Babbling stochastic parrots? On reference and reference change in large language models

Steffen Koch
Department of Philosophy
Bielefeld University
Steffen.koch@uni-bielefeld.de

Abstract

Recently developed large language models (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 LLMs are very likely capable of reference. Not only that: The combination of a popular externalist account of reference and recent experimental data in machine psychology even suggests that LLMs might play a role in shifting what our words refer to.

I.

Large language models (LLMs) such as BERT, GPT-3/4 or Switch-C play an increasingly important role in many areas of our professional and personal lives. These machine learning systems are trained on large corpora taken from the Internet to predict the probability of a token (e.g. a word) based on its preceding or surrounding context (in bidirectional and masked LLMs). LLMs take words as stimuli and output new texts according to the statistical distributions collected from their training data. The texts produced by LLMs have all the characteristics of meaningful sentences. In most contexts, these texts are interpreted in the same way as human speech or text.

Recent work in the philosophy and psychology of AI has begun to question the validity of this impression. Bender et al. (2021) 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" (611). According to them, LLMs are merely babbling "stochastic parrots" that produce the image of human language but in fact fall short of meaningful communication. Similarly, 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. While granting that LLMs are surprisingly successful in many language-involving tasks, Lake and Murphy (2023) also argue that there is a principled limit to their ability to build the knowledge structures that form part of the basis of word meanings.

When debating LLMs' capacity for thought and talk, a number of different philosophical questions need to be distinguished. It is one thing to attribute to LLMs the capacity to *produce* meaningful texts; it is quite another to attribute to them the capacity to *understand* such texts. Moreover, language use has many different facets that must be kept separate. Producing meaningful texts does not yet mean that one has mastered pragmatics, is able to perform all sorts of different speech acts, can construct metaphors, or use language in creative ways.

Here I'll be concerned with a rather narrow pair of questions: Do the words and sentences produced by to-date LLMs refer to things, and if so, how does this feed back into the language that is spoken by human agents?

Reference is an important aspect of language use. We can think of it as the glue between linguistic signs, on the one hand, and the objects, events, and relations they stand for, on the other. What our words refer to determines what we talk about. If our language lacked reference entirely, we wouldn't be talking about anything at all. Getting clear about whether and how LLMs are able to refer informs how we should interpret their outputs.

A prominent argument is that LLMs trained purely on form are fundamentally incapable of acquiring meaning and reference (Lake and Murphy 2023; Bender and Koller 2020). Bender and Koller (2020) illustrate this point with 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 won't be able to recognize a coconut when it sees one. But, according to the authors, the "ability to connect [...] utterances to the world" (5188) is a necessary condition for learning meaning. They conclude that the octopus does not know the meaning of "coconut". 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.

This argument assumes that if one cannot identify the referent of an expression, then one does not know its meaning. But as Piantadosi and Hill (2023) rightly point out, this assumption is problematic. For one thing, there is meaning without a referent ("the present king of France"). For another, one can know the meaning of an expression without knowing what it refers to ("the largest living ant"). Bender and Koller's "octopus test" does not show that LLMs are fundamentally incapable of acquiring meanings.

As I will show in the following pages, more plausible approaches to meaning and reference suggest a more optimistic assessment of LLM reference. The approaches I have in mind here are semantic externalist views of meaning and reference that have been developed in the philosophy of language over the last four to five decades. Following these approaches reveals why achieving reference is less demanding than many authors assume and why the octopus test is not an adequate test for reference.[1]

Somewhat surprisingly, externalist approaches to meaning and reference - now standard in the philosophy of language - have so far received only sporadic treatment in the debate over meaning and content in AI (and LLM in particular).[2] This is a waste of potential, because, as Cappelen and Dever (2021) note, semantic externalism has been most successful for "'neural network' creatures" like us – so why shouldn't we expect it to be successful for neural networks

[1] This being said, it might still be a valid test for other language-involving skills, such as understanding.

[2] Two recent exceptions: Cappelen and Dever’s (2021) book-length treatment of content in AI and a recently uploaded preprint by Mandelkern and Linzen (2023).

and other programs we've created? (66). In addition to arguing for a more optimistic position on LLM reference, another goal of this paper is to make semantic externalism more visible in debates about AI language.

I am not the first author to take an externalist approach to reference in LLMs. Mandelkern and Linzen (2023) also use insights from the externalist tradition to argue that LLMs, like humans, can inherit reference from their training data. However, I will develop the 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. Pace Mallory (2023), I will argue that this is not so. I also consider a possible implication of the resulting view: that LLMs are not mere "parrots" that replicate human language use, but can have a lasting effect on human language through their ability to induce reference change.

The position I advocate here has implications for the much-discussed symbol grounding problem (Harnard 1990). In the context of LLMs, the problem is how a system whose input is limited to data about the co-occurrence of symbols can give meaning to those symbols. While some argue that word meanings (and concepts) must be causally grounded in an extra-linguistic reality through perception, action, or desire (Harnard 1990; Lake and Murphy 2023), and that this stands in the way of an LLM acquiring meaning, others argue that there can be meaningful thought without sensory grounding (Chalmers 2023). The argument I develop here suggests a third way, at least with respect to reference. Reference requires causal grounding, but, despite appearances to the contrary, an LLM's use of a word *is* grounded in extra-linguistic reality through causal-historical chains.

I proceed as follows. In Section II, I use arguments developed by Saul Kripke to show that the strings produced by LLMs are likely to refer to objects in the world. Here I also discuss the issue of whether reference requires referential intentions that preclude LLM reference. In Section III, I raise a complication to the Kripke-inspired view based on observations by Gareth Evans and Michael Devitt. I argue that their suggested fixes to Kripke’s theory works for LLMs as well. In section IV, I relate the implications of the resulting view to new data in the field of "machine psychology," arguing that LLMs may actively contribute to reference change rather than merely following human language use.

II.

How can we approach questions of reference in LLMs? One approach would be to start from the observation that chatbots like Chat-GPT4 clearly give the impression of making statements about things in the world, and then to seek or construct a philosophical theory that predicts and explains the validity of this impression.[3] Another approach would be to treat it as an open question whether LLMs refer. A promising route to this approach would be to consider the conditions under which humans refer, and then to check whether machines like LLMs also satisfy these conditions. This paper follows the latter approach.

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

[3] This approach is endorsed by Cappelen and Dever (2021) with respect to content in AI systems.

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 is part of 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).

Versions of this view were held by Gottlob Frege (1982) and Bertrand Russell (1905). Put a bit more formally, these philosophers believed that for every proper name N, there is an entity e and a property F_N, such that: (i) only e is F_N, and (ii) competent users of N know that only e is F_N. This allowed them to say that being F_N is the meaning (or sense) of N and that the referent of N is the unique object that is F_N.[4] 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.

A view along these lines 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 sider of the speaker, and it is difficult to see how a mindless LLM could meet this requirement. This problem might not be too severe, however, for it might well be possible to construe knowledge in functionalist terms as the statistical data that the LLM associates with individual expressions. A more severe problem is that the LLM has no way of distinguishing which elements of these data belong to the reference-determining information about "Obama" (the "analytic truths", as philosophers say) and which ones do 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? Why? Why not?

There are, however, independent reasons for rejecting the view that reference is established by identifying information that one associates with a word. One way to see this is to ask yourself how *you* would go about deciding which piece of information about Obama is reference-determining and which not. But the clearest case can be made by a thought experiment like the following: 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."[5]

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 was not born in the US. She might ask others if it is true that Barack Obama was not born in the US. Or

[4] John Searle (1958) 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.

[5] A "birther" is a conspiracy theorist who believes that Barack Obama was not born in the US and could therefore not have been elected for president.

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. If she were to say, "Barack Obama was not born in the US," she would be making a false statement about the actual person Obama.

Note that Martha can do these things without having any experience with Obama and without having the slightest idea who he is. The only thing she has heard about Obama - that he was not born in the US - is false. And even if it were true, it would hardly distinguish Obama from many other people. About 7 billion people are not born in the US! Martha has no way of identifying Obama when she sees him. So it seems that successful reference can hardly be a matter of identifying knowledge or experience associated with the thing in question. What else provides the glue between a word and the thing(s) it refers to?

Here is, in rough outline, the story as presented by Saul Kripke in *Naming and Necessity*:[6]

> A baby is born, and his parents name him "Barack Obama." They tell their friends about him. Through various types of conversation, the name is passed from link to link as if by a chain. A speaker at the other end of this chain, who has heard about Barack Obama in the marketplace or elsewhere, may be referring to Barack Obama even though she can't remember who first told her about Obama or who ever told her about Obama. A certain passage of communication, which ultimately reaches the man himself, reaches the speaker. She then refers to Obama, even though she can't clearly identify him. She doesn't know when Obama was president, she doesn't know what Obama stood for or what party he belonged to. She doesn't need to know these things, but instead a chain of communication has been established that leads back to Obama himself, by virtue of her membership in a community that has passed the name from link to link.

This sketch is far from a full-blown theory, but it nicely explains why Martha is able to use the name "Obama" to refer to the actual person, Obama.[7] Martha does not know anything about Obama, and may not even know from whom she first heard of Obama. But she is at the end of a chain of communication that ends with a first use of "Barack Obama" by someone who had direct contact with Obama, most likely his parents. That is all it takes for Martha to refer to Obama by the name "Obama."

If such a view is plausible for human language users, the prospects for the LLM's ability to refer to things look better. Of course, an LLM has never had direct contact with Barack Obama. Nor are LLMs like human language users who simply pick up names in ordinary conversation. But LLMs are trained on large corpora. These corpora are the products of human language users – people who are part of their particular language community, and whose use 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 is sufficient for them to "inherit" the reference relations between the words they use and the objects they denote.

[6] The following is a slightly adapted passage from p. 91.

[7] See e.g. Salmon (1986), Soames (2002), Nimtz (2019), and Cappelen and Dever (2018) for further developments of this view.

This view also allows a direct response to the symbol grounding problem, at least with respect to reference. It may well be that successful reference requires causal grounding in some extra-linguistic reality. But the above view shows how even machines that have no direct contact with the outside world can acquire causal grounding. Kripke's crucial insight is that causal grounding can be mediated by possibly long and far-fetched chains of communication. As long as the expressions in the LLM's training data are causally grounded (even if again mediated), so are the expressions used by the LLM.

Let us go back to Martha and consider a complication to the present view. Suppose Martha becomes more and more fascinated with Obama until she decides to name her new dog after him. When Martha now uses the name "Barack Obama" to refer to her dog, she is not talking about Obama, but about her dog. This suggests that something more is required to ensure that a particular use of "Barack Obama" refers to Obama. Aware of this potential complication, Kripke introduced the additional requirement that the speaker must intend to use the name in the same way as the person from whom she learned it (whoever that may be). With this requirement, this view is able to explain why Martha does not refer to Obama: Martha does not intend to use the name in the same way as the person from whom she learned it.

This additional requirement raises a potential problem for an LLM's ability to refer to things. For although LLMs can be trained on data that has causal relations to things in the world, an LLM's use of a word is not accompanied by any referential intentions. As Mallroy (2023) puts it, "[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" (1084). If intentions are necessary for successful reference, this might imply that LLMs are incapable of talking about objects in the real world.

But whether this complication gets in the way of an LLM's ability to refer depends on how one reads the requirement of having the right referential intentions. On a strong reading, this means 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. Assuming that LLMs cannot have any intentions, this reading would indeed get in the way of LLM reference. But note that such a requirement is quite implausible, *even for human language users*. In most cases, we use names rather automatically, without explicitly forming the intention to use them in this or that way.

A weaker but more realistic reading would be this: if someone picks up the name "Obama" from people who have used it to refer to Obama, and she does not have the intention of using "Obama" to refer to some other person or object, then she refers to Obama as well. The difference with the previous reading is that this requirement is satisfied in the *absence* of conflicting intentions, rather than in the *presence* of consonant intentions. This reading yields the right verdict when we apply it to the dog case. When Martha first uses "Obama" to refer to her dog, that use is accompanied by her intention to use the name to refer to her dog rather than to the person Obama.

This reading of Kripke's additional requirement brings LLMs back into the picture. The requirement that there be no conflicting intentions is consistent with the fact that LLMs cannot have any intentions. So, pace Mallroy (2023), it seems that LLMs satisfy all the externalist's necessary conditions for using names to refer to objects.

III.

Unfortunately, however, Kripke's view as presented above cannot be quite correct. To see this, we need to consider cases of reference change. Changes of meaning and reference occur frequently in all natural languages. The English word "girl" once referred to children of any sex, "meat" once referred to all kinds of solid foods (including vegetarian ones), and "a myriad" once meant ten thousand. Sometimes such changes are intentional. But often they are not, or at least not obviously so.

Unintentional reference changes can also occur with proper names. For example, the name "Madagascar" once referred to parts of the African mainland, but now refers to the island in southeastern Africa. How did this happen? It is hard to say for sure, but according to Isaac Taylor (*Names and their History*, 1898) it has to do with Marco Polo misunderstanding a hearsay report from sailors. While these sailors were actually talking about parts of the African mainland, Marco Polo thought they were talking about the island. When he returned to Europe, he spread this mistaken use of "Madagascar" until it became the dominant usage. Today, it is clear to almost everyone that "Madagascar" refers to the island, not the mainland. It seems that the reference of "Madagascar" has shifted from the mainland to the island, although we can assume that no one ever intended to use "Madagascar" differently than before.

A similar twist could be added to the example of Martha and Obama. Suppose that after Martha returns to the US and hears about Obama, she confuses him with a famous composer, say Claude Debussy. Suppose further that she then goes back to the indigenous tribe she used to live with and tells them all about Obama, the famous composer: the sound of his music, his accomplishments, etc. It now becomes common practice in this community to use "Obama" to refer to Debussy. There is pressure to believe that, at least after some time, the name "Obama" when used by a member of this community actually refers to Debussy rather than Obama. And yet, neither Martha nor anyone else had any intention of using the name "Obama" differently than those before them.

Whether realistic or fanciful, these "slow switching cases" (Burge 1988; Boghossian 1989) are counterexamples to Kripke's version of the causal theory of reference. In Kripke's view, the reference of a term remains constant as long as it is passed from link to link, and no one has any intention of deviating from the way it has been used by those before them. But in the cases just discussed, there is pressure to think that reference changes even though these conditions are met. Kripke was aware of the problem, but noted that it needs "more apparatus than I have developed here [in *Naming and Necessity*]" (Kripke 1980, 163).

What went wrong? There are multiple possible answers to this question, some more Kripke friendly, some less so. According to Gareth Evans, Kripke's mistake is that he 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, 197). 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. 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, 131).

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 correct, 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." In the discussion of Kripke's causal theory above, we noted that if reference required referential intentions, this would preclude an LLM's ability to refer, since LLMs do not have mental states. The same reasoning applies here. 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.[8] 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 and Dever (2021), whether the Evans/Devitt view can be "de-anthropocentrized" via "anthropocentric abstraction":

> "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." (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

[8] See Cappelen and Dever (2021), p. 111f. for how such a view can be applied to other types of AI systems.

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 from training on human language corpora. 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-antropocentrized 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 many questions about details open. We have here been entirely concerned with proper names; more needs to be said about how to apply their views also to certain predicates. Nevertheless, it seems that we have good reason to believe that these views can be modified in an LLM-friendly way while remaining true to their spirit. On the assumption that a view along these lines is roughly correct, we can tentatively conclude that LLMs are, after all, capable of using names (and certain predicates) to refer to things in the world.

IV.

So far we have been concerned with the question of whether LLM-based chatbots are capable of using language to refer to things at all. A key insight from the externalist tradition is that speakers need little or no mental states to do so; it is sufficient that they are part of a linguistic community in which communicative chains relate one to the object in question. Since LLMs are trained on human speech data, they seem to meet this requirement. In this section, I want to focus on the implications that this has or could have for other members of the respective linguistic community - most importantly, for us human language users.

Kripke’s view has the comforting consequence that LLM reference will forever be in line with human reference. Now that we have rejected Kripke’s view, we must face the uncomfortable truth that this is not so. Systematically misapplying a term, or at least applying it differently than before, may induce reference change.

How exactly? This is a bit complicated to spell out, but here is a rough outline.[9] When an LLM uses a referential term differently, over time that term becomes associated with new and different pieces of information – information that comes from different sources. If a large or particularly salient part of the total body of information that speakers associate with a term is based on a different source, the reference will gradually shift from

[9] For more detailed expositions see e.g. Evans (1982,), pp. 388-390; Dickie (2015), ch. 5.2; and Koch and Wiegmann (2022), pp. 36-37.

the original source to the new source. In the case of Marco Polo, this process led to a situation in which the island we now refer to as "Madagascar" became the dominant causal source of the information that speakers associated with "Madagascar". In the case of LLMs, this effect would plausibly be accelerated if the outputs were fed back into the training data. So, if LLMs use particular words in a way that systematically deviates from how humans use it, this might affect what our words refer to.

But do they? It is too early to say for sure. Investigating whether and how specific LLMs show significant divergence in their use of language compared to human agents requires rigorous testing in different domains of judgment, using different LLMs and different testing methods. This research is now a hot topic in the emerging field of "machine psychology" (Hagendorff 2023), which seeks to uncover the inner workings of AI systems, since neither the machines themselves nor their creators can provide adequate insight into them. Surprisingly, however, preliminary data suggests that one of today's most popular LLMs, GPT-4, actually exhibits response patterns that are remarkably different from those of human language users.

Almeida et al. (2023) compared GPT-4 reasoning in moral and legal domains with patterns found in research on human agents. While GPT-4 exhibited many of the biases found in human moral and legal judgments, there were also differences. The authors summarize them as follows:

> In our studies, GPT-4, when compared to humans, among other things: (i) was more likely to ascribe intentionality to morally bad side effects, (ii) was less likely to ascribe intentionality to morally good side effects, (iii) showed a different pattern with regards to the effects of abnormality over causation judgments in disjunctive structures, (iv) was more likely to match the anticipated deontic status of deception, (v) distributed the importance of moral foundations differently, (vi) was more likely to attribute consent in deception cases, (vii) showed more extreme effects of the hindsight bias, and (viii) relied more heavily on text.

We do not need to elaborate on each of these points to see the big picture. One might initially expect an artificial language user like GPT-4 to be more "rational", i.e. less susceptible to bias, than humans. But in some areas the opposite seems to be true. GPT-4 showed a stronger Knobe effect and more hindsight bias than humans. The authors conclude that “GPT-4’s cognition might systematically deviate from that of human beings in ways that cannot be reduced to the “correct answers” effect”.

Assuming, for the sake of argument, that these data are indeed indicative of systematic "cognitive" differences between GPT-4 and humans, what implications would this have for reference in LLMs? An initial, but ultimately false, reason for thinking that it has no interesting implications is this: GPT-4 reaches different conclusions about what is morally permissible, who is causally responsible for what, gives different consent ratings, etc. This suggests, at most, that it has different patterns of reasoning or "opinions," not that it uses language differently. Compare: You and I may have different ways of estimating who will win the next presidential election, and we may come to different conclusions about it. But that does not mean that we use the word "president" or "election" differently.

In human agents, there is a meaningful difference between semantic knowledge on the one hand and world knowledge on the other. Different human agents may diverge in what they believe about the world, but converge in how they use the words necessary to formulate their beliefs. This difference disappears with LLMs. LLMs generate response patterns based on statistical data about language use. All of their "knowledge" is syntactic knowledge, generated from the analysis of linguistic patterns. Depending on one's view of meaning, one might want to count this as semantic knowledge or not. But however we want to call it, there is no meaningful difference between this knowledge and world knowledge. This means that whenever an LLM shows systematically different response patterns than humans, this is a difference in language use.

The data presented above involve rather abstract notions such as moral permissibility or intentionality. It is not yet clear whether there are systematic differences in the way GPT-4 or other LLMs use referential terms such as proper names or kind terms. But the fact that there are systematic differences in other domains should make us cautious about this possibility. Given how much the output of chatbots based on LLMs, such as Chat-GPT, is already relied upon in many domains, this could have a lasting effect on the linguistic communities to which they belong. Just as Marco Polo did when he came to Europe and spread his incorrect use of "Madagascar," the use of LLMs could contribute to reference shifts for an unknown number of referential terms in English and other languages.

V.

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 meaning and 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, LLMs *do* refer. This is true despite the fact that some of the authors working in the externalist tradition use mentalist vocabulary to formulate their views. In the case of Kripke's causal theory of reference, I have argued that it is the absence of conflicting intentions rather than the presence of conforming intentions that ensures the preservation of reference; and in the case of Evans and Devitt, the crucial mentalist terms – information and thought – can be construed in functionalist terms via anthropocentric abstraction.

I have also argued that, just as Marco Polo allegedly did, an LLM that uses words in a systematically different way than human language users might contribute to reference change by shifting the causal source of the information that agents (artificial and otherwise) associate with the term. Recent experimental work suggests that this may indeed be the case with Chat-GPT4, but more evidence is needed to say for sure.

So are LLMs babbling stochastic parrots? That depends. If you think that parrots are incapable of using words to talk about objects, properties, or relations in the world, then the answer is no. But in light of the externalist arguments rehearsed above, you might be willing to consider the possibility that parrots, too, can refer. If this is true, then LLMs may indeed be babbling stochastic parrots - but so what?

Conflict of interest

I have no conflict of interest to declare.

Data availability statement

I do not analyse or generate any datasets, because my work proceeds within a theoretical approach.

References

Almeida, G. F. C. F., Nunes, J. N., Engelmann, N., Wiegmann, A. and de Araújo, M. (2023). Exploring the psychology of GPT4's Moral and Legal Reasoning. arXiv:2308.01264

Bender, E. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bender, E., Gebru, T., McMillan-Major, A. & Smitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21, March 3–10, 2021, Virtual Event, Canada. Doi: https://doi.org/10.1145/3442188.3445922

Burge, T. (1979). Individualism and the Mental, *Midwest Studies in Philosophy* 4, 73-122.

Burge, T. (1988). Individualism and Self-Knowledge, *Journal of Philosophy* 85, 649-63.

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

Cappelen, H. and Dever, J. (2018). *Puzzles Of Reference*. Oxford: Oxford University Press.

Chalmers, D. J. (2023). Does thought require sensory grounding? *Proceedings and Addresses of the American Philosophical Association* 97, 22-45

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

Dickie, I. (2015). *Fixing Reference*. Oxford University Press.

Evans, G. (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume*, 47/1: 187–225.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.

Frege, G. (1892). On Sense and Reference [Über Sinn und Bedeutung], *Zeitschrift für Philosophie und philosophische Kritik* 100.. 25–50.

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. DOI: https://doi.org/10.48550/ARXIV.2303.13988.

Koch, S. & Wiegmann, A. (2022). Folk Intuitions about Reference Change and the Causal Theory of Reference, *Ergo* 8: 25. doi: https://doi.org/10.3998/ergo.2226

Kripke, S. A. (1980). *Naming and Necessity*. Oxford: Basil Blackwell Ltd.

Lake, B. M. and Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review* 130 (2), 401-431.

Mallory, F. (2023). Finctionalism about Chatbots. *Ergo*. https://doi.org/10.3998/ergo.4668

Mandelkern, M. and Linzen, T. (2023). Do language models refer? arXiv:2308.05576v1

Nimtz, C. (2019). Kripkean Meta-Semantics and Generalized Rigidity. *Philosophical Quarterly* 69 (275): 332-353.

Piantadosi, S. Z. and Hill, F. (2023) Meaning without reference in large language models. arXiv:2208.02957v2

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

Salmon, Nathan U. (1986). *Frege's Puzzle*. Ridgeview.

Searle, J. (1958). Proper Names. *Mind* 67 (166): 166-173.

Soames, Scott (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.