Philosophy of AI
Vol. 1, 2025, 71-76
https://doi.org/10.18716/ojs/phai/2025.11963

WHAT DOES IT TAKE TO ESTABLISH REFERENCE IN LLMS? KRIPKE VS. AUSTIN

Steffen Koch[1]

[1] Bielefeld University, Philosophy, Bielefeld, DE

Abstract

Are the texts generated by large language models (LLMs) meaningful, or are they merely simulacra of language? Against a recent trend in AI scholarship that views LLMs as little more than "stochastic parrots," in Koch (2025) I use a Kripke-inspired causal theory of reference to argue that LLMs can use names and kind terms with their usual referential properties. Green (2025), a response to Koch (2025), rejects the causal-theoretic account of LLM-reference and proposes an Austin-inspired alternative. The present paper defends the Kripkean approach and raises objections to Green's alternative.

Keywords

Large Language Models; Chatbots; Meaning; Reference; Kripke; Austin

1. Introduction

Are the texts generated by large language models (LLMs) meaningful, or are they mere language simulacra? Against a recent trend in AI scholarship that sees LLMs as little more than "stochastic parrots," I have argued in a paper in this journal (Koch, 2025) that they are capable of using names and kind terms with their usual referential properties, thus bringing us at least a little closer to the idea that LLM-generated texts are meaningful. My argument for this claim is based on a causal-historical view of reference in the spirit of Kripke (1980). On this view, names and kind terms must be fixed to their referents via ostension or a reference-fixing description and can then be transmitted to other speakers via communicative chains. Based on this view, I have argued that LLMs can inherit reference from their training data, provided that the design and training process of LLMs satisfies the criteria for reference-preservation; and on the basis of observations of their training process, I have argued for some optimism that they do.

In an interesting response to my article, Mitchell Green raises an objection to my account and proposes a supposedly superior alternative. According to Green, my account relies on "needlessly controversial" assumptions about the design architecture of LLMs, that can be easily avoided by an alternative account that relies not on Kripke's historical view of reference, but on John Austin's notion of a phatic act Austin (1975). The recognition of phatic acts, Green argues, allows us to see that syntactically competent LLMs are capable of using language as language, and thus with the same meaning and reference that it normally has.

This paper is a response to Green (2025) in which I will argue for two claims: First, that Green's supposedly simple solution to the problem of LLM-reference is open to serious objections and, moreover, cannot escape the very assumptions it seeks to avoid; and second, that Green's criticism of my own account is at least partly based on a misunderstanding and, in any case, can be addressed without cost. In conclusion, it will turn out that those who wish to build a theoretical framework for LLM-reference should choose Kripke rather than Austin as their companion.

2. Defending the causal-theoretic account of LLM-reference

My argument for the claim that names and kind terms that appear in texts generated by LLMs (LLM-reference, for short), is based on the causal theory of reference (CTR), which goes roughly as follows: a pattern produced by a speaker S is a proper name N referring to object O just in case S is part of a causal chain of speakers whose use of N can be traced back to an initial baptism using N to name O. To be part of such a causal chain, each speaker must intend to use O with the same referent as the speaker(s) from whom they acquired the name.

A crucial question is exactly what kind of intention is required for a speaker to become part of an existing causal chain. In the paper I argue – and Green does not seem to disagree – that what is required is that when a speaker adds the name to her mental lexicon, she intends that this name have the same reference as it had before; but when she uses it later, the absence of conflicting intentions is sufficient for the name to retain its reference.

LLMs are machine learning systems that are trained on large and diverse corpora to predict the probability of a token (e.g., a word) based on its preceding or surrounding context, and then fine-tuned to align their responses with certain human values, such as helpfulness, harmlessness, and honesty. Assuming that the names and kind terms that appear in the training data are referential, the crucial question is whether the training process of the LLM is reference-preserving. This question is particularly pertinent because, as far as we know, LLMs have no intentions. However, I argue that the LLM's lack of intentions is compensated for by its design architecture:

> Whereas human agents typically secure continuity by forming the intention of using the name in the same way as those from whom they have picked it up, it is the design architecture of the LLM that serves this function for them. LLMs are built to pick up and henceforth apply words in just the same way as these words are used in the texts it learned from. This feature of their design architecture secures continuity between what names refer to in the training data and what they refer to when the LLM uses them later on. (Koch, 2025 p. 27)

The idea here is that, despite Kripke's appeal to intentions, reference-preserving intentions are not essential; rather, what we need is just something that ensures continuity between past and future uses. And while humans achieve this continuity through referential intentions, LLMs secure it through the way they are designed. After all, the LLM is designed to mimic human language use as closely as possible; and given its inability to use a name with the intention of referring to anything other than what it previously referred to, its uses of a name will be continuous with those in the training data, and thus reference will be preserved.

Green takes issue with this argument. The passage he is most unhappy with is the following, which he refers to as "LLM Design": "LLMs are built to pick up and henceforth apply words in just the same way as these words are used in the texts it learned from." Green raises two objections to LLM Design, which I will discuss in turn. The first is as follows:

> LLM Design is a claim about how these machines are designed, which is not quite the same as a claim about what they do. However, their capacities are presumably going to constrain their design because the engineers who build them are aware of those capacities. Either way, LLM Design is contentious at best. First of all, at the level of the mechanics of how LLM's are trained, we find them building representations of linguistic units in a continuous vector space, and doing so in a manner that is based on those units' statistical distribution in a large corpus (Millière and Buckner 2024). That is not what appears to occur

> when human speakers learn or use language. Presumably, some psycholinguistic theories of how words are learned and used make appeal to vectors in this way. But then such a theory would have to be shown to be correct, and its competitors refuted, for the "just the same way" claim in LLM Design to be acceptable. (Green, 2025 p. 38)

I think this objection misses the point. I do not believe or claim that humans and LLMs are psychologically similar, or that they use the same or similar information processing techniques. In particular, I do not subscribe to a psycholinguistic theory in which humans "appeal to vectors" in the same way as LLMs. Rather, I argue that LLMs are built by their human inventors to do something, namely, "to apply words in just the same way as these words are used in the texts [they] learned from." That is, LLMs are designed (and fine-tuned) to mimic human language use as closely as possible. This idea is compatible with the fact that they use quite different mechanisms to achieve this.

Again, I think it is helpful to remember the broader dialectic: Kripke introduces the requirement of a reference-preserving intention to rule out cases where a speaker does not continue the original name-using practice, or continues some other practice instead. To preserve reference, we need something to ensure that this is not the case. People are free agents, so it is up to them whether they continue a given practice or not. So what matters is their intention to do so. In contrast, LLMs are not free agents, so it is not up to them whether they intend to continue a given practice or not. Rather, they are designed to continue the name-using practices that are fed to them in their training data. Either way, reference is preserved.

Here is Green's second objection:

> Second, at the level of the performance of these machines, what they do with words, and what human users might do with the very same words, can diverge drastically. One of those texts might have been the record of a human speaker referring to a perceptually salient object, like a beach ball, and describing it as red. LLM lack perceptual capacities and as a result are unable to describe a perceptually salient object as red, or at least to grasp such a description in anything like the way that a normally sighted human user can do. So too, some of those texts might be records of apologies or warnings, and Koch has chosen to refrain from imputing cognitive sophistication to LLM's of a kind that would enable them to perform such acts as these (Koch, 2025). So here again we have an apparent dissimilarity between how LLM's use language and how ordinary human speakers do, and here again we see the implausibility of the "just the same way contention. (Green, 2025 p. 38)

This passage criticizes a view that I do not hold. I am not suggesting that there are no interesting differences between LLMs and humans. Current LLMs have no perceptual capacities, and so they are unable to describe a perceptually salient object as red, etc. Similarly, they may not be able to perform certain illocutionary acts that humans are capable of. But I fail to see how these observations, plausible as they are, stand in the way of LLM-reference. On the contrary, it seems to me that these observations are perfectly compatible with my claim that LLMs are designed to continue the name-using practices with which they have been fed in their training data. After all, it is precisely the point of CTR that speakers need *not* be able to perceptually identify the referents of the names they use. Similarly, it is not clear why an LLM would have to be capable of performing the full range of human speech acts in order to borrow reference from human users. In the human case, it seems entirely unproblematic to accept that someone who, for whatever reason, is unable to perform certain speech acts – such as apologizing or issuing warnings – can nonetheless refer with proper names.

I conclude that my Kripke-inspired argument for LLM-reference survives Green's objections unscathed. Is there an alternative argument for LLM-reference, one that is less controversial and unlimited in scope? This is what Green claims; unfortunately, however, Green's argument to this effect faces objections that are so serious that the argument is ultimately untenable. I will explain why in the next section.

3. Against the phatic acts account of LLM-reference

Green is open to the idea that the Kripke-inspired argument presented and defended in the last section can be salvaged. He writes: "Might there be a way of weakening that claim to something like, "largely the same way," so that the argument still goes through?" to which he answers "perhaps." But Green thinks that we should not waste our time doing so, because there is an easier, i.e., more direct, less controversial, and more general argument for LLM-reference. This argument is not based on Kripke's CTR, but on Austin's notion of a phatic act. Here is what Green says about phatic acts:

> In a *phatic act*, a speaker makes an utterance in a way that is sensitive to the part(s) of speech contained in the grammatical structure (if any) of the expression used. Without knowing its meaning, I might utter the Swahili sentence, 'Mbwa anabweka,' in a way sensitive to the fact that its first term is a noun and the second is a verb, and that the entire string is an indicative sentence […] In so doing, I have performed a phatic act.
>
> One symptom of an act's being phatic is that it can be reported in direct discourse: A third party might report, "Mitch said, 'Mbwa anabweka.'" Further, when so quoted, these words retain their linguistic meaning: In the phatic act of uttering 'Mbwa anabweka,' 'Mbwa' refers to dogs, and so on […] By contrast, if dust settles on the floor in such a way as to spell out the words, 'Mbwa anabweka,' no phatic act has been performed. The reason is that the etiology of that configuration was not sensitive in any way to its grammatical structure. The result is that when a randomly produced configuration of dust on a countertop appears to spell out 'Mbwa anabweka,' there are in fact no words there. (Green, 2025 p. 35)

Based on this characterization of a phatic act, Green goes on to argue that LLMs are quite capable of performing phatic acts, and thus of using bits of language with the same semantic properties – including reference – as those bits have when uttered by human users:

> As we noted, performing a phatic act requires using language in a manner constrained by its grammar. But then a machine designed to convey information by means of conventional language, and in a way sensitive to grammatical structure and different parts of speech, will also perform phatic acts. This is pertinent to recent developments in LLM technology, because we now have strong evidence that these machines can acquire sophisticated syntactic competence based on relatively modest exposure to training data (Millière, 2024). Accordingly, we have substantial grounds for concluding that when LLM's produce what appears to be language, they are (at the very least) performing phatic acts, and thus appearances are not deceptive: the word- and phrase-like patterns that these machines generate really are words and phrases, and carry their standard linguistic meaning when so generated. Thus, when LLM's produce proper names and other referring expressions, those expressions do refer in a way dictated by their linguistic meaning. (Green, 2025 p. 35)

Green believes this to be an approach to LLM-reference that "is not limited to proper names, noun phrases, or kind terms, but to meaningful expressions generally" and that "does not rest on any controversial claims about how similar LLM's are to human users." If Green were right, this account would indeed be superior to mine. Unfortunately, however, the account is untenable.

Green's account sets the bar for performing phatic acts very low. He suggests that all it takes for someone to perform a phatic act is for that person (or machine, for that matter) to produce a pattern "in a way sensitive to grammatical structure and different parts of speech," or, as he puts in the beginning, "in a way sensitive to the fact the first term is a noun and the second is a verb, and that the entire string is an indicative sentence." Green argues that LLMs can do this "because we now have strong evidence that these machines can acquire sophisticated syntactic competence based on relatively modest exposure to training data." In effect, then, Green argues that all that is necessary for someone to perform phatic acts – and thus to use language in a meaningful way – is for the speaker to be syntactically competent.

Now, the problem with this argument is not that it trades in unrealistic assumptions about LLMs, but that it makes unrealistic assumptions about how easy it is to perform phatic acts. In particular, Green seems to suggest that doing so does not require communicative intentions. This point is dialectically relevant because Green and I share the assumption that LLMs cannot have such intentions. This assumption also played a crucial role in my own account of LLM-reference, since it triggered the whole discussion of whether LLMs are able to secure referential continuity without having any intentions. It turns out, however, that speakers cannot perform phatic acts without having certain communicative intentions. This is acknowledged, albeit somewhat implicitly, by Austin himself (italics mine):

> The phatic act is the uttering of certain vocables or words of certain types, belonging to and *as belonging to*, a certain vocabulary, conforming to and *as conforming to* a certain grammar. (Austin, 1975 p. 95)

Austin emphasizes that it is not enough to utter certain vocables or words that actually belong to a particular language (while being sensitive to grammar); one must also utter these vocables and words *as* belonging to that language, that is, with the *intention* that they belong to that particular language. Without such an intention, the sounds may only accidentally resemble the words of a language, which is not enough for them to have the semantic properties of the actual words of the language. This condition for phatic acts is widely recognized in the literature. Here is Indrek Reiland:

> What does performing a phatic act require of the speaker? Although Austin himself doesn't elaborate, it is plausible that the speaker must be to some degree phonologically competent with the language and know that the sounds uttered belong to a language and are meaningful (even though she doesn't have to be semantically competent and grasp their meanings). Furthermore, as Forguson puts it, it plausibly requires that the speaker have something akin to *intentions* to produce a sound that counts as utterance of the sentence of the relevant language. (Reiland, 2024 p. 5)

Note that this point is not only exegetically plausible, but also on its own terms. To use one of Reiland's examples, "if a phonologically and syntactically competent speaker coughs and makes a noise indistinguishable from 'go' then she hasn't performed a phatic act because she didn't have the right intentions" (2024 p. 5); and, to use an example from Searle (2012 p. 35), if an American soldier captured by Italian troops during World War II tries to pass himself off as a German by uttering the phrase "Kennst du das Land, wo die Zitronen blühn?" without knowing what this phrase means, then he commits a phatic act only if he has the intention of producing sounds that count as an utterance of a sentence in German. By parity of reasoning, this leads to the conclusion that even a syntactically competent LLM uttering a sentence in English will have performed a phatic act – and thus used language rather than a mere language simulacrum – only if it has the intention that this be a sentence in English. Bottom line: If LLMs cannot have intentions, then they cannot perform phatic acts.

Now, Green might either be willing to argue that these authors are wrong, and that people (or machines) can indeed perform phatic acts without intentions, despite appearances to the contrary. But that would require a separate argument, and I am not optimistic about the prospects for such an argument. Or, alternatively, he could claim that, contrary to our shared working assumption, LLMs are capable of having the relevant intentions. But then the argument loses its dialectical force, because it is described as less, not more, controversial than my Kripke-inspired argument.

There is another reason why everyone should be skeptical about the idea that the recognition of phatic acts can get us beyond the Kripkean picture. The reason is that the conclusion of Green's argument amounts to a reductio of its premises: if Austin's view really entailed (which I doubt) that people can use bits of language in meaningful ways without knowing what those bits mean, or even without intending them to be expressions of a particular language, this would show little more than that Austin's view is wrong. To see this, consider again the debate over what proper names refer to. According to descriptivists, it is the identifying content that speakers associate with the name. According to Kripke, it is an unbroken communicative chain to which one can attach oneself merely by having the intention to do so. Green's view is that neither is necessary – you just have to be syntactically competent.

This view faces counterexamples. To take just one, suppose that I, a syntactically competent speaker of English, utter a random name, say "Buguhajorishka." I do not know of the existence of a person with that name, nor is there any causal relationship that connects me to such a person. But suppose there is a person with that name. Green's view seems to imply that I refer to that person (or at least I don't see why I shouldn't), but I clearly don't. There is also a deeper reason for rejecting this view. Whatever their merits or demerits, both descriptivism and CTR deserve credit for offering *prima facie* plausible explanations of what connects the use of a name with its referent. Green's view, on the other hand, seems to imply that syntactic competence is sufficient for reference. But how could something as unspecific as syntactic competence provide the necessary glue between a name and its referent?

4. Conclusion

In light of these serious problems with Green's Austin-inspired argument for LLM-reference, I conclude that my Kripke-inspired argument fares better. As we have seen, Green's account fails to live up to its promise of providing a simpler, less controversial route to LLM-reference because it cannot shake off the appeal to intentions that it officially rejects. Moreover, the discussion has shown that there are independent reasons to be skeptical of a view that makes meaning and reference as easy as Green's. For what is lost along the way is a realistic consideration of what successful reference requires, together with a plausible explanation of how it can be achieved.

5. Bibliography

Austin, J. L. (1975). *How to do things with words: the William James Lectures delivered at Harvard University in 1955* (2. ed). Clarendon Press.

Green, M. (2025). Large Language Models and the Varieties of Meaning. *Philosophy of AI*, *1*, 34–40. https://doi.org/10.18716/OJS/PHAI/2025.11652

Koch, S. (2025). Babbling stochastic parrots? A Kripkean argument for reference in large language models. *Philosophy of AI*, *1*, 19–33. https://doi.org/10.18716/OJS/PHAI/2025.2325

Kripke, S. (1980). *Naming and Necessity*. Blackwell Publishers.

Reiland, I. (2024). 'Austin vs. Searle on locutionary and illocutionary acts.' *Inquiry*, 1–26. https://doi.org/10.1080/0020174X.2024.2380322

Searle, J. R. (2012). *Speech acts: an essay in the philosophy of language*. Cambridge Univ. Press. https://doi.org/10.1017/CBO9781139173438