

# LLMs, Turing Tests and Chinese Rooms: the prospects for meaning in Large Language Models

## Content

Discussions of artificial intelligence and computational systems in philosophy of mind (and elsewhere) have been shaped to a large extent by two brilliant and highly influential thought-experiments: Alan Turing's Imitation Test for thinking systems and John Searle's Chinese Room Argument (these two thought experiments are universally well-known, but just for sake of completeness I'll sketch them very briefly in §1). As is extremely well-known, Turing argued for a behavioural test for intelligence. Searle, on the other hand, argued that passing a purely behavioural test, focused at the level of inputs and outputs, could never guarantee that meaning, understanding, or thinking was occurring in a system, since a purely formal, syntactic capacity could underpin the behaviour. In many ways, recent debates about large language models (LLMs) are struggling to move beyond the divide marked by these original, opposing thought-experiments. Thanks to the open access status of some large language models (systems like ChatGPT and Bard), we all now have, available on any suitable laptop or phone, an extraordinarily powerful example of a system which (at least within certain parameters) is capable of passing a Turing Test. Yet those who feel swayed by Searle's Chinese Room are likely to remain unconvinced that the incredible behavioural prowess displayed by these systems is sufficient to show that LLMs deserve the attribution of any richer predicates.

One way to move this debate forward, stressed in several places already (xxx), is to 'look under the hood' of AI systems, asking exactly what (if anything) the elements they manipulate represent and what features are relevant to the computations they perform. While this is certainly an important strategy, and I'll touch on it below, in this paper I want to pursue a different route, exploring the features those sceptical about LLMs might take to ground meaning. In §1 I set out some common ground between Believers and Sceptics, concerning how LLMs work and the kind of features a Sceptic is likely to demand from an LLM before she would be willing to treat it as a semantic engine. In §§2-3, I explore the first potential ground of meaning, which requires a robust relation between linguistic signs and external objects or states of affairs. In section 3 I argue that concerns about worldly connections can be met, so that the outputs of LLMs should be viewed as genuinely meaningful. Then in (§4) I turn to the idea that LLMs require original intentionality before we admit them to the space of meaning. I suggest that this demand is not a prerequisite for meaning per se, but rather for agency or conscious understanding. I'll suggest

that the demands for original intentionality are not met by LLMs, but that nor should we want them to be.

## 1 Turing Tests, Chinese Rooms and LLMs

Ideas generated in philosophy of mind usually remain within the confines of the ivory tower where they were born, but this is certainly not true of Turing Tests and possibly not true of Chinese Rooms. Still, just to ensure everything is on the table, in this section I'll recap very briefly on the two thought-experiments which have done so much to shape understanding in this field.

Alan Turing proposed what he called 'the imitation game' in his 1950 paper 'Computing machinery and intelligence' as an operationalised method to replace the question "Can machines think?". Turing's starting point was the recognition that terms like "thinking" and "intelligence" are difficult to define and are prone to yield divergent interpretations in different contexts. Instead, then, he proposed replacing questions framed in these vague terms with an investigation into a much more tractable question: "Can a machine pass the imitation test?". In the imitation test, a human subject poses questions to an unknown target, via text, and receives answers in the same way. If, given a reasonable length of conversation, the human subject is unable to tell whether they are interacting with another human or with a machine, then, if they are in fact interacting with a machine, that machine is judged to pass the imitation test. While Turing doesn't in fact say very much about the repercussions of passing such a test, the suggestion seems to be that, for all intents and purposes, a successful machine would qualify as a thinking thing. The only other option, Turing seems to think, would be adopting some kind of solipsistic condition on thinking, requiring first-personal conscious awareness, so that we judge thinking to be going on only when we 'know it from the inside'. However, as Turing 1950: ?? points out, this solipsistic perspective is not the one we adopt in everyday life, instead he notes that "it is usual to have the polite convention that everyone thinks". A similar courtesy, it seems, should be extended to behaviourally adept machines.

In Searle's famous thought experiment from 1980, however, behavioural success of the kind envisaged by Turing is held to be *insufficient* for attribution of rich predicates like thinking or meaning to a system. For Searle objects that success at the level of outputs could come from a system which did nothing more than manipulate symbols in terms of their formal, syntactic properties, with no consideration of semantic content, or meaning, occurring within the system.

To see this, he asks us to consider a sealed room with a human subject in it. The subject receives what appear to her to be meaningless marks via an input slot. She then uses a giant look-up table to pair the incoming marks with a further set of marks and she outputs this further set of apparently meaningless marks via an output slot. Unbeknownst to the subject, however, the incoming symbols are questions in Chinese and the series of marks which she outputs constitute answers to these questions (again, in Chinese).<sup>1</sup> Searle asks us to consider whether the individual in the room knows or understands Chinese and the intuitive answer, despite her impressive performance with Chinese questions, is that she does not. Yet, Searle contends, as for the individual in the Chinese Room, so for the computer. Both systems are purely syntactic engines with no semantic properties in play.

So, what are we to say of LLMs? All parties agree that Large Language Models are trained for the task of minimizing prediction errors when deciding on the most probable next word in a sequence. Crudely, faced with the sentence “Emily paid in her cheques at the \_\_”, the task of an LLM is to fill in the blank with the most probable next word, e.g. completing this sentence with the expression “bank”.<sup>2</sup> In order to carry out this task, LLMs utilise what is known as *transformer architecture* (Vaswani et al 2017), which allows expressions of a natural language to be mapped into vectors within a multidimensional vector space. During pre-training, the system is provided with an almost unimaginably huge dataset (hundreds of billions to trillions of token words) which it uses to learn the statistical relationships between linguistic items in the database. The system does this by repeatedly masking words in sentences within its database and ‘making guesses’ (initially entirely randomly) about what the masked word is. Once the correct word is revealed the system then recalibrates the vectors it assigns to words in light of this information. The vector assigned to each word is thus determined by the statistical correlations between expressions across the entire training data. Following this unsupervised learning, there is then a period of human feedback which helps the system fine-tune its guesses for new input strings and then recalibrate resulting vector weights. Finally, vectors are also moderated by the correlations in the input strings themselves, so that, e.g., the vector assigned to a token of “dog” will differ dependent on whether the input string to the LLM is “The big dog...” or “The barking dog...”, etc. (see Jurafsky and Martin 2023 for discussion).

## Footnote

<sup>1</sup> Bender et al 2021 provide an up-dated thought-experiment, involving an octopus intercepting an underwater communication cable, which is analogous to Searle’s.

<sup>2</sup> Since the task is next-word prediction, some Sceptics have argued that LLMs are nothing more than auto-correct systems on steroids; see Marcus xxxx.

While the details get murky quite quickly (with part of the problem being that even computer scientists are unsure about exactly how to understand the internal states of LLMs), it seems that neither the input nor the design of LLMs *directly target* semantic content. The properties that LLMs are trained to recognise are simply statistical occurrence properties. Furthermore, the training data they receive is raw and unencoded (i.e. it has not been labelled in advance by humans using categories like “noun” or “verb”, nor does it contain axioms that could plausibly be taken as assigning meaning to basic symbols, as in classical computational systems). The processing behind LLM outputs thus seems to bypass semantic properties. Yet the surprising result of this kind of processing is that it results in behaviour which gives the *very strong impression* of linguistic understanding. Try interacting with ChatGPT and its fellow travellers and it is hard, sometimes almost impossible, to resist the belief that one is conversing with another person, a thinking thing (the infamous case of Google engineer Blake Lemoine being a stark case in point).<sup>3</sup> As [Sparks p. 8] suggest of GPT-4:

[Its] primary strength is its unparalleled mastery of natural language. It can not only generate fluent and coherent text, but also understand and manipulate it in various ways, such as summarizing, translating, or answering an extremely broad set of questions. Moreover, by translating we mean not only between different natural languages but also translations in tone and style, as well as across domains such as medicine, law, accounting, computer programming, music, and more.

Furthermore, this exemplary performance means that the role LLM outputs play in our lives is increasingly akin to those of an ordinary conversational partner: we take the systems to convey information, to answer questions, and to advance our understanding, in ways that are often indistinguishable from our use of the testimony of other humans.

From the perspective of behavioural tests, then, it is very easy to believe that LLMs should be attributed rich, semantic properties: LLMs look like they respond to and manipulate genuine content, generating meaningful outputs through their understanding of language. In what follows, I'll label those who hold that LLMs deserve the attribution of rich predicates (like representing, meaning, and perhaps even understanding and thinking) 'Believers'. At the most extreme end of Believers are those who hold that by tracking and manipulating facts about the statistical occurrences of words these kinds of system develop “a form of general intelligence, indeed showing sparks of artificial general intelligence...demonstrated by [GPT-4's] core mental capabilities (such as reasoning, creativity, and deduction), its range of topics on which it has

---

Footnote

<sup>3</sup> “The Google engineer who thinks the company’s AI has come to life”, Washington Post, June 11<sup>th</sup>, 2022; <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

gained expertise (such as literature, medicine, and coding), and the variety of tasks it is able to perform (e.g., playing games, using tools, explaining itself).” [Sparks, p.91]

‘Sceptics’, on the other hand, insist that LLMs are *gaming* Turing’s test – passing the behavioural criteria but by the use of strategies that do not, in fact, warrant ascription of richer properties (as Titus 2024: 101174 claims LLMs “do not themselves function to represent or produce meaningful text”). For LLMs to gain admittance to the space of meaning, we need to be sure that they are sensitive to the meaning properties of the signs they manipulate, providing answers which are driven by a grasp of semantic content not merely through sensitivity to common word distributions. Yet, Sceptics point out, LLMs lack the sorts of features we might *prima facie* expect are required for the emergence of genuine sensitivity to meaning: first, they are not embedded in the non-linguistic environment (they can’t see, hear or touch things, and they can’t manipulate objects in the real or virtual world), and, second, they lack the kind of long-term goals and aims which could drive their interactions with the world and with us. Furthermore, their responses lack the sort of consistency and stability indicative of having a point of view or perspective. As [Sparks, p.60] notes:

The ability to explain one's own behavior is an important aspect of intelligence, as it allows for a system to communicate with humans and other agents. Self explanation is not only a form of communication, but also a form of reasoning, requiring a good theory of mind for both yourself (the explainer) and the listener. For GPT-4, this is complicated by the fact that it does not have a single or fixed “self” that persists across different executions (in contrast to humans). Rather, as a language model, GPT-4 simulates some process given the preceding input, and can produce vastly different outputs depending on the topic, details, and even formatting of the input.

Yet the ability to act in the world, in a consistent and principled way reflective of long-term goals and needs, seems central to our conception of agency and it is very tempting to think that agency is a prerequisite for getting meaning off the ground. All this leads to a suspicion that LLMs are nothing more than “stochastic parrots” (to borrow the now well-known phrase of computational linguist Emily Bender and colleagues, 2021) – that they are probability-based mechanisms which just parrot empty words back to us.

So, who is right here? Perhaps predictably, I’m going to argue that there is right on both sides: Believers are right to think that LLMs are special, that their outputs are meaningful, and perhaps even that the systems succeed in representing and manipulating genuinely semantic properties, even though they are trained on an entirely non-semantic task. Sceptics, however, are right to think that meaning in these systems is derivative – LLM outputs mean things because *we* mean things when we use language. For a system to have non-derivative meaning I suggest it

## Content

would have to qualify as an agent (having, in Searle's terms, original intentionality). LLMs don't have that, but neither, I conclude, should we want them to. For creating systems which met this demand would risk drawing artificial systems into the space of moral concern. To see this, in the next two sections I consider LLMs connection with the world, then in §4 I turn to the issue of original intentionality.

## Headline heading

### 2. Semantic content requires contact with things beyond the text

## Content

Perhaps the most obvious objection to the claim that LLMs deserve the attribution of rich predicates like meaning and representing is that such systems lack the kind of connections to the external world which seem necessary to ground semantic content. As Piantadosi and Hill 2022 note, there is a “prominent view...that models trained only on text cannot acquire realistic meanings because they lack reference, or connection to objects in the real world” (see, e.g, Bender and Koller 2020. Lake & Murphy 2023: 411, although they argue against externalist semantics, agree that nothing in the vector assigned to a word in an LLM would allow a user to label an object or describe a scene). The assumption here is that words get their meaning from their connection to things outside the sphere of language, so that, e.g. “dog” means what it does because it refers to dogs or picks out the property of *doghood* (Fodor xxxx). While sentences and utterances are meaningful because they make claims about the world, so that their meaning can be analysed in terms of their conditions of truth (Davidson, Partee). As David Lewis famously put it (1970: 18), it seems that: “semantics with no treatment of truth conditions is not semantics”. Yet, if all a system has available to it is intra-linguistic information, we might suspect that truth won't enter the picture, since the system is out of touch with the external states of affairs where (according to most theories) truth gets its footing.<sup>4</sup>

The worry that externally determined properties like truth don't get any traction within LLMs is reinforced by the well-known issue of LLM fabrication, or “hallucination”: [sparks 82]

[A] key limitation of LLMs [is] their tendency to generate errors without warning, including mathematical, programming, attribution, and higher-level conceptual errors. Such errors are often referred to as hallucinations per their tendency to appear as reasonable or aligned with truthful inferences. Hallucinations, such as erroneous references, content, and statements, may be intertwined with correct information, and presented in a persuasive and confident manner, making their identification difficult without close inspection and effortful fact-checking.<sup>5</sup>

## Footnote

<sup>4</sup> Of course, there are ‘internalist-friendly’ (see §2.1) theories of truth available, such as Coherence theories. However, I'll leave these to one side here.

<sup>5</sup> For a particularly egregious example of LLM fabrication, see Oremus...

## Content

The existence of such fabrications undermines the idea that users can rest epistemic weight on LLM outputs, since falsehoods emerge unbidden from exactly the same background processes as give rise to statements of fact. As ChatGPT's developers, OpenAI, note (pointed out by Titus 2024: ??), fixing this problem is far from straightforward:

ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there is currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.

The Sceptic's argument is thus that a system needs robust externalist connections prior to any attribution of rich semantic predicates (like representation and meaning) and that since LLMs lack such connections they cannot be treated as semantic engines.

In response, however, I think there are (at least) three kinds of response a Believer might make:

1. reject the claim that externalist connections are needed (or at least maintain that referential properties are only part of the story).
2. accept the claim that externalist connections are needed but argue that future iterations of LLMs will have the required connections.
3. accept the claim that externalist connections are needed, but argue that, as things stand, LLMs already realise appropriate word-world relations through their connection to humans.

I'll take these responses in order but focus on (3), since I think it provides the most promising response to the Sceptic.

## Headline

### 2.1 *Distributional, Inferential, and Conceptual Role Semantics*

## Content

A first move would be to reject the call for an externalist semantics in favour of some form of internalist approach. According to internalist semantic theories, words and sentences are made meaningful not through their connection to external objects and situations, but through their position in a linguistic or conceptual network.<sup>6</sup> On this model, reference and truth emerge only later, at the point where *already* meaningful expressions are used to talk about the world. The properties internalists appeal to include distributional properties, entailment relations and

## Footnote

<sup>6</sup> For a recent, sophisticated defence of an internalist approach, where linguistic meanings are understood as instructions for building concepts, see Pietroski xxxx



## Content

position within a conceptual network. The first idea – that the distributional properties of expressions (i.e. the statistical patterns of word occurrences) suffice for meaning – has a long history in philosophy and linguistics, with the linguist John Rupert Frith, writing in the 1950's, capturing the approach in the slogan: “You shall know a word by the company it keeps”. That is to say, ‘distributional semantics’ holds that the meaning of a word is fixed by the pattern of words it commonly co-occurs with, so that, e.g., “dog” means what it does because it commonly appears with words like “barks”, “bites”, “bones”, etc, but less often (or not at all) with words like “justice”, “flies”, “not”, etc. Distributional semantics is clearly a pre-cursor of the approach to natural language processing adopted by LLMs (and by corpus linguistics in general), thus an advocate of Distributional semantics is likely to simply deny the claim (made in §1) that LLMs do not directly target semantic properties. If meaning is given by distributional properties and LLMs model distribution, then LLMs just *are* modelling semantic content. One response to the Sceptic, then, would be to embrace Distributional semantics.

LLMs may also qualify as modelling semantic content on alternative internalist approaches. For instance, Inferential Role semantics holds that meaning is grounded in the entailment relations that exist between sentences: to master “cow” is (to be disposed) to infer from “x is a cow” to “x is an animal”, “x has four legs”, “x gives milk”, etc (Boghossian 1994). As noted in [Sparks, p?], it seems that LLMs are highly adept at capturing these kinds of inferential relations. Or again, Conceptual Role semantics couches the required internalist properties at the level of thought, focusing on the relationships between concepts which presumably underpin acceptable inferences at the level of language (Harman 1987). Given LLMs success with inferential and conceptual network tasks, it could be argued that information about these relations is already tacitly contained within the system. Piantadosi and Hill 2022 seem to embrace this kind of view. They suggest that “relationships between concepts are the essential, defining, aspects of meaning” and they go on to argue that:

The key question for LLMs is whether training to predict text could actually support discovery of conceptual roles. To us the question is empirical, and we believe has been answered in a promising, partial affirmative by studies showing success on tasks that require knowledge of relationships between concepts. Text provides such clues to conceptual role because human conceptual roles generated the text.<sup>7</sup>

Deciding whether LLMs count as semantic engines given either an Inferential or a Conceptual Role approach requires taking a stance on what it is for information to be

## Footnote

<sup>7</sup> We should note that Piantadosi and Hill in fact adopt what philosophers call a *disjunctive* account, holding that both narrow, internalist content and wide, externalist content is ~~needed~~ in a complete account (Segal 2000 provides a defence of this approach, for objections to disjunctivism, see Davies 1991).



represented in a system and this is a controversial question. An analogy might be helpful here: say I set up a system to identify and track mutations in the melanocortin 1 receptor, a gene located on chromosome 16. Let's suppose that these mutations are entirely (rather than merely largely) responsible for the phenotypic trait of red hair. So a system which tracks these genetic mutations will, in effect, track people with red hair. It might be argued that such a system does not actually represent red-headedness or carry information about red-heads, since it simply tracks a property that correlates with or causes this observed trait, without recruiting the higher-level property in any of its internal transformations. On the other hand, since having red hair is (in the current thought experiment) nothing over and above having these genetic mutations, an alternative stance would be that by tracking the mutations the system *de facto* ends up representing information about red-headedness. Analogously, a Believer will want to claim that, from an internalist point of view, tracking statistical distributions turns out not merely to track a *proxy* for semantic properties but in fact succeeds in representing those properties as such, since distributional properties tacitly contain information about inferential and conceptual roles.

For philosophers, this kind of debate will call to mind protracted discussions (between Dretske, Millikan, Papineau, Fodor and many others) about the possibility of a naturalised account of content. Contributors to this debate tend not to adopt an internalist approach to content, instead grounding meaning in the relationship between a sign and some worldly state of affairs, but still the question of how we settle exactly what distal property an internal state represents remains the same. (Just as Dretske xxxx worried about whether the internal state of a frog represents *flies* or *moving black dots*, so this debate about LLMs asks whether an internal state of an artificial system represents conceptual or inferential relations or only statistical facts.) Thus we might think that the same tools used in naturalised externalist frameworks will be available to the Believer here; for instance, spelling out representation in terms of asymmetric causal dependence (so that an internal state of an LLM could be held to represent inferential/conceptual relations not merely distributive properties since the distributional properties are caused by the inferential properties of a sign but not vice versa; see Fodor xxxx) or its proper function within a system (as in the kind of teleological approach advocated by Millikan xxxx). Of course, these questions remain vexed and I can't hope to settle them here. Yet I think it suffices for present purposes to note that if the demand for an externalist semantics is rejected, LLMs will at least be in the running as systems which directly target and represent semantic properties through targeting and representing distributional properties.

An alternative response to the Sceptic's demand for better integration with the non-textual environment would be to accept that LLMs fail on external aspects *at the moment*, but hold that this is a contingent gap which will be filled by future iterations of artificial systems. This kind of move is of course familiar from the 'Robot Reply' Searle considers to his Chinese Room argument: according to this response, what is needed to get meaning into a Chinese Room is the addition of supplementary systems that integrate the Room into its non-linguistic environment (Crane 1991: 127). For instance, a perceptual system could be added, allowing the person in the room to connect incoming symbols with external objects. Artificial limbs or prostheses could be added, so that the person in the room could manipulate their environment as part of their response to incoming linguistic prompts, perhaps together with some form of locomotion, so that the system as a whole could explore its environment. In short, then, the Chinese Room could be embedded within a robot.

This philosophical call for embedding a linguistic system in a multimodal structure has been answered by some engineers, as Lake & Murphy 2023: 412 note:

AI researchers are certainly working on various forms of multimodal learning. A recent flurry of work has focused on integrating vision and language, leading to creative combinations of computer vision and NLP models.

A recent multi-author paper (Butlin et al 2023) pursues this idea, looking in detail at three such wider systems: PaLM-E (Driess et al. 2023), described as an "embodied multimodal language model"; a "virtual rodent" trained by reinforcement learning (Merel et al. 2019); and AdA, a large Transformer-based, RL-trained "adaptive agent" (DeepMind Adaptive Agents Team 2023). While it is early days for this research, some of the results have been impressive, with systems able to recognise and label complex situations that go beyond the specific examples contained in the system's training data (see Lake & Murphy for further discussion). Furthermore, some of these models show "emergent visual-language alignment", where the decoder attends to the relevant part of an image when producing a word (e.g. analysing the umbrella part of an image when producing the word "umbrella" in a caption).

Although this is an extremely fast-moving area and the above projects are exciting, it is clear that the Robot Response cedes the current ground to the Sceptic: if we opt for an externalist view of semantic content and pair this with a requirement that an individual language

## Content

user must *themselves* realise the required word-world relations, then LLMs currently fall short.<sup>8</sup> A different move, however, would be to argue that there is no need to embed an LLM within an artificial multi-modal system, for they already have access to a proxy which allows them to forge the necessary connections with the world: although an LLM lacks a direct ability to interact with the non-linguistic world, it does connect via humans. As [Sparks p.53] contend:

language is a powerful interface, allowing GPT-4 to perform tasks that require understanding the environment, the task, the actions, and the feedback, and adapting accordingly. While it cannot actually see or perform actions, it can do so via a surrogate (e.g., a human).

This brings us to the third response that Believers might make to the Sceptic's demand for appropriate word-world relations, which appeals to the role that humans play in the lifespan of LLMs.

## Headline

### heading

### 2.3 *Derived Intentionality*

## Content

Computers are built and used by people, thus a common response to Searle's challenge has been to appeal to the meaning that programmers or users assign to the symbols manipulated by artificial systems: when a calculator outputs the symbol "4" in response to the input "2+2", this output means *four* because of the intentions of the person who designed or built the calculator and/or because of what the users of the calculator take the sign to mean. This is to appeal to what Searle termed 'derived intentionality': the symbols of an artificial system inherit meaning from their connection to human thinkers and their practices.

The appeal to derived intentionality takes on a special force, however, with respect to LLMs, for the signs manipulated by these systems just *are* the signs of ordinary natural languages that we, human speakers, manipulate in order to convey information to one another. These symbols are connected with human communicative systems both through the database of human-generated text that the system is exposed to during pre-training and through the interaction with speakers once the system is in use. Furthermore, these natural language signs are manipulated by LLMs in ways that look promising from the point of view of meaning. If we were faced with an artificial system which took English words as input but produced as output some kind of word salad – a rag-bag collection of English terms spewed out by a purely random mechanism, say – the idea that the signs produced in the output retained the meaning they would

## Footnote

<sup>8</sup> As Searle objects: "The first thing to notice about the robot reply is that it tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relation with the outside world".

have in the mouths of ordinary English speakers would be on much shakier ground. Yet the outputs of LLMs respect syntactic rules (even though at no point is the system given any explicit syntactic training or instruction) and they make contextual sense.<sup>9</sup> So, if we think the sentence “Dogs bark” means what it does when uttered by an English speaker because it combines words with a particular meaning in a particular well-formed syntactic structure, it seems a Believer might similarly hold that the sentence “Dogs bark” when produced by an LLM means what it does because it is the result of combining meaningful words in a particular well-formed syntactic string. LLM outputs are meaningful, according to this line of thought, because they stand on the shoulders of ordinary language users, re-using words and structures that have been antecedently made meaningful by us.

Searle himself, of course, thought that what mattered for understanding was, fundamentally, *original intentionality* and a Sceptic might object that the kind of derived aboutness discussed above is not really sufficient for meaning.<sup>10</sup> Rather what is needed is the ability to invest signs with meaning *de novo*, an ability which Searle holds arises in biological brains and things which have the same causal powers as human brains. I’ll return to the demand for original intentionality in §4, but in the next section I want to explore further the idea that derived intentionality might be sufficient to support genuine semantic attributions to LLMs (not least because the position of LLMs with regard to word-world connections may in fact mirror the position human interlocutors are often in).

### Headline heading

### 3. Derived intentionality as sufficient for meaning?

#### Footnote

<sup>9</sup> Piantadosi 2023 holds that the emergence of syntactic rules within LLMs serves to refute the Chomskian nativist model, according to which children must come equipped with innate knowledge of some fundamental syntactic principles in order to be able to acquire a language, given the limited linguistic training that infants receive (see Chomsky’s ‘poverty of stimulus’ argument). He writes that “The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics”. However, while Piantadosi is right that LLMs refute a very strong nativist argument (which holds that acquisition of syntactic rules would be impossible without innate knowledge), they don’t (at least currently, I think) challenge a somewhat weaker, but apparently more relevant, poverty of stimulus argument, which holds that the *limited input available to infants* means that they require innate knowledge. LLMs are built using astoundingly large amounts of data at the pre-training phase, amounts which no human (infant or otherwise) could comprehend. So the fact that LLMs are able to acquire sensitivity to grammatical rules from this vast data set without prior syntax-specific programming does not entail that infants, exposed to so much smaller a dataset, could acquire such sensitivity without innate knowledge (this argument is spelt out clearly in Milliere 202?). See also Landauer and Dumais 1997.

<sup>10</sup> Searle himself (Mind, Language and Society: 94) in fact maintains that derived intentionality gives the correct account of linguistic meaning, while original intentionality holds only at the level of thoughts and beliefs. Somewhat similarly, Fodor xxx holds that there is no such thing as semantics for natural languages, with meaning only really occurring at the level of the Language of Thought.

Intention-based semantic theories hold that meaning is grounded in the intentional practices of agents, so that, e.g., “dog” means what it does because English speakers intend “dog” to refer to dogs and an intentional practice has grown up of using this term in this way. There are, however, different ways to conceive of the relationship between meaning and intentions. Borg (20xx) introduces a distinction that might be helpful here, between what she terms ‘A-style’ and ‘B-style’ versions of intention-based semantics.<sup>11</sup> According to ‘A-style’ intention-based theories, speaker intentions play a preconditional role: in order for a sign to mean something there has to be a practice of using that sign with that meaning amongst intentional agents.<sup>12</sup> A-style theories are thus likely to appeal to generalised, conventional intentions (as in Schiffer’s 1972 rendition of the Gricean project). On this kind of A-style account what matters for an expression coming to have a given meaning in a given community is that the expression be used by one speaker with the intention of conveying a certain meaning and that this use be picked up by the community, so that it becomes conventional to intend to convey that meaning by using that word. A-style approaches provide an answer to a constitutive question concerning the kind of thing linguistic meaning is, grounding it in the intentional practices of language users, but they need not demand that producers of token expressions have particular intentions in order for signs to be meaningful.<sup>13</sup> According to A-style accounts, although intentional agents (with a practice of using certain signs to represent things) are needed to get a meaningful system off the ground in the first place, once the system is up and running, token signs from the system can be meaningful without any input from the intentions of the current speaker (under an A-style intention-based account, a speaker who says “dog” but intends to mean *cat* will fail, at least literally, to mean *cat* since the meaning of this sign is fixed by the conventional fact that speakers generally intend to convey the meaning *dog* by saying “dog”).

B-style accounts, on the other hand, hold that a linguistic utterance is simply a good piece of evidence for determining the intentions of a speaker and that it is only once someone has grasped these speaker intentions that they can be credited with a grasp of semantic content. So, to discover the meaning of some sentence “s” as produced by U we need to look to what U

## Footnote

<sup>11</sup> A somewhat similar distinction, between *meaning as a public object* and *meaning as a first-personal mental report* is drawn by Nowak and Michaelson 2022.

<sup>12</sup> An alternative move, in the same space here, would be to dispense with intentions entirely and pursue the kind of naturalised approach to content alluded to at the end of §2.1. So, for instance, following teleological accounts such as Millikan xxxx, we might try to define a linguistic practice without appeal to intentions (e.g. in terms of a sign’s *proper function* amongst a community of speakers). Naturalised accounts of content face well-known problems (the most pressing of which for LLMs is perhaps the presence of content in systems without an evolutionary history; see discussions of ‘Swampman’ xxxx), thus I’ll leave open the question of whether such accounts will ultimately be viable, opting instead to focus on a richer, derived intentionality model.

<sup>13</sup> Although some accounts do impose this additional requirement; I return to this point below.

intends to convey, what thought she is trying to get across by uttering what she does. B-style accounts take the locus of linguistic meaning to be utterances;<sup>14</sup> while A-style accounts focus on the type-level, taking semantic content to attach to sentence types. Prima facie, if semantic content is, first and foremost, a property of linguistic types then it seems that the door is open to token LLM outputs sharing in semantic properties in just the way that token human utterances do.<sup>15</sup>

A-style accounts rely on a notion of *semantic deference*, according to which a speaker can produce a given linguistic token with a specific meaning not because she cognises that meaning herself (though she may do) but because her use is part of an established practice and she defers to that practice. Semantic deference is most often discussed as part of the kind of ‘social externalism’ defended by Burge xxxx, where a speaker defers to experts in her community for the meaning of technical terms. In Putnam’s famous example (Meaning of Meaning, 1975), a speaker may be unable to distinguish elm trees from beech trees, and yet her utterance of the term “elm” still refers to elms (and not to beeches) because she defers to experts in the community who can tell these two kinds of tree apart. Social externalism allows a speaker to express a concept (and perhaps even be credited as possessing that concept) even if she has only a very partial grasp of what the concept involves. Hunter 2003: 734 frames this idea as follows:

(T) If the experts in S’s community use a word, W, to express concept C, and S is minimally competent with W and defers to the experts in its use, then S has C too.

The kind of deference envisaged here need not, it seems, be limited to technical terms and experts in particular fields. Instead, A-style accounts hold that it captures a much more general feature of linguistic meaning: words in general mean what they do because intentional agents kick-start a social activity, so that a token expression produced by a speaker now means what it does because there is a community practice of using that term with a certain meaning and the current speaker defers to this practice.

This kind of A-style, deference model is embedded in many approaches to semantic theorising which fall under the broad heading of ‘formal semantics’. So, for instance, take

#### Footnote

<sup>14</sup> It seems that B-style accounts fit well with approaches which stress the underdetermination of semantic content by words and structure alone (so-called ‘contextualist’ models; see Searle xxxx, Recanati xxxx). In these accounts, it is tempting to think that an appeal to current speaker intentions will be necessary prior to semantic evaluation, e.g. to work out that the speaker who says “The bridge is strong enough” means that it is strong enough *to withstand the storm, or to take the cars*, etc. Though see Borg xxxx for objections to the idea that speaker intentions can play this role.

<sup>15</sup> This would be to deny Searle’s claim about derived intentionality, that “formal symbol manipulations by themselves don’t have any intentionality; they are quite meaningless; they aren’t even symbol manipulations, since the symbols don’t symbolize anything. In the linguistic jargon, they have only a syntax but no semantics”.



‘minimal semantics’: according to minimalism, all well-formed, declarative sentences (relativized to a context of utterance) express truth-evaluable content which is fixed via the lexico-syntactic constituents of the sentence (plus their manner of composition) alone, where those constituents are, for the most part, not themselves context-sensitive.<sup>16</sup> So, for instance, minimalism claims that there is a truth evaluable content to be recovered for sentences like “All philosophy students study logic” and “Meera is ready” just on the basis of the words and structure of these sentences (i.e. the first sentence literally claims that every philosophy student, without qualification, studies logic, while the second literally states simply that Meera is ready, without indicting what she is ready for). Even though these contents are not the ones that ordinary speakers who utter these sentences standardly convey, nevertheless minimalists argue that we need to recognise this kind of literal, minimal, type-level content as truth-evaluable semantic content (e.g. to explain our practices around holding speakers responsible for what they say, as well as for drawing distinctions like that between lying and misleading; see Borg explanatory roles). Recovery of minimal content does not (at least in some versions) require access to speaker intentions; in a slogan from Borg 2004: ? grasping semantic content is a matter of “word-reading not mind-reading”.<sup>17</sup> So, could we assign sentence-level, minimal semantic content to the outputs of LLMs?

One potential objection to this suggestion is that LLMs engage in autoregressive processing, which imparts more context-sensitivity to the bearers of meaning than semantic minimalism allows. As noted in §1, in an LLM the vector assigned to a word in a current input string is moderated by the expressions that come before it (so that the vector assigned to “dogs” will differ if it is preceded by “all” or “angry”, etc). Yet this might seem to suggest the systems are treating word meanings as much more context-sensitive than the current appeal to type-level semantic properties allows. However, I think this objection is probably too quick. A primary aim of type-level, formal semantics accounts like minimalism is to preserve the compositionality of meaning for complex linguistic items, whereby the meaning of a sentence is fixed by the meaning of the words it contains together with their mode of combination. Yet it is possible that the contextual vector moderation LLMs undertake merely reflects the compositional process. That is to say, rather than reflecting a change to the meaning of a word at the token level, vector moderation could reflect the combinatorial process of integrating word meanings across an

## Footnote

<sup>16</sup> The relativisation of sentence-types to a context of utterance is to allow for genuinely context-sensitive terms, such as demonstratives and indexicals. According to Borg xxxx: ch.?, grasping the semantic content of sentences containing these expressions requires a hearer to introduce an appropriate singular concept the content for which is fixed by deferring to the intentions of the producer of the token.


<sup>17</sup> Cappelen and Lepore 2005: ?? advocate a version of minimalism that requires access to current speaker intentions prior to recovery of semantic content.


entire string. As Fodor and Lepore 1998: 111-2 note (in their discussion of treatments of polysemy and the purported need, which they deny, for a generative lexicon of the kind defended by Pustejovsky):

It's news if the *lexicons* of natural language are generative; lots of people (ourselves included) think they are more or less lists. But it's no news that VPs are generative, or that their semantics must somehow integrate the meaning of a verb with the meaning of its arguments.

The difficulty of deciding whether the kind of context-sensitivity introduced by autoregressive processing fits with or challenges a type-level approach to semantic content reflects a point already noted in §2, namely that there are currently no conclusive views on what the internal states of LLMs represent. Given this, however, I think it would be too soon to conclude that the mere fact of autoregressive processing shows that LLM outputs are not realising type-level semantic content.

A second challenge to the idea that LLMs can be treated as dealing in minimal content emerges from the fact that the outputs of LLMs are not always reflective of this type of content. That is to say, although in response to some prompts ChatGPT does deliver outputs which reflect a sensitivity to the strict, literal content of the prompts it receives, in others the outputs seem to suggest a sensitivity to richer, more contextually determined content. To see this, consider the following exchanges:

 small small  
if Jill is getting ready for a party and she tells Fred "Every bottle is in the fridge", is what she says false if there are bottles at the shop which are not in a fridge?

 small  
Yes, what Jill says would be considered false if there are bottles at the shop that are not in a fridge. Her statement "Every bottle is in the fridge" implies that all bottles, without exception, are located inside a refrigerator. If there are bottles at the shop that are not in a fridge, then her statement does not accurately represent the situation, and it is false in that context.

small  
EM You  
if jill drank 3 beers and drove home, which did she do first?

Content

small  
ChatGPT  
small  
The sentence "Jill drank 3 beers and drove home" does not specify the chronological order of events. It could be interpreted in two ways:  
small  
1. Jill drank three beers and then, at some point afterward, drove home.  
small  
2. Jill drank three beers while she was at the location where she eventually drove home from.  
small  
To determine the actual sequence of events, you would need additional context or information. Without that context, the sentence itself doesn't specify whether she drank the beers before driving home or not.  
📄 🍌 🗨️ ↻

In both these exchanges, ChatGPT responses pick up on the minimal content of the input sentence. In response to other prompts, however, LLM outputs seem to focus more on what we might think of as the contextually enriched content of the prompt.<sup>18</sup> For instance, tell ChatGPT that “Linda took out her key and opened the door” and the system’s response incorporates the bridging inference (that Linda used her keys to open the door). Yet this bridging inference is relevant at the token, not the type, level (that is to say, the sentence form “p and q” does not always invite this kind of bridging; see Carston 2002):

small  
EM jill took out her keys and opened the door. Was the door locked before jill opened it?

Yes, the sentence "Jill took out her keys and opened the door" implies that the door was locked before Jill opened it. Taking out keys and using them to open a door is typically done to unlock and then open the door. If the door was already unlocked, there would be no need for Jill to use her keys to open it.

In addition, sometimes ChatGPT delivers answers in line with content commonly thought of as richly pragmatically modulated content, i.e. picking up on Gricean implicatures, as in the following:<sup>19</sup>

footnote

<sup>18</sup> For discussion of this contextually enriched content (labelled ‘explicature content’ by Sperber and Wilson 1983), see XXXX.

<sup>19</sup> For a discussion of the differences between minimal content, explicatures and implicatures see Borg ???.

small

EM

my friend gemma asked bill if he wanted to have dinner with him and bill said he had to work. What did he mean by this?

Content



small

When Bill said he had to work in response to Gemma's invitation to have dinner, it typically means that he has a work commitment or obligation during the time Gemma suggested for dinner. This could mean that he has a job-related task, a meeting, or simply that he needs to be at work during the evening hours.



small

Bill's response suggests that he is unable to join Gemma for dinner because of his work responsibilities. It doesn't necessarily imply anything beyond that unless there is additional context or information that clarifies his intentions.

As ChatGPT's response notes, Bill's utterance in the imagined context typically suggests that he can't have dinner with Gemma. ChatGPT here reflects content that Bill would commonly be taken to imply by his utterance, but such implications have standardly been thought of as highly contextualised, attaching to token utterances in very specific contexts, not as part of the general, type-level meaning of sentences (Grice xxxx).

So, does the fact that the responses issued by ChatGPT in the above examples capture both minimal and pragmatically enriched content show that we should not treat LLM outputs as possessing minimal semantic content? I think there are two points to note about this. First, the claims are about differences in how ChatGPT is processing linguistic prompts (i.e as reflecting rich contextual information in the latter cases but focusing on a more minimal, type-level content in the former). However, this is not directly relevant for the current claim, which is that LLM outputs can be understood as expressing minimal content. That is to say, the claim is that the sentences LLMs produce can be understood as expressing their literal, type-level meaning, regardless of whether what they are reporting is the strict, literal content of the prompt or some more contextualised content. Secondly, the fact that LLMs seamlessly report on aspects of content that have traditionally been taken to be driven by context rather than by strict, literal meaning, is unsurprising. LLMs are massively holistic systems, where every answer they produce reflects the entire database they were exposed to during pre-training. Since the kind of contextually enriched contents reflected in the second two outputs from ChatGPT above reflect the way in which these kinds of sentences are commonly used by ordinary speakers, we should expect that what are commonly thought of as pragmatic inferences find their way into the system. It is because when keys and doors are mentioned, the most common relationship involves the keys being used to open the door that ChatGPT embodies the bridging inference in its response above. What this shows is that sensitivity to distributional properties across a large enough body of text will result in outputs which are sensitive to (at least some) pragmatic

inferences, just as sensitivity to distributional properties results in a sensitivity to minimal content. The fact that LLM outputs often reflect pragmatically enhanced versions of the input strings they receive speaks against any claim that the distributional properties LLMs map are fixed by the literal meaning of words and sentences alone, but this need not be part of the current claim (that LLM outputs can be understood in terms of the minimal content they express).

So, should we claim that LLM outputs are meaningful because they possess type-level content arrived at through a kind of semantic deference (so that the expressions produced by LLMs count as meaningful simply because they are tokens of types of symbols that belong to a meaningful communicative system, put together in grammatically respectable ways)?<sup>20</sup> I think the answer to this question is likely to depend on one's views on what is required for belonging to a linguistic practice, for some versions of A-style approaches re-import an appeal to current intentions at this juncture and any such move would place even this kind of derived content out of reach of an LLM. Take, for example, Kripke's theory of names (xxxx) which holds that the referent of a name is fixed by an initial act of 'baptism', where an agent who is in contact with an object introduces a linguistic label for it. A subsequent user of the name then refers to the object named by the original baptiser (even if the current user is not in contact with that object themselves) just in case the current speaker's use of the name belongs to a chain of past uses which can be traced back to the original baptism. Importantly, however, the current user needs to *intend* her utterance to belong to that chain of uses. Thus current speaker intentions do matter for fixing meaning for Kripke, even though they are not determinative of reference. If intentions are needed for semantic deference (in the way that the Kripkean model suggests) then type-level semantic content is not available for LLM outputs, since the system cannot intend its outputs to be part of a social practice.<sup>21</sup> On the other hand, if a current token expression can be taken as belonging to a chain of uses regardless of whether the speaker intends this or not, I would suggest that LLM outputs can be held to express genuine, type-level content.

Finally, however, two questions remain: should LLMs be said to grasp or understand the semantic content expressed by their outputs? And can an LLM be said to be asserting type-level

Footnote

<sup>20</sup> See Cappelen and Dever...

<sup>21</sup> This is why (I think) Searle holds that the symbols manipulated by the Chinese Room are meaningless, even though linguistic expressions possess ~~derived~~ rather than original intentionality (see *Mind, Language, Society*, 94). He claims that "that the mind imposes Intentionality on entities that are not intrinsically Intentional by intentionally conferring the conditions of satisfaction of the expressed psychological state upon the external physical entity" (*Intentionality*: 27). That is to say, even though words have, at best, ~~derived~~ intentionality, for them to possess even this aspect, for Searle, they need to be backed up by current intentional psychological states (rather than by membership in a social practice).

semantic content (do LLMs ‘mean what they say’)? In answer to the first question, I think whether we take large language models to grasp semantic content will depend on what we say about the internal states of the systems and the stance we take on what is required to represent a given property. The inner states of LLMs are generated by the distributional properties of words alone, but, as Lake and Murphy note, these distributional properties track what seem to be genuinely semantic properties like inferential role and conceptual connections (in the same way that, in the thought experiment suggested in §2 certain mutations realise red-headedness, or as a system which tracks a certain arrangements of atoms might be held to be representing the property of fragility). Similarly, an advocate of the derived intentionality model of LLM outputs might hold that the intentional practices of speakers are embedded in the distributional facts of the massive training database to which LLMs are exposed in such a way that internal manipulations which turn on these kinds of statistical properties also, de facto, turn on genuine meaning properties. In this way, it might be possible to argue that vector weightings can be understood as representing meanings not merely distributions. However, as the discussion of naturalised theories of content at the close of §2.1 alluded to, the question of what it is for an internal state to represent one distal property over another is vexed and I think that currently the jury must remain out on whether LLMs engage in the kind of genuinely content-driven internal transitions which would be needed to warrant an attribution of grasp of semantic properties to the system.

Finally, what about the illocutionary force of LLM outputs: do LLMs assert type-level semantic content? [cf. Cappelen and Dever ‘whole hog’] I think that the answer to this question is ‘no’. In line with many philosophical theories (such as Grice xxxx and Williamson xxxx), I think that truth has a key role to play in communication and assertion. But as already noted in §2, truth does not seem to get a grip in LLMs. This can be seen in the tendency of such systems to output plausible sounding falsehoods alongside truths and to go on to support any such claims by ‘fabricating’ evidence. This tendency reflects the fact that, as we might put it, the norm of LLM outputs is plausibility not truth. I think this means that, even though we should count LLM outputs as meaningful we should not count them as assertions. We should thus ensure that any of their outputs which make it into the public sphere are marked as such, indicating that a higher degree of vigilance would be needed before the semantic content is endorsed.

To summarise, then: the Sceptical worry that meaning requires an appropriate connection to the world can be met in three ways: by adopting an internalist approach to meaning, by embedding LLMs in robotic systems, or by adopting a type-level, deference model of semantic content. This latter approach, I suggest, in fact captures the way in which meaning



attaches to the vast majority of human linguistic performances. By producing well-formed strings of token signs which belong to an extant system of human communication, both the sentences we utter and the sentences LLMs produce get to have semantic content (assuming entry to the linguistic practice is not intention-dependent).<sup>22</sup> The main difference between us and the machines, then, as far as linguistic meaning is concerned, is that we unquestionably represent semantic properties *qua* semantic properties and we also (often) mean what we say.<sup>23</sup>

#### 4. Original intentionality

Searle (Mind language society 94) writes that “intrinsic Intentionality is observer-independent – I have a state of hunger regardless of what any observer thinks. Derived Intentionality is observer-dependent – it is only in relation to observers, users, and so on, that, for example, a sentence of French has the meaning it has”. He then holds that it is human brains alone which are capable of giving rise to this kind of original intentionality. This idea of an observer-independent, inner mental life where a subject takes one thing to stand for another is of course far from uncontroversial. Dennett, for instance, famously argues that *all* intentionality is observer-dependent (coming about through the adoption of the Intentional Stance), while Boden 1988 objects that the idea of original intentionality is ineliminably mysterious, relying on a dualist approach to the mind that should be rejected. This paper is obviously not the place to try and resolve such absolutely fundamental questions about the human mind, but let’s suppose that Searle is right here (at least about intrinsic intentionality, without the chauvinistic limitation to human brains), what would that mean for LLMs? First, to stand any chance of having original intentionality emerge in an artificial system it seems clear that it will need the kind of rich embedding in a non-linguistic environment envisaged in the Robot Reply (§2.2). It is too early to know exactly what LLMs will look like once successfully embedded in multi-modal systems, but at the very least we can conclude that *as things stand* LLMs are not candidates for intrinsic intentionality (or for consciousness). Secondly, intrinsic intentionality requires more than just contact with things beyond the text, it requires a subject of experience, so that worldly

<sup>22</sup> Of course, another way to ensure that entry to the practice does not require access to speaker intentions would be to provide a naturalised account of practices, as in Millikan-style teleological approaches which do away with speaker intentions entirely, focusing instead on the conventional communicative practices within a community of speakers (so that, e.g., “dog” refers to *dogs* because there is a practice amongst English speakers of using this expression to convey information about dogs and a practice amongst hearers of the term of taking it to convey this kind of information; see Millikan xxxx).

<sup>23</sup> Even if, in line with Fodor 20xx, we were to accept that ultimately talk of semantic content for a public language is derivative on the possession of semantic content by concepts and thoughts, I think it would turn out that talk of the meaning of LLM outputs is no less robust than the ascription of meaning to human utterances, for in both cases meaning at the level of language turns on derived intentionality.

## Content

interactions reflect the long-term goals and aims of a system. This in turn would result in a system which not only responds to the prompts it is given but which initiates (interesting) content-driven interaction on its own, displaying a kind of stability and consistency in its (linguistic and non-linguistic) actions that we might take to be reflective of having a point of view or character. Whether or not suitably extended LLMs could come to display this kind of behaviour is unclear, but I suggest that we should be wary in advance about whether it is behaviour we want to bring about. For a system which was able to interact with its environment (including its human environment) in this point-of-view driven way would surely be a system which was in the running for moral consideration. Rich embedding in an environment, the pursuit of long-term goals and the making of consistent judgements are not, I've suggested here, things which we should require of a system prior to treating its outputs as genuinely meaningful, though they are the kinds of capacities required prior to the attribution of properties like understanding and consciousness.

## Headline heading

### 5. Conclusion

## Content

First, I think that LLM outputs are deserving of attributions of semantic content since they comprise well-formed sentences of meaningful human languages. As such they express type-level content, through a process of semantic deference and derived intentionality. Second, whether we take large language models to grasp the semantic content their outputs express will depend on what we say about the representational properties of the internal states of the system and the properties that are relevant for how the system manipulates those internal states. If all that matters for internal transitions turns out to be the distributional properties of expressions, then I think we should refrain from saying that LLMs represent semantic properties or have any level of understanding of meaning (the surprising fact will then be that a system which lacks any kind of linguistic understanding can behave *as if* it has linguistic understanding). On the other hand, there may be (teleological or other) accounts of content under which the internal states of LLMs can be understood as representing not merely distributional properties but also the semantic properties which give rise to the distributional facts. If such an approach proves successful, then we might want to claim that LLM outputs are meaningful and that LLMs themselves treat them as meaningful.

Whether the internal states of an LLM can be understood as performing content-driven (rather than merely statistics-driven) computations remains extremely controversial (see Hattiangadi xxxx for a sceptical view) but whatever stance we take on this question, I think, thirdly, that we should deny that LLMs are in the business of asserting the content which attaches to the expressions they produce. LLMs do not aim at the truth in the way that would be required for them to count as genuine conversational partners or as asserting or making claims. Fourth and finally, Sceptics are right to think that LLMs lack original or intrinsic intentionality. They are not agents and they are not conscious. This doesn't, I've argued, stop them from meaning things but it may mean that it makes no sense to talk of them as *understanding* or (contra Turing) *thinking* (at least if thinking is held to involve original intentionality). Yet this is something we should be glad of, for the cost of the alternative would be a radical overhaul of our relationship with artificial systems.

**References**

Bender, E. M.; Gebru, T.; McMillan-Major, A. and Shmargaret, S. (2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Boghossian, Paul A. "Inferential Role Semantics and the Analytic/Synthetic Distinction." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 73, no. 2/3 (1994): 109–22. <http://www.jstor.org/stable/4320465>.

Harman, G. '(Non-Solipsistic) Conceptual Role Semantics' in Lepore (ed) *New Directions in Semantics* (1987).

Davies, M. 1991. Individualism and perceptual content. *Mind* 100: 461-484.

Piantadosi, S. 2023. Large language models refute Chomsky's approach to language. <https://lingbuzz.net/lingbuzz/007180> (accessed November 2023).

Segal, G. 2000. *A Slim Book about Narrow Content*. Cambridge, MA.: MIT Press.