Philosophy of AI
Vol. 1, 2025, 19-33
10.18716/ojs/phai/2025.2325

BABBLING STOCHASTIC PARROTS? A KRIPKEAN ARGUMENT FOR REFERENCE IN LARGE LANGUAGE MODELS

Steffen Koch

Bielefeld University, Department of Philosophy

Abstract LLMs perform surprisingly well in many language-related tasks, ranging from text correction or authentic chat experiences to the production of entirely new texts or even essays. It is natural to get the impression that LLMs know the meaning of natural language expressions and can use them productively. Recent scholarship, however, has questioned the validity of this impression, arguing that LLMs are ultimately incapable of understanding and producing meaningful texts. This paper develops a more optimistic view. Drawing on classic externalist accounts of reference, it argues that LLM-generated texts meet the conditions of successful reference. This holds at least for proper names and so-called paradigm terms. The key insight here is that the LLM may inherit reference from its training-data through a reference-sustaining training mechanism.

Keywords:
Large Language Models; Chatbots; Meaning; Reference; Semantic Externalism

1. Introduction

Large language models (LLMs) such as BERT, GPT-4 or Gemini play an increasingly important role in many areas of our professional and personal lives. These transformer-based machine learning systems are trained on large and diverse corpora taken from the Internet – incl. encyclopedias, academic articles, books, or websites – to predict the probability of a t oken (e.g. a word) based on its preceding or surrounding context, and they are then fine-tuned (often via reinforcement learning from human feedback, RLHF) to align their responses with certain human values, such as helpfulness, harmlessness and honesty.[1]

LLMs are remarkably successful in performing a wide range of language-involving tasks: They write better essays than the average undergraduate student (Herbold et al., 2023), program better than the average software engineer (Bubeck et al., 2023; Savelka et al., 2023), rank in the 80-99th percentile on graduate admissions tests (OpenAI et al., 2023), and solve many difficult mathematical problems (Zhou et al., 2023) while expressing their solution in the form of a Shakespearean sonnet. In light of these remarkable feats, it is natural to think that to-date LLMs have mastered language and therefore to interpret the texts they produce in much the same way as human text or speech.

Recent work in the philosophy and psychology of AI has begun to question this interpretation. Bender et al. (2020) warn us that there is a "tendency of human interlocutors to impute meaning where there is none" and that this "can mislead [...] the general public into taking synthetic text as meaningful" (p. 611). According to them, LLMs are merely

1 See, among others, Millière & Buckner (2024) and Gubelmann (2024) for detailed explanations of how LLMs work.

babbling "stochastic parrots" that produce the image of human language but in fact fall short of meaningful communication. Mallory (2023) argues that LLMs do not actually produce meaningful texts, and that to treat them as such is to engage in a kind of useful fiction. Lake & Murphy (2023) argue that there is a principled limit to an LLM's ability to build the knowledge structures that form part of the basis of word meanings. Gubelmann (2024) argues that LLMs are not yet capable of performing speech acts.

This paper defends a more optimistic view of the LLM's prospects for using language. Specifically, I will argue for the following two related claims: (i) many expressions produced by to-date LLMs refer to objects, events, relations and properties in the worlds, and (ii), when they do, they also refer to the *same* objects, events, relations and properties as our uses of these expressions. Throughout, my discussion will be focused on semantic (as opposed to speaker's) reference. Thus, I will not consider to whom or what an LLM *intends* to refer by *using* particular lexical items; instead, I will focus on whether these items refer *themselves*, independently of any specific communicative intentions.[2]

Why does it matter whether LLM-generated texts refer? Reference is an important feature of language. We can think of it as the glue between linguistic signs, on the one hand, and the objects, events, relations, and properties they stand for, on the other. It is only in virtue of the fact that our words refer to things that we can make comprehensible and truth-evaluable claims about them. If the words and sentences produced by LLMs lacked reference, or referred differently, then it would be wrong to interpret them in the way we often do. In considering whether LLM-generated texts refer, I also hope to contribute to the larger question of whether LLMs produce language or a mere language "simulacrum": something that looks or sounds like language, but lacks any trace of meaning. Not everything that *looks* like language therefore *is* language. If I say "Kripke was a philosopher," this is a sentence of English and it means *that Kripke was a philosopher*. By contrast, if a random letter generator produces the sequence "Kripke was a philosopher," this is not a sentence of English, much less one that means that Kripke was a philosopher. The broader question I aim to contribute to in this paper is whether LLMs are more like this random letter generator or more like me.

LLM reference is puzzling because it falls outside the remit of traditional treatments of reference. For one thing, an LLM is not like a human agent. Unlike LLMs, human agents have actual experiences with the outside world, and they have rich inner lives consisting of, among other things, knowledge, beliefs, desires, associations, and communicative intentions. Theories of reference that are designed for human agents can (and do) appeal to these and potentially other factors to explain reference. For another, the texts generated by an LLM are also unlike books, newspapers or research papers in that they are not authored by human agents. Whereas it makes sense to say that the words written in a book refer to people, objects, relations etc. in the outside world partly in virtue of their human authors' properties – e.g., their beliefs, intentions and social relations –, such an explanation does not work for LLM-generated texts.[3]

Considerations along these lines have given rise to a widespread skepticism about an LLM's capacity to acquire reference. Bender & Koller (2020) compare the situation of an LLM to that of a fictional superintelligent octopus that learns English by eavesdropping on people on land. Even if the octopus learns to perfectly simulate how people on land use the word "coconut," Bender and Koller argue, it would still not know the meaning of "coconut." This is because the octopus has never been on land, and therefore lacks something that Bender and Koller take to be a necessary condition for mastering the meaning of a word: the ability to connect utterances containing this word to the world (p. 5188). For the same reason, they hold, purely statistical data about the distribution of word forms across corpora do not enable an LLM to acquire meanings.[4]

2 See e.g. Kripke (1977) for an account of the difference between semantic reference and speaker's reference. See the end of Sect. 4 for further discussion.

3 See Van Woudenberg et al. (2024) for a discussion of authorship is LLMs.

The problem raised by Bender and Koller is reminiscent of the much-discussed symbol grounding problem (Harnad, 1990). As Millière & Buckner (2024) formulate the problem,

> for symbols in NLP systems to have intrinsic meaning, there needs to be some grounding relation from the internal symbolic representations to objects, events, and properties in the external world that the symbols refer to. Without it, the system's representations are untethered from reality and can only gain meaning from the perspective of an external interpreter (p. 15).

In the context of LLMs, this problem concerns how a system whose input is limited to data about the co-occurrence of symbols can give meaning to those symbols (Mollo & Millière, 2023). Critics hold that word reference must be causally grounded in an extra-linguistic reality through perception, action, or desire (Harnad, 1990; Lake & Murphy, 2023) and that this stands in the way of LLM-generated texts to have reference.[5]

These arguments question the validity of our impression that LLM-generated texts describe how things stand in the world. At the same time, however, they invoke substantial and controversial assumptions about reference. As I will show in the following pages, more plausible approaches to reference suggest a more optimistic assessment of whether LLM-generated texts refer. The approaches I have in mind here are semantic externalist views of reference that have been developed in the philosophy of language over the last four to five decades. Following these approaches reveals, among other things, why achieving reference is less demanding than many authors assume, why we can be quite optimistic about reference in LLM-generated texts, and why the octopus test is not an adequate test for reference.

I am not the only author to take an externalist approach to reference in LLMs. Mandelkern & Linzen (2024) also use insights from the externalist tradition to argue that LLMs, like humans, can inherit reference from their training data. However, I will develop t his argument more fully, distinguishing between different versions of externalism and considering in more detail whether they allow the application to LLMs. A crucial point here is whether externalist accounts of reference require any particular intentions on the side of the speaker that LLMs might be incapable of forming. While Mandelkern and Linzen leave this question open, I will argue, pace Mallory (2023), that this is not so.[6]

I proceed as follows. In Sect. 2, I begin my case for LLM-reference by outlining an externalist theory of reference for proper names in the spirit of Kripke (1980). It will emerge that LLM-reference crucially depends on what kind of intentions are required for successful reference. In Sect. 3, I therefore discuss the issue of referential intentions in more detail, arguing for a rather ecumenical view. Sect. 4 concludes the argument by applying this result to the case of LLMs, arguing that they meet all remaining requirements for successful reference. Finally, Sect. 5 and 6 suggest generalizations of the conclusion obtained in Sect.4. In Sect. 5, I show that, contrary to first impression, the same conclusion can be obtained also on hybrid theories of reference that diverge from the Kripkean orthodoxy. In Sect. 6, it is argued that the same conclusion extends way beyond proper names to the much wider class of paradigm terms.

One preliminary before we start: In what follows, I side with other authors in assuming that to-date LLMs do not have mental states, at least not in the sense in which we do. While this assumption might not be entirely uncontroversial, there is broad consensus about it (Gubelmann, 2024; Millière & Buckner, 2024). If you are skeptical, note that this assumption only makes my endeavor more challenging. If I can succeed in showing that LLMs may refer even under the assumption that they lack mental states, their prospects for reference will only became better once we drop this assumption.

---

4 Bender & Koller (2020) do not distinguish between meaning and reference, so it is unclear what implications they would take their arguments to have about reference; but there is little to suggest that they would not be similarly skeptical with respect to reference. See (Piantadosi & Hill, 2022) for critical discussion of this argument.

5 See Chalmers (2023) for an argument to the contrary.

6 See Cappelen & Dever (2021) for a semantic externalist treatment of AI content that does not specifically address LLMs.

2. Large language models and the causal theory of reference

How can we approach questions of reference in LLMs? A plausible approach is this: First, we consider the conditions under which humans refer; second, we check whether machines like LLMs also satisfy these conditions. So what enables humans to use words and sentences to talk about non-linguistic entities like people or trees? To make the discussion concrete, what enables me to use "Barack Obama" to talk about the actual person Barack Obama?

An initially plausible view is this: Even though I haven't had any direct physical contact with Obama, I've had a lot of Obama-related experiences. I've seen him on TV, read about him in the news, discussed his politics with friends and family, seen lots of pictures of him, etc. In short: I have accumulated a body of detailed knowledge about him. All of this information, or some significant part of it, constitutes the meaning of "Barack Obama." And because this information applies only to the person Obama, and not to, say, Hillary Clinton, "Barack Obama" refers to Obama, not to Clinton (or anyone else). In short, then, the meaning of a term is a body of information that speakers associate with it, and its referent is the object of which this information holds true. Versions of this view were held by Frege (1892) and Russell (1905).

According to Kripke's famous generalization of this view, it can be put roughly as follows:[7] For every proper name N, there is an entity e and a description D_N, such that: (i) D_N applies only (or at least best) to e, and (ii) competent users of N associate N with D_N. This allowed them to say that the description D_N is the meaning (or sense) of N and that the referent of N is the unique object that satisfies D_N.[8] This view entails that meaning determines reference, and that using a name N to refer to something requires that one knows how to distinguish this thing from other possible referents of N – a condition we also find in Bender and Koller's suggested octopus test.

Although this is only a rough outline of a view, it immediately casts doubt on an LLM's ability to use words like "Barack Obama" to talk about things in the world. For one thing, the view requires knowledge on the side of the speaker. Knowledge, however, is a mental state, and we are here working under the assumption that LLMs do not have mental states. For another, even if the LLM *were* capable of knowing, it would still have no way of distinguishing which elements of its world knowledge are part of the reference-determining description associated with "Obama" (the "analytic truths") and which ones are not. Is it part of the meaning of "Barack Obama" that he became US president in 2009? Or that he received 53% of the votes on November 4 in 2008? It is very hard to see how an LLM that is purely trained on syntactic form should go about answering these questions.

There are, however, independent reasons for rejecting the view that reference is established by an identifying description that one associates with a word. This can be seen by considering the following Kripke-inspired thought experiment:[9] Imagine a person, Martha, who is in the unlikely situation of never having heard of Obama. (Perhaps Martha went to live with an isolated tribe of indigenous people somewhere in the South Pacific 25 years ago.) Now suppose that the first thing Martha ever hears about Obama comes from the mouth of a conspiracy theorist, a "birther," who says: "Barack Obama was not born in the United States."[10] After this encounter, Martha seems to be able to use the name "Barack Obama" to talk about Obama. For example, she might repeat to others that Barack Obama

7 It is controversial whether Kripke's interpretation especially of Frege is accurate. See e.g. Yourgrau (2012) for discussion. Here I will not delve into Frege exegesis but simply work with Kripke's useful and influential characterization of his view.

8 John Searle remarked that F_N need not be a single property, but may also be a bundle of diverse properties, such that N denotes the unique object that has most (or the most salient) of the properties contained in F_N (Searle, 1958).

9 The argument that this thought-experiment motivates is called "argument from ignorance and error;" see Kripke (1980).

10 A birther is a conspiracy theorist who believes that Barack Obama was not born in the US and, therefore, can't be the legitimate president.

was not born in the US. She might ask others if it is true that Barack Obama was not born in the US. Or she might ask who Obama is, if he is still alive, what he has done, etc. When Martha says these things or asks these questions, she is talking about Obama – even though she has never met Obama and knows virtually nothing about him. The only thing she has heard about Obama – that he was not born in the US – is false. Hence, successful reference can hardly be a matter of an associated definite description that one associates with the name in question.

To be sure, this Kripkean argument is not the nail in the coffin of descriptivism. Modern descriptivists have explored, and continue to explore, ways of answering this challenge; for example, by adjusting the kind of description that constitutes the link between a name and its referent (Jackson, 1998) or by arguing that (a certain version) of descriptivism follows from assumptions that are shared even by its critics (Kipper & Soysal, 2022). I will not go into these here. Suffice it to say that Kripke's argument from ignorance and error along with his modal argument against descriptivism have convinced many to reject the descriptivist approach in favor of some version of semantic externalism: the view that reference is partly determined by factors that are not necessarily accessible to the speaker. Here, causal theories of reference as developed by Devitt (1981) Evans (1973) or Kripke (1980) have been especially influential.

Here is how Kripke (1980) describes the general idea behind this view:

> Someone, let's say, a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain. A speaker who is on the far end of this chain, who has heard about, say Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can't remember from whom he first heard of Feynman or from whom he ever heard of Feynman. He knows that Feynman is a famous physicist. A certain passage of communication reaching ultimately to the man himself does reach the speaker. He then is referring to Feynman even though he can't identify him uniquely. He doesn't know what a Feynman diagram is, he doesn't know what the Feynman theory of pair production and annihilation is. Not only that: he'd have trouble distinguishing between Gell-Mann and Feynmam. So he doesn't have to know these things, but, instead, a cham of communication going back to Feynman himself has been established, by virtue of his membership in a community which passed the name on from link to link […]

Needless to say, a story along these lines could equally be told about Martha and Obama. Obama acquired his name from his family, who contributed to spreading it across the population. Through a potentially very long chain of communication, the name finally reaches Martha. Although Martha does not know Obama personally, nor anything about what he did, it is by virtue of her membership in a community of speakers who passed the name on to her that she can use the name to refer to the person Obama.

The causal theory of reference consists of two separate and equally important elements:

> *Reference-Fixing:* The referent of a proper name is initially fixed in either of two ways: by an ostensive 'baptism', where the referent of the name is present, or by a stipulation that it is to be whatever satisfies a certain description (Kripke, 1980)

> *Reference-Transmission:* Once reference is established, it can easily be transmitted from speaker to speaker via chains of communication. For a speaker to become part of an existing chain of communication, it is sufficient that she hears someone who is already part of the chain use the name and that she intends to use it just as this person did (Kripke, 1980)

This is more of a sketch than a full-blown theory, as Kripke was the first to admit (Kripke, 1980, pp. 93, 96–97). To develop it further, more needs to be said about communicative chains, what it is for a person to use a name in a certain way, or what the relevant intention consists of. We will come back to these issues shortly. Before we do so, however, let us consider the prospects of such a view for whether LLM-generated texts, or more precisely: the proper names that appear in such texts, refer. We can do this by considering whether LLMs are able to satisfy the conditions imposed in *Reference-Fixing* and *Reference-Transmission* respectively.

Just like the Octopus mentioned by Bender and Koller, LLMs lack any direct contact with the world. Because of this, they are unable to introduce names via ostension, e.g., by declaring that the name "Obama" shall henceforth apply to *this* person. LLMs could, in principle, introduce names by giving definite descriptions and allow users to pick up these names and use them with the same reference. If they did so, it is *prima facie* plausible that they would thereby introduce a new proper name into our language. This being said, reference-fixing is not the case that those interested in LLM-reference should set their primary focus on. More important is the question whether an LLM's use of an *already familiar* name refers at all, and if so, whether it has the same reference as it has when we use it. To answer this question, we need to determine whether LLMs are capable of becoming part of an ongoing communicative chain that ultimately leads back to a particular referent.

LLMs are not like human language users who may simply pick up names in ordinary conversations. However, LLM-based chatbots *do* interact with humans through chat messages. If a human language user who is familiar with a certain name were to use this name in a chat message, and this name were then picked up and re-used by the chatbot it could, at least in principle, inherit its reference from the human language user.[11] But more importantly, LLMs are trained on large corpora. These corpora are the products of human language users – people who are part of a given language community, and whose uses of names like "Barack Obama" are therefore links in communicative chains that ultimately go back to Obama himself. The fact that LLMs are trained on such data could be sufficient for them to "inherit" the reference relations between the words they use and the objects they denote. Whether this is true crucially depends on what exactly it takes for someone or something to become part of an ongoing communicative chain. Here is, in more detail, what Kripke says about the relevant requirement:

> When the name is 'passed from link to link', the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it. If l hear the name 'Napoleon' and decide it would be a nice name for my pet aardvark, I do not satisfy this condition. (Kripke, 1980)

One might think that the requirement of intending to use a name with the same reference as the person from whom one heard it raises a potential problem for an LLM's ability to refer to things. As I said earlier, I am working under the assumption that LLMs do not have mental states, and *a forteriori* cannot form any intentions. If they are unable to form intentions, however, then there seems to be no way for them to have any referential intentions either. This point is also raised by Mallory (2023), who argues that "[w]hatever causal chain ties the output of a bot back to the tokens in a corpus or dataset, it is not secured by intentional repetitions on the part of the machine" (p. 1084).

Assessing whether this concern is warranted depends on what kind of referential intentions are needed for reference-transmission. This leads us into a brief excursus about referential intentions.

11 Most current LLMs such as Chat-GPT do not incorporate the prompts that are given to it into its training data. For this reason, reference that is acquired through chat messages would necessarily remain restricted to the particular chat.

3. Excursus: Referential intentions

The passage just cited from Kripke raises some difficult interpretative as well as systematic issues, discussing which will be important for assessing LLM reference. A first issue is whether what is required is the *presence* of *consonant* referential intentions, or rather the *absence* of *conflicting* referential intentions. On the first option (a), the requirement demands that whenever someone uses a given name, say "Obama," then in order to refer to Obama that language user must have the explicit intention of using "Obama" in exactly the same way as the person from whom she heard it. More precisely yet, there are two versions of this demand: (a1) the person must have the *de re* intention of using the name to refer to a particular person, where this person is, in fact, the same as the one that other language users intend to refer to by using this name; or (a2) the person must have the *de dicto* inten-tion of using the name to refer to the exact same person as other language users do (whoever that person may be, cf. Koch & Wiegmann (2022). The *de re* reading is stronger than the *de dicto* reading, for it requires the speaker to know the person they are referring to. On the second option (b), the requirement merely demands that users of a name do not have the intention of using the name differently than those from whom they picked it up.[12]

A second issue is whether the relevant requirement concerns (α) the initial time of borrowing the reference of a name or rather (β) the speaker's later uses of the borrowed name. The passage quoted from Kripke as well as many interpreters seem to favor (α) (Devitt, 2006, 2008, 2015; Michaelson, 2023; Raatikainen, 2020) while others opt for (β), thus claiming that also a speaker's later use of a name must be accompanied by the relevant referential intention (Kipper & Soysal, 2022; Searle, 1958). Although this option has, to my knowledge, not been explicitly considered by interpreters on either side, it is of course possible that different requirements hold for the initial borrowing and the later uses of the term.

What are we to make of this? Let us first assess the different options that concern the speaker's later uses of a borrowed name (β). I take it as given that (β-a1) is too strong a requirement for it to be plausible. It is the quintessence of Kripke's causal theory of reference that speakers need *not* know the referent of a name, nor any identifying descriptions. A re-quirement to intend to refer to a particular person by using a name would countervail the agenda of this theory.[13] For similar reasons, I reject (β-a2). Although this requirement is considerably weaker, it still seems too strong. People typically use names rather automatically, without having any particular positive referential intentions. To be sure, by uttering a sentence, I might and plausibly do have certain communicative intentions – e.g., to tell you something about Obama. But this does not mean that each and every of my uses of "Obama" is accompanied by the positive intention to use this name in exactly the same way as the people from whom I picked it up. I therefore conclude that a requirement along the lines of (β-b) is sufficient for reference preservation. Thus, once a speaker has learned a name, reference is preserved so long as she does not form the intention of using the name differently than those from whom she picked it up.

Let us now consider what sort of referential intentions, if any, are required at the initial time of borrowing a name. Here, one might think that it is rather plausible that positive intentions are required. After all, one does not automatically become part of a communicative chain, simply by beginning to use a name that others have used before. What is needed is some kind of mechanism that ensures continuity between how users have used and con-

12 Again, this could either mean (b1) that they do not have the *de re* intention of using the name to refer to a person who is in fact not the one that is picked out by the other's use of the name, or (b2) that they do not have the *de dicto* intention of deviating from prior use (whatever this prior use might have been). Since there is no principled problem with an LLM lacking a certain intention, the difference between those two options is irrelevant for present purposes.

13 Notably, this point is accepted even by those who argue that Kripke's view carries strong descriptivist commitments. For instance, Kipper & Soysal (2022) describe the requirement such: "for a candidate (external) relation to determine reference, the speaker must intend to refer to the things to which they stand in this relation" (p. 5). This is in line with the *de dicto* reading, but not the *de re* reading.

tinue to use the name and one's own future use of this name. I will grant this much, and thus side with authors such as Devitt (2006, 2008) or Raatikainen (2020), who also propose such a requirement. In the case of human agents, referential intentions may plausibly instantiate the required mechanism. Partly for the reasons given above, I am skeptical about (α-a1), since it would require the name-borrower to have the intention to refer to a particular person, which in turn seems to require knowledge of who that person is. (α-a2), on the other hand, seems to be a plausible necessary condition for successful reference-borrowing, at least as far as human language users are concerned (more on this below).

Before turning back to LLMs, let me address an objection that Kipper & Soysal (2022) have recently formulated against the position I am taking here. These authors argue against the view that positive referential intentions are required only at the time of the initial borrowing. They write:

> Devitt's suggestion [that the requirement is about reference-borrowing rather than later uses of the name] isn't credible [...] Just as a speaker can stipulate a term to have a certain reference when they first hear it, they can perform such a stipulation later. But Devitt's suggestion seems to imply that this is impossible. Accordingly, if a speaker once had the intention to defer to others' usage, they won't be able to use this term with a different reference later, even if they want to. Such a view would entail that we only have control over the meanings of our words when we first encounter these words, which seems no less absurd than the view that we have no control at all over those meanings (p. 6).

This objection misses the mark. Devitt's (and my) point is not that, once a speaker has borrowed a name, reference will remain constant even if the speaker later forms the intention to use the name differently. The point is, rather, that later uses of a name need not be accompanied by consonant referential intentions for them to still refer. Put differently, Devitt (and me) agree with Kipper and Soysal that conflicting referential intentions are *sufficient* for different reference; but this does not mean that consonant referential intentions are *necessary* to preserve reference.

To summarize the results of this excursus, the picture I endorse is as follows. For a speaker to become part on an ongoing communicative chain about a particular proper name *N*, that speaker must do something to ensure continuity between *N*'s prior uses and their own future uses of *N*. Forming the intention to use *N* in the same way (whatever that may be) as the other users of *N* is a plausible way of doing this. Once this is done, however, and the name has become part of the speaker's mental lexicon, it is sufficient for successful reference that the speaker does not form the intention to use the name differently than before. This requirement is thus satisfied by the absence of conflicting referential intentions rather than by the presence of consonant ones.

4. Reference in LLMs: Proper names

What are the implications of this discussion for whether the names that occur in LLM-generated texts refer? If I am right, then for a name to refer, it is not generally necessary that the user of this name has any particular referential intentions. Since we assume that LLMs cannot have any intentions, this is good news for LLM-reference. However, the discussion has also shown that reference-borrowing requires a mechanism that ensures continuity between other's prior uses and one's own future uses of a name. Does this requirement stand in the way of LLM-reference?

When it comes to LLMs, reference-borrowing happens within the LLM's training-phase. Very roughly, to-date LLMs are trained in three steps. In step one, "pre-training", the LLM is trained on large cleaned up data sets to become a 'next token predictor.' In step two, "instruction training", the LLM is trained to understand instructions and output plausible answers to these instructions. This stage is necessary because, without it, the LLM might respond to questions such as "What is the capital of France?" with yet another question, such

as "What is the capital of Germany?." In step three, "reinforcement learning through human feedback," the LLM is further fine-tuned on the basis of human feedback to align better with the intended value specification, e.g., to be harmless, honest, and helpful.

The step that is most essential for the LLM's language acquisition is step one. Here, the LLM is fed with chosen corpora. This process is not too far afield from how humans pick up formerly unknown names through texts or human speech. However, above we have seen that more is needed to become part of an ongoing communicative chain about a particular proper name: one must do something to secure continuity between prior uses of this name and one's own future uses. 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. So, even though an LLM cannot form any intentions, the way it is built ensures that the reference that is undoubtedly present in the texts from which it is trained is transmitted to it. If this is right, then no intentions are needed for the LLM to generate texts with the same reference patterns as the texts from which it is trained.

Is this still a view that Kripke could agree with? I think it clearly is. Let us remind ourselves of the reason *why* Kripke introduces the requirement of having the intention to use the name in the same way as those from whom one has picked it up. Kripke is quite explicit that this requirement is to ensure continuity between prior uses of the name and the way that the speaker will use it henceforth. He wants to rule out cases in which a speaker intentionally and knowingly deviates from standard usage. Now one way to achieve this is by requiring consonant referential intentions on the side of the speaker. Another one, more apt for intention-less AI systems, is to build the system in such a way that such cases cannot occur. And this is precisely what I claim to be the case with respect to LLMs.[14] I therefore conclude that, based on widely shared Kripkean premises about the reference of proper names, a good case can be made that the names that occur in LLM-generated texts refer.

Before moving on, let me say a few words about how the conclusion we have reached relates to the question of whether LLMs *themselves* are capable of referring, i.e., whether they can speaker-refer to things or persons with their outputs. As I noted at the outset, my concern here is with semantic reference rather than speaker's reference. The broader concern that has dricen my inquiry is whether LLMs produce language or mere language simulacra, and an important piece of this larger puzzle is to find out whether the texts they produce bear referential relations to things in the world. In line with this, the Kripkean view I have developed here is one that specifies the conditions of semantic reference rather than speaker's reference. This holds despite the fact that Kripke's view trades in referential intentions. According to Kripke, semantic reference requires a very general referential intention – namely, the intention to use a name in the same way as those from whom one has picked it up. But this general intention is not to be confused with the specific intentions that track the speaker's reference (Kipper & Soysal, 2022; Kripke, 1977).[15]

There remains, of course, the question of LLMs and speaker's reference. The cases that gave rise to the distinction between semantic reference and speaker's reference are those in which the two are erroneously separated: a speaker does not intend to deviate from standard usage, but the person he or she intends to refer to is in fact not the person to whom the name semantically refers. I have argued that LLMs cannot have specific referential intentions, but that their training mechanism still allows them to become part of communicative chains. Depending on one's broader commitments, this could mean either that

14 Note also that this is a contingent feature of LLMs. In principle, an AI system *could* be trained so as to deviate from human language use. The point is that they are in fact not, for the above-mentioned step one of an LLM's training process, designed to make it a perfect token-predictor, guarantees for the required continuity.

15 Kripke stresses the difference between these two kinds of intentions when he discusses the Madagascar case (Kripke, 1980).

LLM-generated texts exhibit semantic reference without speaker reference, or it could mean that for LLMs, semantic reference and speaker reference are necessarily aligned, so that the LLM always speaker-refers to what its words semantically refer to. Although I have sympathies for the latter view, I will not argue for it here, and so I want my contribution to be understood in terms of semantic reference rather than speaker's reference.

In the remainder of the paper, I want to extend the argument developed over the pre-ceding sections in two directions. First, I will argue that a similar result can be obtained for other variants of the causal theory of reference that are often seen as adversaries to the Kripkean picture. Second, I will argue that the result extends beyond proper names to other semantic categories as well, most notably paradigm terms.

5. Reference in LLMs: Beyond Kripke

Kripke's influence on contemporary discussions of reference in the philosophy of language can hardly be overestimated. But even among those who are on board with his general idea, there is some controversy about how to spell out the details, and in particular about the role that descriptions play in reference-determination. Some authors argue that cases of unintentional reference change provide counterexamples to Kripke's causal theory of reference, because Kripke's view implies that reference remains constant unless someone has the intention of using a name differently.

According to Gareth Evans, the problem is that Kripke focuses on the wrong causal connection. Rather than being concerned with the causal connection between the original use (or "baptism") and our contemporary uses, we should consider the causal connection between the object itself and the information associated with a name (Evans,1973). Whatever turns out to be the object that is the "dominant causal source" of this information, understood as a set of belief-like mental states, this object is the referent of that name. This view accommodates cases of reference change by allowing that the information one associates with a name, and thereby the object that constitutes the dominant causal source of it, may change over time.[16] Something similar is suggested by Michael Devitt, who opts for a version of the causal theory that allows for multiple groundings. In Devitt's view, the reference of a proper name is the object that causally grounds a particular subset of a speaker's thoughts, namely those that dispose her to use the name (Devitt, 1981).[17]

Two points are worth emphasizing. First, comparing Evans' and Devitt's views with Kripke's, we can see that both bring back into play the descriptive content that speakers associate with terms. But unlike Frege and Russell, Evans and Devitt do not hold that meaning (or associated content) *determines* reference. What a term refers to is not the object picked out by the associated descriptive content, but the object that is at the end of the causal chain that led you to have that information. The crucial relation, then, is a causal relation, not a relation of semantic fit.

Second, both Evans' and Devitt's views incorporate Kripke's idea of reference transmission through communicative chains. The information one associates with a term may well come from testimony rather than from direct contact with the thing in question. Reference may be transmitted through possibly long chains of testimony that ultimately terminate in first-hand experience.

Assuming for the sake of argument that a view along the lines of Evans and Devitt is indeed more apt than Kripke's – what are the implications for reference in LLMs? A first problem is that both views are couched in mentalist vocabulary. Evans speaks of "the information that one associates with a name," which he clearly understands in mentalist terms as knowledge or beliefs. Devitt mentions "the thoughts that cause one to use a name." Since

16 See Evans (1982); Koch & Wiegmann (2022) for a more detailed exposition of how reference change is explained on Evans' view.

17 See Michaelson (2023) for an attempt to defend Kripke's view against Evans-style counterexamples by endorsing "futurism" about names, that is, the claim that what a name refers to at a given time t may partly depend on things that happen after t. See Ball (2020, 2024); Jackman (1999, 2005, 2020) for similar views.

we are here working under the assumptions that LLMs have no mental lives and are thus unable to have mental states like knowledge or beliefs, this could stand in the way of LLM-generated texts to exhibit reference. So, if the Evans/Devitt view is correct, and this view is indeed committed to the claim that mental states are necessary for reference, then there is no reference in LLMs.

But it is not clear that Evans and Devitt are so committed. The fact that they both mention mental states in their formulation of the view does not mean that mental states are indispensable. Evans and Devitt were concerned with *human* language and had no claim to cover reference in LLMs or AI in general. The crucial question for us should not be whether Evans and Devitt mention mental states, but whether a version of their view – one that is different in letter but similar in spirit – can be formulated without them. Or, to use a phrase recently coined by Cappelen & Dever (2021), whether the Evans/Devitt view can be "de-anthropocentrized" via "anthropocentric abstraction," which the authors characterize as follows:

> In anthropocentric abstraction, we take existing externalist accounts of content determination and abstract away from […] contingent and parochial features of human communication to reveal a more abstract pattern that is realizable in many kinds of creatures. (Cappelen and Dever, 2021; p. 70).

In the case of Evans, the crucial concept we need to de-anthropocentrize is information. Evans understood information as a set of belief-like states, including knowledge, beliefs, and potentially other contentful mental states. But there is a very clear sense of "information" in which this term refers to an abstract type that can be tokenized in different formats. For example, we can say that *Barack Obama is a former US president* is a piece of information that is stored in my notebook, in my long-term memory, and on my computer, even though each of these media realizes this information quite differently. In the case of LLMs, a likely candidate for the information associated with a name is roughly *data about the statistical distributions of the words surrounding that name*, stored on a server. This data is accumulated in step one of the LLM's training phase from human language corpora. Nothing stops us from saying that whatever turns out to be the dominant causal source of the information that people whose texts are included in these corpora associate with the name in question is the referent of the LLM's use of the name.

Devitt's view could be de-anthropocentrized in a similar way. Here the crucial concept is thought rather than information. Though thoughts are less susceptible to a non-mentalist reading than information, they can be understood in functionalist terms in the same way. An appropriate LLM analog of *thoughts that cause one to use a name* might again be *data about the statistical distributions of the words surrounding that name*. Modified in this way, the question of what a term as used by an LLM refers to is pushed back to human language use, since the data in question come from human language corpora.

These rough sketches of how to de-anthropocentrize Evans's and Devitt's respective views leave some questions about details open. Nonetheless, there is reason to assume that even their variants of the causal theory of reference suggest that LLMs that are trained on human language corpora may use names to refer to things in the world.

6. Reference in LLMs: Beyond proper names

The previous discussion circled entirely around proper names. If I am right, then proper names that are used by an LLM refer exactly as when they are used by us – provided that the LLM is trained on sufficiently large and qualified data sets. But proper names make up only a small fraction of the words used by an LLM. In this section, I will argue that what holds for proper names equally holds for the much wider class of so-called paradigm terms, including also natural kind terms such as "water" or "gold" and social kind terms such as "woman," "money," or "point guard."

Already in *Naming and Necessity,* Kripke observed that there is a close connection be-

tween proper names and natural kind terms. Proper names are *rigid designators*, i.e., they re-fer to the same object or class across all possible worlds (Kripke, 1980, p. 46). The same is true of natural kind terms. Here, what secures stability across all possible worlds, Kripke argued, are the kind's essential properties – those properties in virtue of which a token instantiates a given kind. In the case of water and other chemical kinds, this is most likely it's chemical composition, i.e., consisting of H_2O. One consequence of this view is that we often don't know whether something belongs to a given natural kind or not. It took until the 1780s to discover the chemical composition of water, and it is far from obvious to lay people what chemical structure a given substance has. Because of this we cannot be sure what is or is not in the extension of a natural kind term such as 'water'. But if it is not our beliefs or state of information that determines what is in the extension of a natural kind term, then what is it?

According to Kripke, it is basically the same mechanism of reference-fixing and reference-transmission that we already discussed with respect to proper names. Here is Kripke:

> In the case of proper names, the reference can be fixed in various ways. In an initial baptism it is typically fixed by an ostension or a description. Otherwise, the reference is usually determined by a chain, passing the name from link to link. The same observations hold for such a general term as 'gold'. If we imagine a hypothetical (admittedly somewhat artificial) baptism of the substance, we must imagine it picked out as by some such 'definition' as, 'Gold is the substance instantiated by the items over there, or at any rate, by almost all of them'.[…] I believe that, in general, terms for natural kinds (e.g., animal, vegetable, and chemical kinds) get their reference fixed in this way; the substance is defined as the kind instantiated by (almost all of) a given sample. (Kripke, 1980, p. 136)

If this is right, then there is no relevant difference between proper names and natural kind terms with respect to how reference-fixing and reference-transmission work. Speakers who introduce natural kind terms might have somewhat different intentions, i.e., to use the term to denote all and only those things that are of the same kind as the sample. But since LLM-generated texts typically acquire reference through reference-transmission rather than reference-fixing, this difference is irrelevant for present purposes. Just like proper names, natural kind terms may be passed on from speaker to speaker (or from speaker to machine, for that matter). And if, as I've argued above, the training phase of an LLM ensures that the conditions for reference-transmission are typically satisfied with respect to proper names, the same holds for natural kind terms.

This observation broadens my case for reference in LLM-generated texts from mere proper names to all sorts of natural kind terms, including those referring to chemical kinds ("water," "gold," "jade"), biological kinds ("tiger," "elm tree," "pneumococcus"), and physical kinds ("atom," "electron," "photon"). But we can go further still. For, as Nimtz (2017) convincingly argues, "the modal and epistemic peculiarities commonly considered distinctive of natural kind expressions are in fact traits shared by paradigm terms in general" (p. 125). A further benefit of this view is that the metasemantics of paradigm term does not hinge on any potentially controversial view about the metaphysics of kinds. All that it requires is that the object in question be relatively objective.[18] Paradigm terms are all predicates whose application conditions are (i) relationally determined, (ii) object-involving, and (iii) actuality-dependent (p. 126). These three conditions specify the paradigm term's value structure. "Is water," for example, is a paradigm term with something like the following value structure: <is the same liquid as, the liquid in a given sample, in the actual world>. Paradigm terms plausibly extend beyond natural kind terms, for they depend more on the

18 See Dupré (1981) for an influential criticism.

way a predicate is introduced than on the nature of the thing for which it is introduced.

Sally Haslanger proposes a view very similar to Nimtz'. Haslanger argues that "the basic strategy of natural kind externalism need not be confined to natural kinds;" instead, "[e]xternalism is an option whenever there are relatively objective types," where "objectivity is not only to be found in the natural world" (Haslanger, 2006, p. 109). This leads Haslanger to adopt:

> Objective type externalism: Terms/concepts pick out an objective type, whether or not we can state conditions for membership in the type, by virtue of the fact that their meaning is determined by ostension of paradigms (or other means of reference-fixing) together with an implicit extension to things of the same type as the paradigms. (ibid.)

The conclusion to be drawn from this discussion is this: As Kripke argued himself, natural kind terms can be introduced and passed along in pretty much the same way as proper names; and as demonstrated by Nimtz and Haslanger, the same reasoning extends also to the broader class of paradigm terms or terms for objective types respectively. So, to the extent that proper names which appear in texts generated by to-date LLMs refer, the same is true of the much wider class of paradigm terms (or terms for objective types, as Haslanger would put it). This is a very significant extension of the argument presented in the previous sections.

7. Conclusion

The recent success of LLMs raises difficult questions about whether our tendency to take their output at face value can be trusted. Some scholars warn that we should be cautious about attributing meaning and reference to LLMs. Because LLMs lack any contact with the real world, these scholars argue, they cannot fully grasp what natural language expressions mean or refer to. Some even go so far as to call LLMs mere "babbling stochastic parrots" that may mimic real language use without actually mastering it. While I agree that the capacity of an LLM to acquire reference should not be taken for granted but thoroughly investigated, I have argued here for a more optimistic position, at least with respect to reference. By the lights of classical externalist approaches to reference, at least the proper names and the paradigm terms that feature in LLM-generated texts *do* refer. The key insight here is that, just as with human language users, reference can be transmitted from an LLM's training data to its later uses of the expression in question.[19]

8. References

Ball, D. (2020). Relativism, metasemantics, and the future. *Inquiry*, *63*(9–10), 1036–1086. https://doi.org/10.1080/0020174X.2020.1805710

Ball, D. (2024). *Definition and dispute: a defense of temporal externalism*. Oxford University Press. https://doi.org/10.1093/oso/9780198906186.001.0001

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2303.12712

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: philosophical foundations*. Oxford University Press.

19 With thanks to Christian Nimtz, Jakob Ohlhorst, Andrea Raimondi, and the editors and reviewers of this journal for their valuable input. My work on this paper was generously supported by a grant from the DFG, grant number 533861171.

Chalmers, D. J. (2023). Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models. *Proceedings and Addresses of the American Philosophical Association*, *97*, 22–45.
Devitt, M. (1981). *Designation*. Columbia University Press. https://doi.org/10.7312/devi90836
Devitt, M. (2006). Responses to the Rijeka Papers. *Croatian Journal of Philosophy*, *6*(1), 97–112.
Devitt, M. (2008). Reference borrowing: A response to Dunja Jutronić. *Croatian Journal of Philosophy*, *8*, 391–366.
Devitt, M. (2015). Should Proper Names Still Seem So Problematic? In A. Bianchi (Ed.), *On Reference* (pp. 108–144). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198714088.003.0007
Dupré, J. (1981). Natural Kinds and Biological Taxa. *The Philosophical Review*, *90*(1), 66–90. https://doi.org/10.2307/2184373
Evans, G. (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume*, *47*(1), 187–225. https://doi.org/10.1093/aristoteliansupp/47.1.187
Evans, G. (1982). *The varieties of reference* (J. H. McDowell, Ed.; Repr). Clarendon Press.
Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift Für Philosophie Und Philosophische Kritik*, *100*(1), 25–50.
Gubelmann, R. (2024). Large Language Models, Agency, and Why Speech Acts are Beyond Them (For Now) – A Kantian-Cum-Pragmatist Case. *Philosophy & Technology*, *37*(1), 32. https://doi.org/10.1007/s13347-024-00696-1
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6
Haslanger, S. (2006). *I — Sally Haslanger*: What Good are Our Intuitions? *Aristotelian Society Supplementary Volume*, *80*(1), 89–118. https://doi.org/10.1111/j.1467-8349.2006.00139.x
Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9
Jackman, H. (1999). We Live Forwards But Understand Backwards: Linguistic Practices and Future Behavior. *Pacific Philosophical Quarterly*, *80*(2), 157–177. https://doi.org/10.1111/1468-0114.00078
Jackman, H. (2005). Temporal externalism, deference, and our ordinary linguistic practice. *Pacific Philosophical Quarterly*, *86*(3), 365–380. https://doi.org/10.1111/j.1468-0114.2005.00232.x
Jackman, H. (2020). Temporal externalism, conceptual continuity, meaning, and use. *Inquiry*, *63*(9–10), 959–973. https://doi.org/10.1080/0020174X.2020.1805706
Jackson, F. (1998). Reference and Description Revisited. *Noûs*, *32*(S12), 201–218. https://doi.org/10.1111/0029-4624.32.s12.9
Kipper, J., & Soysal, Z. (2022). A Kripkean argument for descriptivism. *Noûs*, *56*(3), 654–669. https://doi.org/10.1111/nous.12378
Koch, S., & Wiegmann, A. (2022). Folk Intuitions about Reference Change and the Causal Theory of Reference. *Ergo*, *8*, 25. https://doi.org/10.3998/ergo.2226
Kripke, S. (1977). Speaker's Reference and Semantic Reference. *Midwest Studies in Philosophy*, *2*, 255–276. https://doi.org/10.1111/j.1475-4975.1977.tb00045.x
Kripke, S. (1980). *Naming and Necessity*. Blackwell Publishers.
Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review*, *130*(2), 401–431. https://doi.org/10.1037/rev0000297
Mallory, F. (2023). Fictionalism about Chatbots. *Ergo*, *10*, 38. https://doi.org/10.3998/ergo.4668
Mandelkern, M., & Linzen, T. (2024). Do Language Models' Words Refer? *Computational Linguistics*, *50*(3), 1191–1200. https://doi.org/10.1162/coli_a_00522
Michaelson, E. (2023). The Vagaries of References. *Ergo*, *9*. https://doi.org/10.3998/ergo.3115
Millière, R., & Buckner, C. (2024). *A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2401.03910
Mollo, D. C., & Millière, R. (2023). *The Vector Grounding Problem* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2304.01481
Nimtz, C. (2017). Paradigm Terms: The Necessity of Kind Term Identifications Generalized. *Australasian Journal of Philosophy*, *95*(1), 124–140. https://doi.org/10.1080/00048402.2016.1155226
OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu,

P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2303.08774

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* (arXiv:2208.02957). arXiv. https://doi.org/10.48550/arXiv.2208.02957

Raatikainen, P. (2020). Theories of reference: What was the question? In A. Bianchi (Ed.), *Language and Reality From a Naturalistic Perspective: Themes from Michael Devitt* (pp. 69–103). Springer.

Russell, B. (1905). II.—On Denoting. *Mind*, *14*(56), 479–493.

Savelka, J., Agarwal, A., An, M., Bogart, C., & Sakr, M. (2023). *Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses*. https://doi.org/10.48550/ARXIV.2306.10073

Searle, J. R. (1958). II.—PROPER NAMES. *Mind*, *LXVII*(266), 166–173. https://doi.org/10.1093/mind/ LXVII.266.166

Van Woudenberg, R., Ranalli, C., & Bracker, D. (2024). Authorship and ChatGPT: a Conservative View. *Philosophy & Technology*, *37*(1), 34. https://doi.org/10.1007/s13347-024-00715-1

Yourgrau, P. (2012). Kripke's Frege: Kripke's Frege. *Thought: A Journal of Philosophy*, *1*(2), 100 – 107. https://doi.org/10.1002/tht3.15

Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2023). *Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.07921