Language Models and Externalism: A Reply to Mandelkern and Linzen

Gary Ostertag

Icahn School of Medicine at Mount Sinai & The Graduate Center, CUNY
`gary.ostertag@mssm.edu`

Do texts generated by language models (LMs) refer? Mandelkern and Linzen (2024) argue that externalist principles point to an affirmative conclusion. What grounds reference, according to their externalism, is a term's "natural history". For example, 'water' refers to H_2O among English speakers, and not to the phenomenally indistinguishable chemical XYZ, because H_2O, and not XYZ, is implicated in the natural history of 'water'. Appealing to the literature on contrastive explanation, I show that a term's natural history does not generally ground its referential properties. Thus, Mandelkern and Linzen's quick route to the referentiality of LM-generated texts fails.

Do texts generated by language models (LMs) refer? There seems ample reason to be skeptical. Since their inputs "are only strings of symbols," it would seem that LMs "cannot produce referential words: *reference* cannot be derived from *form*." While the quotation is from Mandelkern and Linzen (2024), they go on to suggest that this pessimistic attitude is mistaken. In their view, the externalist tradition emanating from Kripke (1980) and Putnam (1975) provides the resources to challenge the "very tempting argument" just described. As they argue, externalism, with its emphasis on the natural history of a symbol, opens up the possibility that an LM-generated text can indeed refer, since it may lay claim to the same sort of natural history. Moreover, if LM-generated texts can be shown to refer, then we can be reasonably optimistic concerning the question whether such texts are meaningful.[1] (If they cannot, then a large chunk of generated discourse—indeed almost all of what we care about—will be meaningless.) As I argue, however, the natural history of a term, in the austere sense assumed by Mandelkern and Linzen, cannot ground the reference relation. Since they give clear expression to a growing consensus among philosophers of language engaging in these questions, it is worth considering Mandelkern and Linzen's central argument in detail, especially given that the discussion has come to the attention of researchers in AI and natural-language processing.

---

[1] See Chalmers (2023). But see also Bender et al. (2021), Ostertag (2023), and Titus (2024) for reasons to remain skeptical.

1. Do Facts About Natural History Ground Reference?

Since Kripke (1980), it has seemed to many that some form of externalism about reference is inevitable. The fact that I refer to Nixon in uttering *Nixon resigned in disgrace* is due to causal-historical facts involving that name. In particular, reference is mediated by a causal-historical chain linking the utterance to an event in which the bearer, Nixon, was "baptized" with that name. Since a name's entering my vocabulary "does not require substantive experience, beliefs, or capacities vis-à-vis its referent" (Mandelkern and Linzen 2024, page 1194), the question is raised whether an LM's tokening of the above sentence refers to Nixon. One line of resistance to a positive answer stems from the claim, cited above, that "reference cannot be derived from form." But this oversimplifies the dialectic. As Mandelkern and Linzen point out, "LMs' inputs are not bare strings of symbols, but strings of symbols with certain natural histories that connect them to their referents" (page 1191). Once this fact is taken into account, they argue, we have an effective response to the skeptic about LM reference.

But this correction involves an oversimplification of its own: Nothing in Kripke's understanding of the natural history of a name or kind term allows for such optimism. Indeed, Kripke's picture would seem to preclude LMs' participating in the sort of name-using practices required for their outputs to refer.[2] Kripke himself, although he occasionally spoke of a "causal or historical connection" between current use and baptismal event, was careful to avoid causal vocabulary when describing how reference is transmitted. As he wrote:

> 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. (Kripke 1980, page 96)

Consider the practice of using the name *Nixon* to refer to the former president. On Kripke's view, what links the referential acts constituting this practice is the speaker's intention, at each point, to preserve the reference of the person from whom they first learned the name. (The precise construal of the intention need not detain us here.) This suggests that what grounds the fact that, e.g., my current use of *Nixon* refers to Nixon is that (*a*) the person from whom I learned the name used it to refer to Nixon and that (*b*) *I intend my current use to co-refer with that use*. Referential intentions, however they are ultimately construed, are thus an ineliminable part of what grounds proper-name (and natural-kind-term) reference.

This does not automatically imply a negative answer to our question. One general strategy of response is to show that LMs can, in fact, be said to intend to refer in conformity with communal practice. While Mandelkern and Linzen are cautiously optimistic about the prospects for such an approach, they acknowledge that it is an open question whether LMs can literally be said to possess intentions of the relevant sort. Alternatively, one might part with Kripke here and deny that an attitude of deference to the community is integral to our referential practices (Cappelen and Deutsch 2024). If either view turns out to be correct, then nothing should stand in the way of our saying that LM-generated tokens refer.

---

2 For the moment setting aside the thorny question of what it could even mean for an AI tool to engage in "a practice". See, however, Cappelen and Dever (2021, Chapter 5), who attempt to provide a "de-anthropocentrized" natural history, one applicable to AI tools.

But this is a big *if*. An alternative path to an affirmative answer is more direct, dispensing with the need to route reference through states like intending or believing. Mandelkern and Linzen provide a succinct statement of the approach:

> What grounds reference is not beliefs, experiences, and discriminatory capacities of the kind that LMs obviously lack. Rather, what grounds reference is the *natural histories* of words: the causal-historical links between a speech community's use of a word and the word's referent. (Mandelkern and Linzen 2024, page 1194; note omitted)

Let's call this view—that the natural history of a name grounds its reference—**naïve externalism**. On such a view, it should be "strikingly easy for a speaker to use a word referentially" (Mandelkern and Linzen 2024, page 1194). To demonstrate this, they present the example of Luke, who assertively utters *Peano proved that arithmetic is incomplete*, based entirely on his reading an article whose sole content is that very sentence. As long as his use of *Peano* is appropriately situated vis-à-vis Peano, Luke can use that name to refer—even though the information on which that usage is based is false.

Another case they use to show the irrelevance of specific beliefs to reference is Putnam's well-known Twin Earth example (Putnam 1975). Recall that Twin Earth is just like Earth except for the fact that the transparent liquid that flows in its rivers and streams and is rapidly seeping from its glaciers and icecaps is not H_2O but something else—XYZ. Since we are supposing that Twin Earth is superficially just like Earth, it follows that Twin Earthians go around using *water* to refer to the liquid that flows in its rivers and streams, etc., just as we use *water* to refer to the liquid that flows in *our* rivers and streams, etc. The difference is that the liquid flowing in their rivers and streams is not what flows in our rivers and streams. Because of this, when they use *water* their tokens refer to their local liquid (XYZ) whereas when we use that term, our tokens refer to *our* local liquid (H_2O).

> What makes it the case that *water* in Earthian English refers to H_2O, while *water* in Twin Earthian refers to XYZ? It is *not* the beliefs of the speakers about the liquids' respective chemical compositions, since, by assumption, they do not yet know the chemical compositions of their respective watery stuffs... Rather, it seems to simply be the fact that the stuff in the natural history of the word *water* on earth was made of H_2O, whereas the stuff in the natural history of the of the word *water* on twin earth was made of XYZ. (Mandelkern and Linzen 2024, page 1195)

For Mandelkern and Linzen, all that is required for *water* to refer to H_2O (XYZ) among Earthians (Twin Earthians) is for the term to have the appropriate natural history. Similarly, all that is required for Luke's use of *Peano* to refer to the mathematician is for it to have the right causal-historical pedigree.

2. *Why?* Questions and Constitutive Explanation

As indicated, while Mandelkern and Linzen do not explicitly endorse naïve externalism, they hold that it provides a promising route to a positive answer to our leading question—whether LMs' outputs refer.

> Even if the inputs to LMs are simply strings of text, they are strings of text with natural histories, *and that may suffice for LMs' words to refer*.

> By contrast [to Cappelen and Dever], we will argue that the standard externalist
> accounts are at least prima facie compelling with respect to LMs. (2024, pages 1191,
> 1193; emphasis added)

My claim in what follows is to show that their optimism is misplaced: Naïve externalism gives us no grounds for holding that LMs' outputs refer.

I begin by noting a crucial slip in their argument. To see this, we first need to get clear on the question Mandelkern and Linzen take the externalist to be answering, which, in turn, will require a brief excursion into contrastive explanation.

Here is their question, slightly modified to cohere with standard accounts of *Why?* questions (Mandelkern and Linzen 2024, page 1195).[3]

Q1.   Why does *water* refer to H_2O among Earthians, while *water* refers to XYZ among Twin Earthians?

The first thing to note is that Q1 is really the conjunction of two questions:

Q1a.   Why does *water* refer to H_2O among Earthians?

Q1b.   Why does *water* refer to XYZ among Twin Earthians?

I will focus on the naïve externalist's answer to Q1a, although the same problems will arise with respect to their answer to Q1b.

Why? questions are intimately linked to the explanations they seek: We conceive of explanations as answers to such questions. What the contrastivist adds is that *Why?* questions are always, either implicitly or explicitly, contrastive in nature, taking the form: *Why P as opposed to Q*? Accordingly, explanations cite a reason or cause both for the fact that *P* and for the fact that *not-Q*.

Consider, as an example, (1):

(1)   Why did John get paresis?

As indicated, the contrastivist holds that a question of the form *Why P?* is always incomplete; it is elliptical for a question of the form *Why P rather than Q*? In the case of (1), the contrasting proposition, or "foil", will be supplied by context. If we fill in some background, we can see how this works: Paresis can only arise in individuals with untreated syphilis, occurring, even then, in only one out of ten cases. In addition, in any given instance of untreated syphilis it is indeterminate whether or not the particular case will develop into paresis. Let's assume that all this is part of the common ground and, in addition, that John has paresis *and* that José and Maryan do not. Relative to this background, it is plausible to take (1) as asking (2):

(2)   Why did John, *as opposed to José and Maryan*, contract paresis?

---

3 I have also formulated the question so that it concerns the reference of *water* among Earthians rather than what *water* refers to in "Earthian English." Presumably, nothing about the natural history of *water* is relevant to what it refers to in the abstract entity we here refer to as "Earthian English."

An additional element in the contrastivist's account concerns the proper form an *answer* to such questions must take. It addresses the question: If *Why?* questions are invariably contrastive, what must their answers look like? Lipton's well-known account of explanation provides an answer:

> To explain why *P* rather than *Q,* we must cite a causal difference between *P* and *not-Q*, consisting of a cause of *P* and the absence of a corresponding event in the history of *not-Q*. (Lipton 2004, page 42)

Intuitively, the fact that John was the only one among the three with untreated syphilis explains why John was the only one among the three who developed paresis. And this meshes perfectly with Lipton's account, giving us a (partial) cause of John's developing paresis together with the absence of "a corresponding event" in the history leading up to José and Maryan's current state of non-paresis. Importantly, for our purposes, the explanation shows us the form that an answer to a *Why?* question should take.

If contrastivism is on the right track, then Q1 is elliptical for a contrastive question. Given the context, it can be seen to be elliptical for either of the following:

Q2.   Why does *water* refer to H_2O, and not to XYZ, among Earthians?

Q3.   Why does *water* refer to H_2O among Earthians, but not among Twin Earthians?

On Lipton's account, an answer to Q2 must cite a causal difference between the fact that *water* refers to H_2O among Earthians and the fact that *water* does not refer to XYZ among Earthians. (A similar requirement holds for Q3, which I will set aside here.)

With all this in place, we are now ready to ask: *Does the externalist succeed in actually grounding reference?* For present purposes, this becomes: *Does the externalist succeed in responding satisfactorily to Q2?*

Before proceeding, note that a correct answer to Q2—that is, a contrastive explanation—provides a "difference maker": something that grounds the fact that *water* refers to H_2O but which is absent from the fact that *water* does not refer to XYZ. It is not required that the difference maker provides a complete ground of the fact that water refers to H_2O, only that it provides a necessary condition (partial ground) of the fact. And this is precisely what the externalist is equipped to deliver. To see this, recall (2):

(2)   Why did John, *as opposed to José and Maryan*, contract paresis?

The answer is: *Because John had untreated syphilis, whereas José and Maryan did not.* Here, the property of having untreated syphilis is the difference maker: It is what John has (and part of what makes it the case that he developed paresis) and José and Maryan lack.

The externalist can provide a parallel response to Q2. Here's how: Assume that XYZ never occurs in this corner of the universe. (Putnam assumes merely that it is not "plentiful" on Earth.) Since H_2O, but not XYZ, is implicated in the natural history of *water* (as used by Earthians), we have an explanation as to why *water* refers to H_2O but does not refer to XYZ, and thus an answer to Q2: *Because H_2O is implicated in the natural history of 'water' whereas XYZ is not implicated in the natural history of 'water'.* Here,

the property of being implicated in the natural history of *water* is the difference maker: H_2O has it (and that's part of what makes it function as the referent of *water*), XYZ does not.[4]

But even if this serves to answer Q2, the strategy fails to yield an answer to Q4:

Q4.   Why does *water* refer to H_2O, and not to oxygen, among Earthians?

To explain why P rather than Q, we must cite a difference maker: a property that H_2O has but oxygen lacks. But, although H_2O is implicated in the causal history of *water*, and thus partially grounds that reference, oxygen is equally implicated in the causal history of *water*. We thus fail to meet "the difference condition" (Lipton).

Again, it might be thought that citing the fact that H_2O is implicated in the natural history of *water* would "stop the buck" in both Q2 and Q4; but it can do so only if this fact is also a difference maker. In Q2 it is: H_2O, but not XYZ, is implicated in the natural history of *water*. But not in Q4: The parallel "difference maker"—that H_2O, but not oxygen, is implicated in the natural history of *water*—is false.

3. Fine-tuning the Difference Maker

For the naïve externalist, facts about the natural history of *water* ground its reference. While this view is far more austere than anything proposed by the initial architects of externalism, it appears to underwrite the intuition—shared by many—that LMs' outputs refer. Yet, the foregoing shows that the naïve externalist is incapable of answering Q4—a necessary condition, it would seem, on any successful theory of reference. Is there a way of embellishing naïve externalism so that it remains naïve and yet provides an answer to our question?

Translated to address our current concerns, the question is: Can the naïve externalist provide something more specific than *being implicated in the natural history of 'water'* in answering Q4? And, in fact, a more fine-grained difference maker can be gleaned from the following claim: "Luke's use of *Peano* refers to Peano. . . thanks to the fact that his use of *Peano* traces back via causal-historical lines to Peano himself" (page 1195). Applied to the case in question, we get, as a candidate difference maker, *being such that 'water' traces back via causal-historical lines to it*.

This gives us the following answer to Q4:

A1.   Earthians' use of *water* traces back via causal-historical lines to H_2O; it
       does not trace back via causal historical lines to oxygen.

But *being such that 'water' traces back via causal-historical lines to it* is easily seen *not* to be a difference maker: If *water* traces back via causal-historical lines to H_2O, then it traces back via causal-historical lines to oxygen as well. So A1 fails as an answer to Q4. (Things do not improve if we add details about the event at which the causal chain terminates.

---

4 Although I'll focus on Q2 in what follows, the answer to Q3 is: *Because H_2O is implicated in the natural history of Earthian's use of 'water' whereas H_2O is not implicated in the natural history of Twin Earthian's use of 'water'.* Here, the difference maker is the property of *being such that H_2O is implicated in its natural history*: Earthian's use of *water* has it, Twin Earthian's use of *water* does not.

For example, *being such that 'water' traces back via causal-historical lines to a baptismal event involving it* applies equally to H_2O and oxygen.)

Perhaps we can invoke deference to the community, or, if not, to a set of experts? Since the Earthians cannot intelligibly be said to defer to themselves, deference to "the community" is a non-starter. A difference maker that invokes the relevant experts, i.e., chemists, is more promising—for example: *being such that chemists are disposed to use 'water' in its presence*. This gives us the following answer:

A2.   Earthians defer to chemists in their use of *water*. Chemists are disposed to use *water* just when they are in the presence of H_2O; they are not disposed to use *water* just when they are in the presence of oxygen.

This, however, needs to be finessed. If chemists use *water* when they are in the presence of H_2O, they also use *water* when they are in the presence of oxygen. Since every event in which they token *water* is one involving H_2O, every such event also involves oxygen.

But chemists arguably have a more fine-grained disposition: to use *water* only in the presence of pure samples of H_2O, and never in the presence of pure samples of oxygen. Moreover, our deference is based on this more rarefied usage. Here we get:

A3.   Earthians defer to chemists in their use of *water*. Chemists use *water* only when in the presence of pure samples of H_2O; they do not use *water* only when in the presence of pure samples of oxygen.

Note that it would be too strong to say that chemists are disposed to use *water* when, and only when, they are in the presence of pure samples of H_2O—in a word, that their use of *water covaries* with the presence of pure samples of H_2O. What A3 states is that chemists refrain from using *water* in the absence of pure samples of H_2O. It does not require, implausibly, that they use the term whenever such pure samples are present.

In fact, A3 is still too strong. Even chemists will, on occasion, "speak with the vulgar" and use *water* in the presence of impure samples of H_2O—to request a glass of drinking water, for example. One can rule out such possibilities by invoking the conversational context, as in A4:

A4.   Earthians defer to chemists in their use of *water* with respect to chemists' purely scientific communications. These chemists have the following disposition: in the course of scientific communications, to use *water* only when in the presence of pure samples of H_2O; they do not have an analogous disposition with respect to pure samples of oxygen.

This easily accommodates a stray use of *water* in the presence of an impure sample of H_2O. Since the chemist's request for a glass of drinking water is, technically, not part of a purely scientific exchange, it does not constitute a counterexample to A4.

Even if we grant the assumption that context will make it clear which sense of *water* the speaker intends, there is the question as to whether this context-sensitivity comports with naïve externalism. According to A4, the larger community does not defer to expert usage *per se*, but to expert usage *relative to scientific exchanges*. This requires an explanation as to how deference is made to the just right class of uses, and it is

hard to see how our naive externalist, who eschews talk of "beliefs, experiences, and discriminatory capacities," can accomplish this.

These worries are compounded by the observation that, even with respect to strictly scientific communications, there is significant variation with respect to chemists' use of *water*. As Weisberg (2006) notes:

> Chemists' ordinary use of natural kind terms are highly context sensitive.... Say a chemist needs to use a warm water bath to keep a reaction at a particular temperature. Her request for more warm water from an associate will be taken to mean that she wants a substance that is composed primarily of H_2O molecules with an isotopic abundance somewhere in the normal background range. On the other hand, if she is doing a very isotopically sensitive kinetic study, her request for water would be interpreted in a different way... (Weisberg 2006, page 343)

If this is correct, *water* does not refer to a unique chemical kind, even among chemists.[5] But then the idea that the experts use *water* when, and only when, they are in the presence of pure samples of H_2O is a fiction—even if we restrict the claim to scientific exchanges. The upshot is that the naive externalist remains incapable of answering Q4.

We are now in a position to see the "slip" I referred to in Section 2: Mandelkern and Linzen answer what appears to be a contrastive question, Q1, but which is in reality a conjunction of two incomplete questions. Choosing randomly between the two, we saw that the externalist can provide an adequate (if perhaps unsatisfying) answer to one completion (Q2), but fails entirely to answer another completion that is equally pressing (Q4).

The externalist thus provides only a partial ground of *P* in Q2—the fact that *water* refers to H_2O among Earthians. While citing this partial ground (namely, that Earthians' use of *water* traces back via causal-historical lines to H_2O) might strike us as unsatisfying in the case of Q2, it at least makes a move in the answering game. In contrast, a partial ground of *P* in the context of Q4 is not simply unsatisfying: It fails to be a difference maker and thus fails to answer the question. The upshot is that externalism, as understood by Mandelkern and Linzen, cannot ground reference.[6]

4. Conclusion

Naïve externalism—the thesis that the natural history of a name or kind term grounds its referential properties—presents an alluringly simple answer to our leading question, whether LMs can refer. While I have argued that natural history alone cannot ground a

---

5 See Hendry (2006, pages 869–874), LaPorte (1996, pages 118–121), and LaPorte (2004, pages 103–110) for similar arguments.

6 One might take issue with my appeal to the contrastivist framework and in particular reject the claim that Q1a is, in fact, elliptical for the contrastive question, Q2. If it is not, the fact that the externalist cannot answer Q4 should be of no particular concern. No one should expect that the correct answer to a simple (i.e., non-contrastive) *Why?* question will provide the difference-making answer to *every* contrastive expansion of that question. The externalist has thus done their job if they have addressed Q1a as originally stated.

  But a look at Mandelkern and Linzen's answer to Q1a should quickly dispel this idea ("Because Earthians' use of *water* traces back via causal-historical lines to H_2O"). While this is technically an answer, it can hardly lay claim to being an *explanation*—certainly not once it is part of the common ground (which it now is) that Earthians' use of water traces back via causal-historical lines to an event involving oxygen. NB: I am not here assuming a contrastive analysis of Q1a, but merely appealing to an intuitive judgment—which I assume the externalist shares—that the answer is not explanatory in the relevant context. Thanks to Antonella Mallozzi for pushing me to address this objection.

name's reference, I have been neutral on whether or not an externalism construed either along explicitly Kripkean lines, or an externalism that invokes community but, *pace* Kripke, eschews deference, can provide an appropriate grounding for LM reference. We can express these options with the following conditional: *if we can make sense of LMs possessing referential intentions, or if deference to usage is not essential to reference, then there is no principled barrier to LMs' outputs referring*. In this case, my concern is not with the conditional, which I accept, but with the antecedent, about which I have doubts. In the case of the conditional that Mandelkern and Linzen are, in effect, asserting (namely, *that if naïve externalism is true, then there are no principled obstacles to the claim that LMs' outputs refer*) my concerns are different. Here, my doubts are with the conditional itself. As I have argued, I see no obvious route from antecedent to consequent.

In any case, as I hope to have shown, the path to the claim that LMs' outputs refer remains daunting—a far cry from the easy route that the naïve externalist promised.

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. `https://doi.org/10.1145/3442188.3445922`

Cappelen, Herman and Max Deutsch. 2024. Reference without deference. In Ernie Lepore, and Una Stojnić, editors, *Oxford Handbook of Contemporary Philosophy of Language*. Oxford University Press. `https://doi.org/10.1093/oxfordhb/9780192856852.013.33`

Cappelen, Herman and Josh Dever. 2021. *Making AI intelligible: Philosophical Foundations*. Oxford University Press. `https://doi.org/10.1093/oso/9780192894724.001.0001`

Chalmers, David. 2023. Could a large language model be conscious? *Boston Review*. `https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/`.

Hendry, Robin. 2006. Elements, compounds, and other chemical kinds. *Philosophy of Science*, 73(5):864–875. `https://doi.org/10.1086/518745`

Kripke, Saul A. 1980. *Naming and Necessity*. Harvard University Press.

LaPorte, Joseph. 1996. Chemical kind term reference and the discovery of essence. *Noûs*, 30(1):112–132. `https://doi.org/10.2307/2216306`

LaPorte, Joseph. 2004. *Natural Kinds and Conceptual Change*. Cambridge University Press. `https://doi.org/10.1017/CBO9780511527319`

Lipton, Peter. 2004. *Inference to the Best Explanation*, second edition. Routledge. `https://doi.org/10.4324/9780203470855`

Mandelkern, Matthew and Tal Linzen. 2024. Do language models' words refer? *Computational Linguistics*, 50(3):1191–1200. `https://doi.org/10.1162/coli_a_00522`

Ostertag, Gary. 2023. Meaning by courtesy: LLM-generated texts and the illusion of content. *The American Journal of Bioethics*, 93(10):91–93. `https://doi.org/10.1080/15265161.2023.2249851`, PubMed: 37812115

Putnam, Hilary. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193.

Titus, Lisa. 2024. Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, 83(101174). `https://doi.org/10.1016/j.cogsys.2023.101174`

Weisberg, Michael. 2006. Water is *not* H_2O. In Davis Baird, Eric Scerri, and Lee McIntyre, editors, *Philosophy of Chemistry: Synthesis of a New Discipline*. Springer, pages 337–345.