to a surprising degree, satisfy its semantic and pragmatic constraints – adhere to Sellars' rules of material inference which are typical of advanced language games. In other words, the model functions as a reasoning device, manifesting an ability to give and justify reasons in a human-like way. The most interesting point, however, is that this ensues without any syntactic or reasoning rules being *explicitly supplied* to the model.

To appreciate the importance of this point, it is useful to introduce a three-partite analysis of rule following behavior [14; pp.134-137]. The first type, which can be called *explicit* rule following, is when a system S (propositionally) knows that R is the right rule and applies it. The second type, which can be called *implicit*, is characterized by the fact that S actually *does* follow R but it does not (propositionally) know that it does. The third type, which can be called *apparent* rule following, applies to cases where S merely behaves *as if* it follows R. Obviously, there is no question about an LLM *not* being able to follow a rule in the first sense – in fact, according to the UTM, not even human language users typically conform to such a wide repertoire of rules (and type of rules) consciously and deliberately. The issue is rather whether an LLM can follow rules in the second, implicit sense. If not, its rule following behavior should be qualified as merely law-*like* – in accordance with a rule, without this rule being followed in a sense that would differ from the sense in which, say, the Earth "follows" the rule of gravity when orbiting around the Sun.

For Horwich, the necessary condition for a language using system following a rule in the second, implicit sense is that its relevant activity be "governed by an ideal law" (as explicated above). Moreover, for such rule following to take place, the rule does not have to be *formulated* in any way, nor *understood* by the user, nor result in user's *conscious decision* to behave in accordance with that formulation [7; p.50]. With these restrictions out of the way, LLM seems, at least *prima facie*, to satisfy the mentioned condition. Horwich's "ideal law" of use for a given word – its meaning$_u$ – can be interpreted as "built into" the model during the training phase, i.e., by application of the appropriate machine learning algorithm. In view of the sheer number and diversity of sentences (and contexts) the model is exposed to during training, this interpretation seems reasonable, even though it is hard, if at all possible, to specify the type of relation between the law and the learning algorithm. In any case, considering its conversational efficacy across various contexts, the model's use of word tokens reflects the communally established, i.e., statistically typical use. In fact, for most types of words, one could claim (paraphrasing Horwich) that the meaning of a given word-token in LLM's "idiolect" almost perfectly overlaps with its meaning in the communal language.

However, its behavioral efficacy notwithstanding, the crucial question is what makes LLM's law-like behavior more than merely rule-*conform*. If there is no plausible answer to that question, would not it be more appropriate to compare its behavior to a bee dance – which is also determined by a kind of algorithm, albeit a genetically pre-programmed one – than to human language user interacting with other users within a community? As we have seen, Sellars took a lot of pains to deal with this puzzle. In his solution, special emphasis is given to the specific way language rules are learned and applied in a community – to the fact that they can be transmitted as "ought-to-dos" and then gradually transformed into "ought-to-bes", i.e., full-blown norms.

Horwich, confronted with the same problem, proposed the "ideal law" solution. However, in his later works [7; pp.14-15, 7; pp.125-126, 8; pp.117-118], as if feeling that this condition was not sufficient, he considered another characteristic of language users which, in combination with their exposure to linguistic rules, would justify viewing human linguistic activity as implicit rule following. This characteristic is the ability for *self-correction* – of acting "against one's initial inclinations" in cases of deviations from the rule. Basically, this is the same process of conditioning assumed by Sellars in his explanation of the origin of linguistic norms. Notwithstanding Horwich's own doubts about self-correction being necessary for rule following [8; pp.133-134], it is interesting to note that LLM's potentials for meeting this condition are severely limited. LLM is not a robot – it has neither sensors for receiving signals from the outer world nor "effectors" and "externals" that would enable it to act in that world. Therefore, having no need for language-entry and language-exit rules, it is dependent solely on statistical patterns ("language-language transition rules") when "making decisions" about which rule to follow in order to arrive at the best possible output for a given input. Therefore, the only way it can "act against" its previous "decision" – revise its initially calculated output – is when it is explicitly prompted to do so[8]. Otherwise, lacking a meta-layer which would enable it to evaluate and eventually revise its own "acceptances" – approximations to the ideal law of use – in real time, the case for an LLM as a rule following system seems weaker that it would otherwise be.

There is another aspect of human linguistic behavior that might also present a problem for the analogy between a human language user (as an authentic rule-follower) and an LLM. It is our astonishing flexibility in dealing with ambiguities and non-literal forms of meaning – implicature, metaphor, metonymy, irony, sarcasm, propositional meaning etc. At least for now, these forms of meaning cannot be aptly simulated by an LLM. In fact, they seem to present a problem for Horwich's theory too, despite his optimism regarding the prospects for deriving non-literal forms of meaning from his general theory (which, as he readily admits, applies to "unambiguous word types"). Some authors would see this as a manifestation of another, more general and more obvious deficiency pertaining to all use theories of meaning. It is their inability to differentiate between a user that *understands* what he/she says and a one that merely manifests, "simulates", meaningful behavior [3; p.83]. We do not share that view, but we recognize its intuitive appeal.

## CONCLUSION

After stating our initial motivation for conducting this inquiry, in the second part of the paper we explicated what we see as central claims of the non-normative version of the UTM. Firstly, we presented Sellars' conception of language games by invoking the distinction between pattern-governed and pattern-conform behavior and by showing how, according to Sellars, the former can develop from the latter, i.e., how language rules arise, become entrenched and transmitted in a community. We also specified three types of these rules, especially emphasizing (with our initial hypothesis in mind) the rules of "material inference" as a subtype of "interlinguistic transitions" guiding user's verbal responses to other user's verbal acts.

Secondly, we indicated how these notions can be complemented and further developed by a radically non-normative, deflationist version of the UTM. As a representative of such a theory we have chosen the one proposed by Paul Horwich. The backbone of his reductive conception of meaning is the idea that there is a rule, an "ideal law", that determines how a word type is used by individual speakers across different sentences and different environmental contexts, whereby people in their linguistic behavior "implicitly" follow this rule. Their implicit rule following is manifested in their law-like "acceptance" of sentences in which a given word occurs. This fact, the assumed "use-property" of a word, *is* the very *non*-semantic feature that

constitutes its meaning. We have attempted to clarify the most important tenets of this theory, starting from the very *acceptance condition* – how it is to be interpreted, how it can be reconciled with the normative characterization of meanings and how it applies to various types of words and expressions, throughout different context and for performing different tasks, with a special emphasis on tasks requiring the application of inference rules.

With this version of the UTM in mind, we have drawn a parallel between an ordinary language user and an LLM-type of chatbot with the purpose of testing our initial hypothesis – that the reductive, non-normative approach to meanings might gain certain support from this parallel. Avoiding technical details (as much as this was possible) we have examined whether Horwich's notion of implicit rule following and Sellar's notion of pattern-governed behavior can be applied to an artificial language user. As our discussion revealed, this cannot be done without important caveats. This is why the language-like behavior of an LLM *cannot* simply be taken as an argument in favor of the non-normative version of UTM. Nor can it be taken as an argument against alternative, representationalist theories of meaning. In fact, proponents of the latter theories could invoke the relevant differences between an LLM and a human language user – lack of understanding, spurious ability of self-correction, inability to successfully cope with the non-literal forms of meaning – to argue, by way of a *reductio*, against the UTM.

However, we would like to end our paper by relativizing this conclusion. The sheer possibility of law-like linguistic behavior as exhibited by the extant generation of LLMs makes one wonder about characteristics like understanding, meaning-intentions and normative attitude being irreducible features of language use and indispensable for the generation of meanings. If an AI system can capture the typical, communally established patterns of language use and apply them to various tasks, including those requiring sophisticated reasoning, and with an astonishing sensitivity to the interlinguistic contextual cues, is it that much inconceivable that the use-properties corresponding to these patterns actually *are* the (non-semantic) facts constituting their meanings? It is our feeling that this inconceivability will diminish with the rising prospects of an advanced type of artificial agents – e.g., a robotic version of an LLM supplied with sensors and peripherals – which will also exhibit certain degree of sensitivity to *extra*linguistic (environmental) features of language use. And with these prospects rising, the representationalist will be more pressed to specify the conditions under which she would be willing to attribute linguistic understanding to an artificial language user. On the other side, a proponent of a UTM would have to answer the question about the degree to which her acceptance of this theory is forcing her to attribute such an understanding to an LLM[9].

## REMARKS

[1]For overview, see chapter 6 in Lycan's Philosophy of Language [3; pp.79-89].

[2]Horwich devotes much time and space to this – for him crucial – aspect of his deflationist theory of meaning. Basically, it is a simple idea that truth of a sentence or proposition is not a kind of substantial relational property – as assumed by correspondence, verificationist or pragmatic theories of truth – but a simple fact that we tend to accept certain sentences or propositions on the grounds of the general principle: 's means that p → (s is true ↔ p)'. Thus, " 'w is true of x' does not stand for a substantive (i.e. potentially-reducible) relation. Rather, we understand it – *relative to a prior understanding of meaning-attributions such as 'w means DOG'* – through our acceptance of conditionals such as:

w means DOG → w is true of all and only dogs

w means CAR → w is true of all and only cars … and so on" [8; p.104].

[3]Here is a formulation from his later work: ... a public sentence is accepted by a person if his disposition to utter it is correlated with his being in a mental/neural state that grounds one of his premises in theoretical and practical reasoning" [8; p.31].

[4]Regarding the history of the debate on meaning as rule following, starting with Kripke's seminal work [9], it is tempting to view this kind of pattern governed behavior as a paradigmatic case of implicit rule following. For Horwich, however, "although the terminology of 'rule following' may be fairly natural here, it is not compulsory. ... [I]t seems plausible to suppose that there is no determinate fact as to whether so-called 'implicit rule following' (and hence meaning) is really a matter of rule following" [7; p.50].

[5]Some critics have understood acceptance exactly thus – as an exclusively "psychological notion rather than a recognizible form of actual social behavior" [3; p.88]. Although many Horwich's formulations point in that direction, we think that the interpretation of this key theoretical notion can be accommodated to a "behaviorist" reading – in the sense that a language user accepts a sentence if he/she behaves as if he/she is in the corresponding mental state or brain/hardware state

[6]The technical details about the functioning of LLMs were supplied from several sources [10-12].

[7]Actually, it calculates internal relations between tokens in real time by applying the so-called "transformer's attention mechanism": it first "analyzes" the user's input to "determine the context" – by identifying keywords, syntactic structures, and semantic relationships between them. During a single conversation session, previous items are also included in the thus supplied "context".

[8]A more apt way to interpret the absence of a self-repair function in LLMs would perhaps be to say that the model has fulfilled this task "in advance", i.e., in the training phase during which it has been exposed to machine-learning algorithm – the equivalence of the conditioning process (as described by Sellars) to which a language learner is exposed to adopt the communally established norms. The crucial difference is, of course, that the language learner by adopting a norm also learns to react to cases of its violation, i.e., adopts the self-repair function, while the LLM acquires a kind of (obviously imperfect!) "immunity" against rule violation.

[9]We are especially indebted to one of the reviewers of our article for this last point.

## ACKNOWLEDGMENTS

The research that resulted in this article was conducted within the framework of two scientific projects, both hosted by the Institute of Philosophy in Zagreb: "Intentionality and Modes of Existence" (IP-2022-10-5915) funded by the Croatian Science Foundation and "Antipsychologistic Conceptions of Logic and their Reception in Croatian Philosophy" (APsiH 2024) reviewed by the Croatian Ministry of Science and Education and financed through the National Recovery and Resilience Plan of the European Union.

Barbara Babič has done her part of the research when preparing her B.Sc. Thesis "The Problem of Rule Following in Thought and Language" at the Faculty of Croatian Studies of the University of Zagreb, supervised by Tomislav Janović.

We are also thankful to the reviewers of the manuscript version of this article for their comments and suggestions.

## REFERENCES

[1] Speaks, J.: *Theories of Meaning*.
In: Zalta, E.N. and Nodelman U., eds.: *The Stanford Encyclopedia of Philosophy*. 2024, https://plato.stanford.edu/archives/win2025/entries/meaning,

[2] Wittgenstein, L.: *Philosophical Investigations*.
Translated by: Ansvombe, G.E.M. MacMillan, New York, 1953,

[3] Lycan, W.: *Philosophy of Language: A Contemporary Introduction*. 3rd edition.
Routledge, New York & London, 2019,
http://dx.doi.org/10.4324/9781315146119,

[4] Sellars, W.: *Some Reflections on Language Games*.
Philosophy of Science **21**(3): 204-228, 1954,
cited from: Scharp, K. and Brandom, R., eds.: *In the Space of Reasons: Selected Essays from Wilfrid Sellars*. Harvard University Press, Cambridge & London, pp.28-56, 2007,

[5] Brandom, R.: *Making It Explicit*.
Harvard University Press, Cambridge, 1994,

[6] Horwich, P.: *Meaning*.
Clarendon Press, Oxford, 1998,
http://dx.doi.org/10.1093/019823824X.001.0001,

[7] Horwich, P.: *Reflections on Meaning*.
Clarendon Press, Oxford, 2005,
http://dx.doi.org/10.1093/019925124X.001.0001,

[8] Horwich, P.: *Truth–Meaning–Reality*.
Clarendon Press, Oxford, 2010,

[9] Kripke, S.: *Wittgenstein on Rules and Private Language*.
Blackwell, Oxford, 1982,

[10] ByteByteGo: *How ChatGPT Works Technically: ChatGPT Architecture*.
https://youtu.be/bSvTVREwSNw?si=vm77b8wGBA1Yr6Kb,

[11] Wolfram, S. *What Is ChatGPT Doing ... and Why Does It Work?*
Wolfram Media, Champaign, 2023,
http://dx.doi.org/10.31855/bc47ee6b-75c,

[12] Liu, Y., et al.: *Understanding LLMs: A Comprehensive Overview from Training to Inference*.
Neurocomputing **620**(2), No. 129190, 2025,
http://dx.doi.org/10.1016/j.neucom.2024.129190,

[13] Devitt, M. and Sterelny K.: *Language and Reality: Introduction to the Philosophy of Language*.
Blackwell, Oxford, 1982.