TRANSNATIONAL PRESS® LONDON

The Agonist - May 2024 - Volume: 18, No: 1, pp. 7 – 17
ISSN: 2752-4132 (Print) ISSN 2752-4140 (Online)
journals.tplondon.com/agonist

Received: 31 January 2024 Accepted: 26 May 2024
DOI: https://doi.org/10.33182/agon.v18i1.3261

Nietzschean Language Models and Philosophical Chatbots: Outline of a Critique of AI

Anthony Kosar[1]

Abstract

Developers of the deep learning algorithms known as large language models (LLMs) sometimes give the impression that they are producing a likeness to the human brain: data-processing 'neural networks' are 'taught' to recognize patterns in language and then, based on this pattern recognition, create or generate new content in the form of natural, humanlike speech, writing, images, etc. The results have been unsettling to some; less appreciated are the metaphysical assumptions underlying the attribution of any meaningful agency whatsoever to an algorithm. In this essay, Nietzsche's thoughts on the "seduction of grammar" form the basis of one possible critique of generative AI – a critique, moreover, which exposes our society's current fixation with LLMs for what it is: a fetishization and humanization of new technologies.

Keywords: *Nietzsche; generative AI; LLMs; metaphysics of language; common philosophy of grammar*

Whereas Nietzsche's conscious reflections on his 'writing machine'[2] ended when the latter finally broke down beyond repair,[3] his thoughts on language spanned his entire writing career, from at least as early as his student days all the way through the writings of 1888. In an early set of notes from 1869/70, language is understood metaphorically as "a complete organism [*ein ganzer Organismus*]": living and breathing through us, it is almost parasitic in its grip on us, for whom language is truly indispensable. The metaphor emphasizes the natural necessity of language in contrast to anything inorganic and artificial, such as machines, for example, which are far more transitory in comparison, Nietzsche's irreparable writing machine a case in point both metaphorically and literally.

There are, however, good reasons to believe that this latter assumption is short-sighted, especially now with machine learning on the rise, a term used to refer to the subfield of AI responsible for the deep learning algorithms on which chatbots such as OpenAI's ChatGPT and Google Gemini are based: large language models (LLMs). In fact, the very tendency to describe deep learning algorithms in biological terms – each of which is supposed to be an artificial iteration of its natural counterpart: intelligence, activity, neural networks, natural language – suggests rather that the algorithms themselves might one day become conscious[4] and take on a life of their own, in much the same way that language seems to have done so already for Nietzsche: "too complicated" to be "the work of an

[1] Anthony Kosar, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany.
E-mail: anthony.kosar@philosophie.uni-freiburg.de
[2] The reflections include poetry, as well. See NL 1882, 18[2].
[3] Cf. Kittler 293–310.
[4] On the different possible kinds of consciousness in AI, see Hildt 2–3.

individual" and yet "for the masses much too unified, a complete organism."[5] Either that, or we would do well to take Nietzsche's own metaphor of language – especially when applied to LLMs – as just that: a literary device of our own making which we ourselves ultimately wield.

There is some truth to both ways of reading Nietzsche's metaphor. For one, we really can't overstate the power language has over us. There is simply no way that we could ever just think it away – and in fact, there are some who would even argue that thought itself would cease to exist without it. As the 19th-century linguist and early scholar of comparative religion F. Max Müller observed, "Language and thought are inseparable. Words without thoughts are dead sounds; thoughts without words are nothing. Thinking is silent speaking; speaking, thinking aloud. The word is the thought incarnate" (Müller 330–331, translation my own).[6] One could take this insight of Müller's a step further and consider how the cultural techniques of reading and writing, along with their corresponding materialities, have become so intertwined with our thinking patterns as to become virtually inseparable from them. Perhaps we will one day have a similar realization about the role of the now ubiquitous chatbots – that is, once we've learned to stop worrying and love the bot.

Were we to take Nietzsche's metaphor of language more strictly *as metaphor*, however, then we are forced to ask whether it really applies to generative AI in quite the same way. Some clarity is needed on what we mean by the latter and why this has so quickly begun to be perceived as a threat to some. Developers of LLMs sometimes give the impression that they are producing a likeness to the human brain. Data-processing 'neural networks' are 'taught' to recognize patterns in language and then, based on this pattern recognition, create or generate new content in the form of natural, humanlike speech, writing, images, etc. This is achieved by "pretraining" the networks on massively large datasets "to predict a hidden part of an input sentence—a method called 'self-supervised learning.'" (Mitchell and Krakauer 1) Through this, the model will then generate new content that has not been fed into the system directly.[7] The results have been unsettling to some; less appreciated are the metaphysical assumptions underlying the attribution of any meaningful agency whatsoever to an algorithm. What could it mean for the latter to be active and intelligent – to say nothing of it being sentient, conscious or self-aware?

I maintain that preliminary answers to these questions can be found by reflecting with Nietzsche the idle psychologist[8] on some of the ways in which language leads us astray –

[5] Cited in Crawford 222 and 225. The notes are titled "*Vom Ursprung der Sprache*" and can be found both in the original German as well as in Claudia Crawford's English translation in her monograph, *The Beginning of Nietzsche's Theory of Language* (222–226). Crawford understands the metaphor to derive indirectly from Eduard von Hartmann's *Philosophie des Unbewussten*, where the role of instinct in conceptual formation is underscored. Thus for Nietzsche, it is on the basis of human instinct that conscious thought qua language ultimately emerges. See especially Crawford 17–21 and 42–50. Incidentally, the metaphor is also found in Wilhelm von Humboldt. Cf. Chomsky, *Cartesian Linguistics* 74.

[6] On Nietzsche's readings of Müller and the influence of the latter on his thoughts on the relationship between language, metaphysics and mythological thinking, see Zavatta, "Die in der Sprache versteckte Mythologie" 295–298.

[7] For a more detailed overview of the training of large language models, see Hinton, "Will digital intelligence replace biological intelligence?"; Piantadosi and Hill, "Meaning without reference in large language models" and Wei et al. 3–6. Insight into the "black box" of the transformer neural network in particular can be found in Uszkoreit et al., "Transformer."

[8] As opposed to the philosopher of the *Übermensch*, for example. Nietzsche was perhaps more given to leisurely pursuits than his overly bombastic writing style would suggest, as the original title to *Twilight of the Idols* attests: "A Psychologist's Idleness." Cf. Sommer, NK 6/1 197 and 211–213.

specifically the way it allows us to imagine ourselves as free and sovereign individuals[9] within a bustling world of other similar 'agents' (atoms, forces and the like). As there doesn't seem to be a 'way out' of language – assuming we wanted out – it behooves us to become aware of the mechanisms that are at play in it (grammar, tropes, etc.) so that we might have a little more control over them. Nietzsche offers a philosophically fruitful path for doing just that – one which is no less relevant to our interaction with chatbots, to which we relate solely linguistically, after all.

Nietzsche's most succinct remarks on the inherent "prejudices" of language[10] are found dispersed throughout his late writings.[11] Here in particular, the "seduction of grammar" comes into play as the mechanism which misleads us into an erroneous view of ourselves as 'free agents.'[12] For Nietzsche, however, there is neither free nor unfree will, only relatively stronger and weaker *wills* by necessity.[13] That we think of ourselves otherwise speaks to a kind of petrified *habitus* which can perhaps by now never be thrown off completely – one, moreover, with strong mythological undertones. When criticizing what he understands to be our singular sense of subjectivity, Nietzsche takes up a then popular concept in religious studies: the fetish. Language, as Nietzsche understands it, is a fetishizing activity (*Fetischwesen*).[14] Through it we posit a world of unchanging substances *behind* everything that we experience, thereby altering completely what it is that was experienced in the first place. We *distort* the endless flow of becoming in order to make it fit a particular mold, i.e. *appear* stable. Since becoming is presumably all there is, any truth which is expressed in language is therefore suspect at best, an irony to which Nietzsche alludes in the following metalinguistic reflection from *Twilight of the Idols*:

> Change, alternation, becoming in general were formerly taken as proof of appearance, as a sign of the presence of something which led us astray. Today, on the contrary, we see ourselves, as it were, entangled in error, *necessitated* to error, to precisely the extent that our prejudice in favor of reason compels us to posit unity, identity, duration, substance, cause, materiality, being; so sure are we, on the basis of a strict reckoning, *that* the error is to be found here. The situation is the same as with the motions of the stars: in that case error has our eyes, in the present case our *language* as a perpetual advocate. Language belongs in its origin to the age of the most rudimentary form of psychology: we find ourselves in the midst of a rude fetishism when we call to mind the basic presuppositions of the metaphysics of language – which is to say, of *reason*. It is that [fetishism] which sees everywhere doer and deed; that [fetishism] which believes in will as cause in general; that [fetishism] which believes in the "I," in the I as being, in the I as

[9] Our subjective identity lends itself equally well, moreover, to the image of unfree, oppressed victims – who are only oppressed, however, to the extent that there is believed to be an oppressor.
[10] The "prejudices [*Vorurtheile*]" in question are literally the *judgments* we make as native speakers *before* any critical thinking sets in, every last one of our reflexive, unreflective *Vor-urteile*.
[11] Though the thoughts are already present in *The Wanderer and His Shadow*: "Jedes Wort ist ein Vorurtheil" (WS §55, KSA 2.577).
[12] Cf. BGE §16–17 (KSA 5.29–31), GM I §13 (KSA 5.279–281) and TI "Errors" §3 (KSA 6.90–91).
[13] Cf. BGE §21, KSA 5.35–36.
[14] Cf. Sommer, NK 6/1 298–300.

> substance, and which projects its belief in the I-substance onto all things – only thus does it create the concept "thing"… (TI "'Reason'" §5).[15]

The grammatical structures in question for Nietzsche are those of subject and verb, doer and deed, which make it so that even nonliving entities like the sun can only really be talked about *as though* they were living agents, rising and falling at will even when we have since learned to think about them otherwise. Nietzsche's anthropology of knowledge explains this tendency as a perpetual projection of our own self-image onto the outside world:

> The oldest and longest-lived psychology was at work here – indeed it has done nothing else: every event was to it an action, every action the effect of a will, the world became for it a multiplicity of agents [*Thätern*], an agent ("subject") slipped itself into every event [*schob sich allem Geschehen unter*]. Out of himself man projected his three "inner facts," that in which he believed more firmly than in anything else: will, spirit, ego – only from the concept "ego" did he take the concept "being," he posited "things" as possessing being according to his own image, according to his concept of the ego as cause. No wonder he later always rediscovered in things only *that which he had put into them*! (TI "Errors" §3)[16]

That we can talk about a 'thing' in the first place is due entirely to the fact that we have already made it like us to some degree: an agency in its own right. And yet to assume that we ourselves already possess such agency is itself the result of a psychology that has perhaps from the very beginning falsely accepted a theory of causality for which only the will is causal – an assumption Nietzsche himself is no longer 'willing' to make:

> The will no longer moves anything, consequently no longer explains anything – it merely accompanies events, it can also be absent. The so-called 'motive': another error. Merely a surface phenomenon of consciousness, an accompaniment to the act, which conceals rather than exposes the *antecedentia* of an act (TI "Errors" §3).[17]

No doubt our LLMs, too, have undergone an anthropomorphic transformation in our hands, yet the likeness of such software to the human brain – if there really is any – begs another question: if we aren't really as free as our grammar 'seduces' us into believing we are, but merely believe in our ideas with the same natural necessity as a spider spinning its web,[18] then what light might this shed on generative AI, which supposedly does the same as we do, but far more efficiently?[19] Was Nietzsche perhaps saying that we were all automata after all?

A closer look at his conception of freedom and the will is first needed before questions like these can be conclusively answered. Manuel Dries has gone very far in elucidating Nietzsche's understanding of freedom as a first-personal experience of resistance rather

---

[15] Hollingdale, R. J., translator. *Twilight of the Idols and The Antichrist*. By Friedrich Nietzsche, Penguin Books, 1968, translation modified. Cf. KSA 6.77. On the meaning of the "reason" which Nietzsche sets in quotation marks (*"Vernunft"*), see Sommer NK 6/1 285–286.

[16] Hollingdale's translation, significantly modified. Cf. KSA 6.91.

[17] Hollingdale's translation, modified slightly. Cf. KSA 6.91.

[18] TL §1, KSA 1.885. On Nietzsche's "regulative fictions" as "consciously willed semblance," see Vaihinger 771–790.

[19] Geoffrey Hinton, the so-called "godfather of AI," is convinced that the neural networks at the base of LLMs in particular are far better versions of our own brains, from which the former are supposedly inspired. See "Will digital intelligence replace biological intelligence?"

than as the metaphysical *creatio ex nihilo* of an absolutely free will. Far from denying that we act freely, Nietzsche conceives of freedom naturalistically in the language of drives and resistances overcome by the organism.[20] A particular grammar might very well convince us that there is something distinct in the endless flow of becoming called the subject, and that this subject is solely responsible for a particular kind of event in the world by the name of actions.[21] For Nietzsche, however, "there is no 'being' behind the deed, its effect and what becomes of it; 'the doer' is invented as an afterthought, – the doing is everything" (GM I §13).[22] The biggest critique he might have leveled at the growing fascination with LLMs therefore certainly would have had less to do with what these can or cannot be said to "understand" than with a point of pride of what it means to be a free human subject in the first place. It may seem trivial to state that humans are categorically different from chatbots, however much the output of the latter may *seem* human to some, and yet pessimists are surely also right that any categorical differences between us and our machines will hardly matter anymore if the algorithm finds that all it needs to fulfill a task is more and more control – and ultimately gets what it's after. A forcefield of resistances to be overcome is precisely how Geoffrey Hinton conceives of AI, which uses language effortlessly and efficiently even without a metaphysical self calling the shots.[23] If "consciousness" is what is so central for Nietzsche's conception of agency, as Dries argues, then artificial consciousness would seem more relevant than ever now for those who take the latter to be a threat to their own autonomy.

To return to the language faculty specifically – since this is how LLMs' agency is defined – the question could just as well be raised as to what we mean by autonomous language use, and whether LLMs can rightfully be said to 'speak freely.' It was the creative aspect of language which led Noam Chomsky – following Descartes – to conceive of a kind of language organ inexplicable in purely mechanistic terms, one which would at the very least imply – if it could never outright prove – that each individual's capacity for language is innate rather than learned.[24] Since generative AI merely predicts the most likely combination of words, it does not really create ideas by Chomsky's estimation.[25]

Though Nietzsche never went so far as to posit a language organ, his sustained interest in art and culture repeatedly begs the question of what constitutes genuine creative activity and output. For all their indeterminateness, LLMs are rather predictable, something which any good artist will work against. For all their humanness, LLMs lack a personality expressed in their 'work' (a 'soul' in another sense). However much we may be willing to attribute agency to chatbots, this same agency is not only unthinkable without the engineers who design LLMs to begin with but would fall apart without all the human actors working behind the scenes to keep the machine running. As Anna Wiener aptly observes, "simulated chat obscures the reality of what it takes to create, train, update, and maintain large language models, which are, at least for now, hugely expensive and resource-

[20] See especially 144–149.
[21] Cf. Williams 8.
[22] Diethe, Carol, translator. *On the Genealogy of Morality*. By Friedrich Nietzsche, Cambridge University Press, 2006. Cf. KSA 5.279.
[23] Cf. "Will digital intelligence replace biological intelligence?"
[24] Cf. *Cartesian Linguistics* 59–77. In her essay "Nietzschean Linguistics," Benedetta Zavatta sketches an alternative Nietzschean linguistic paradigm based on individual experience through acculturation.
[25] Cf. Chomsky et al., "The False Promise of ChatGPT." Hinton reacts to the criticisms of Chomsky and his coauthors in "Will digital intelligence replace biological intelligence?"

intensive" ("The Age of Chat"). According to Wiener, it is precisely this very mundane, human work that is hidden, and intentionally so if the mirage is to be sustained. How much we are willing to go along with this "fantasy," as Wiener later calls it,[26] remains to be seen. Nietzsche's service perhaps lies in urging us to ask the less intuitive question of whether we are keeping another fantasy alive regarding our own language faculty.

One can suspect that Nietzsche, though his thinking about the human conceives of the latter on a continuum with nature rather than as a categorically separate entity,[27] would have had at least some sympathy with Chomsky on the point of genuine creativity. It seems wrong to suppose that we aren't doing anything but *predicting* with language when we speak and write. And yet the question of how particular grammars structure our thinking problematizes this suspicion somewhat, for even our apparently most individual thoughts can't escape linguistic conventions established long before our time.

> The strange family resemblance of all Indian, Greek, and German philosophizing speaks for itself clearly enough. Where there are linguistic affinities, then because of the common philosophy of grammar (I mean: due to the unconscious domination and direction through similar grammatical functions), it is obvious that everything lies ready from the very start for a similar development and sequence of philosophical systems; on the other hand, the way seems as good as blocked for certain other possibilities of interpreting the world. Philosophers of the Ural-Altaic language group (where the concept of the subject is the most poorly developed) are more likely to "see the world" differently, and to be found on paths different from those taken by the Indo-Germans or Muslims: the spell of particular grammatical functions is in the last analysis the spell of *physiological* value judgments and racial conditioning [*Rasse-Bedingungen*] (BGE §20).[28]

Nietzsche is far from conceding in this passage that a single "universal grammar" – Chomsky's language organ – guides all of our thinking. Rather, it is the particular language in which each of us was raised and in which we spend most of our lives which structures our thoughts the most – contingent factors if there ever were any. This is evident even in Nietzsche's own thoughts on the role of the 'subject' for thought in general: had he been raised in a non-Indo-European language, it's very possible that his conception of primordial thought would have fit a pattern different from that for which the 'agent' (*Thäter*) was everywhere to be found, and not rather unstable and unaccountable process-events. Though our thoughts are not determined by a program set in advance by a programmer, they are still the result of our socialization – in Nietzsche's words, the conditions of the race (*Rasse-Bedingungen*), specifically as these work with and against our

[26] "All of this infrastructure buttresses a fantasy. Technologists have long dreamed of having interpersonal relationships with programs. Recently, Sam Altman, the C.E.O. of OpenAI, reminisced to the *Wall Street Journal* about being a child, peering into his Macintosh, and having the 'sudden realization' that 'someday, the computer was going to learn to think.' (The *Journal*'s use of the word 'realization' suggests fact, rather than conjecture; it's not yet clear whether L.L.M.s, or subsequent technologies, will be able to 'think' in any recognizable or meaningful way.)" (Ibid.) The WSJ article referenced is from March 2023: https://www.wsj.com/tech/ai/chatgpt-sam-altman-artificial-intelligence-openai-b0e1c8c9?mod=Searchresults_pos1, last accessed on 6 May, 2024.

[27] Cf. Abel 6–11.

[28] Norman, Judith, translator. *Beyond Good and Evil*. By Friedrich Nietzsche, Cambridge University Press, 2001. Cf. KSA 5.34–35.

physiology. It was only through this socialization, moreover, that we humans came to consciousness in the first place, and with this, language.[29]

Much like our machines then, we, too, are determined – if not by a divine determinant, then by our long and often bloody prehistory, preceded by the far more vast and elusive history of evolution.[30] Our 'program' is human society itself, without which we would have no sense of ourselves as agents. This, however, is precisely what the machines lack, which is not to suggest that we necessarily have an advantage over them for having society, only that whatever knowledge they could ever possess will be something entirely foreign to us – though they use the same words, speak the same language.

That they don't quite speak the same language, however, is apparent in any number of ways, perhaps the most glaring of which is LLMs' utter lack of moral sense. The most scathing critique of LLMs' amorality[31] that I've come across so far comes from Chomsky himself, who together with his coauthors takes up Hannah Arendt's famous concept of the "banality of evil" to describe ChatGPT as a kind of glorified autocomplete which merely "summarizes the standard arguments in the literature […], refuses to take a stand on anything, pleads not merely ignorance but lack of intelligence and ultimately offers a 'just following orders' defense, shifting responsibility to its creators" ("The False Promise of ChatGPT").[32] The disregard for its own reputation is another characteristic of the LLM's amorality, a feature which ultimately bars it from being a social person for Jacob Browning: "LLMs are not social persons because they are insensitive to norms and thus prone to bullshitting, inconsistency, offensiveness, and irresponsibility" ("Personhood and AI"). Moral persons that we are (which is to say that we do care to some degree at least about where we stand with others), Browning concludes that, since "it is unclear (to say the least) how to make a machine that cares about its reputation, suffers from sanctions, and takes risks when making claims," it is unlikely that a machine will ever "understand *us*."

No matter how well LLMs may parrot moral truisms, they can never undergo by way of simulation the same history that has informed our moral imagination, depicted in the *Genealogy of Morality* in its emergence as a long and cruel process of socialization at the end of which one finally learns to regard oneself as a 'debtor' to the tribe, as it were – regardless of whether or not one is in fact a culprit (*Thäter*) in the literal sense. Only through such

[29] Cf. GS §354, KSA 3.590–593.
[30] Cf. D §18, KSA 3.32.
[31] That is, if it is not rather a contradiction in terms to speak of a non-human agent's lack of moral standards.
[32] Interestingly, it's the preprogrammed, almost Kantian aspect of Chomsky's theory of language which anthropologist Chris Knight believes to have been especially attractive to earlier computer scientists, who gained with it philosophical backing for their work: "My own suspicion is that, for Chomsky's institutional milieu, his ideas just had to be true. Endorsing Chomsky meant endorsing his picture of language as a digital computational device. To any computer scientist, that was an attractive idea. Chomsky's programme promised to elevate a generation of military-sponsored computer scientists to the status not merely of electronics engineers but philosophers in the tradition of Plato and Descartes, geniuses delving into the greatest of all mysteries – the ultimate nature of human language and mind. Right or wrong, it was clearly too attractive a vision to be lightly set aside. Even to this day, despite decades of disappointment and failure, the vision still enjoys passionate support" ("The two Chomskys"). Knight attempts in his essay to account for what appears to be a split personality in Chomsky – the linguist on the one hand, political activist on the other – on account of the popularity which the theory of generative transformational grammar had among computer scientists and those in the American military industrial complex. In other words, the guilt which Chomsky must have felt – according to Knight – at his ideas being used by the American war machine ultimately allowed Chomsky the political activist to become even stronger. Thankfully, for Chomsky's sake, his linguistic theories have never proven to be very effective as weapons.

blood and cruelty did 'man' become a 'person' in any recognizable sense,[33] and with this newfound sense of self learn to distinguish between necessary and accidental,

> to think causally, to view the future as the present and anticipate it, to grasp with certainty what is end and what is means, in all, to be able to calculate, compute – and before he can do this, man himself will really have to become *reliable*, *regular*, *necessary*, even in his own self-image, so that he, as someone making a promise is, is answerable for his own *future*! (GM II §1)[34]

For Nietzsche at least, society and its "social straightjacket"[35] are central to our language, without which one may very well imagine completely different – and to us at least, altogether alien – forms of thought than those that seem possible to express. These alien forms are amply provided by our LLMs.

What does it mean to be an agent that does not experience, for whom (or for which) there isn't really anything that it's *like to be* that agent?[36] We would do well, I think, to remember that AI is a tool not unlike other cultural techniques such as reading and writing. To compare it to the human brain is no less misleading than if we were to use an encyclopedia for the analogy. Thankfully we've learned by now that even books can be misleading, based as they are on grammars which by no means reflect the world as it truly is. On the other hand, no insight would be possible without the concepts and structures that give it form, that *make it form* in the first place, without which our thoughts would remain a nearly imperceptible flicker. That we have these forms with which to think at all is what makes the training of LLMs possible, which for some have, as if overnight, become too much like us for comfort. And yet, since LLMs' predicted probabilities of word sequences – which cannot truly be said to be 'utterances' – have no lived history behind them, they lack a meaning that our utterances can't help but possess simply by virtue of the fact that these speak from, for and to us. Without this, AI-generated writing often reads like a Mad Libs, where words only really make sense as parts of speech.

Could a machine ever have a truly original thought? But this admittedly merely sidesteps the question of originality. We could just as easily ask whether our machines do not in fact show us how little originality there is in each one of us. Do we really do anything different than LLMs, and if so, how could we know? Whether we are, in fact, doing anything besides predicting the next most probable word in a sequence is one thing; how meaning is constituted by us and machines is another. For us, at least, the reference of our meaning often goes beyond the linguistic system itself, pointing to something 'out there' in the extra-linguistic world of experience. Josef Simon describes this feature of our language as a metaphysical aporia – one, moreover, which is likely impossible for us to overcome by now without giving up on language altogether:

> By the concept of a language is meant a *system* of signs which are supposed to mean "something" extralinguistic in their composition according to rules "internal to" this system – rules which likewise systematically exclude signs

[33] Cf. GM II §4–8, KSA 5.297–307. The edited anthology of Benne and Müller problematizes the dichotomy 'person-subject' found in Nietzsche in illuminating ways.
[34] Diethe's translation. Cf. KSA 5.292.
[35] GM II §2, KSA 5.293.
[36] To borrow a phrase from Thomas Nagel's seminal essay, "What is it like to be a bat?" (165–180)

> foreign to it. In this respect the *concept* of a language is a metaphysical concept. It refers at once to one of the "essential" aporiae, one could perhaps say *the* fundamental aporia of metaphysics: signs linked to rules are, in this internal linkage, at the same time supposed to refer to external, "objective" relations. The "form of the representation" should be regulated in a way that is inherent to the system, and precisely *therein* "correspond" to "external" relations (7, translation my own).

These observations of Simon's well predate the advent of LLMs, which do happen to represent a perfectly closed linguistic system without any means of reference outside itself.[37] Seeing as the language currently at LLMs' disposal is a human one translated into numerical tokens and back, the only conception of language any LLM could possibly come up with as of now would be one utterly foreign to what actually goes on inside the black box, for in order for a model to conceive of a non-metaphysical, non-referential language unlike the one – or many – it 'speaks,' it would first have to become aware of itself and then compare what it does with what we do when we speak, read and write. As of now, any novel form of understanding found in LLMs will likely be limited to their "superhuman predictive ability" (Mitchell and Krakauer 4), lacking as they do a self-knowledge which we – its programmers – either possess or strive for. Why or how we do this, however, remains the mystery of our own black box: human consciousness.

Like language, AI can sometimes seem to be its own organism. Not even its engineers quite understand it fully: no matter how much they train it on data, how it organizes this data to produce what it does remains the mystery of the black box. The ability of AI to learn new skills beyond what it was trained on has also shown how abilities not present in smaller language models can emerge in unforeseen ways in much larger ones.[38] Where the goal of these models is artificial general intelligence, this is both promising and unsettling at once, for without a genuine understanding of what goes on inside these machines, there is no telling what havoc AI systems might wreak.

And yet, though the fear mongering is no doubt at least in part due to a fetishization of algorithms, as though these were in fact autonomous agents who could harm us if they so willed, Nietzsche's early metaphor for language as its own organism is admittedly no less fetishistic. Dependent as we are on language, it is not its own agent with its own will, but only a part of the human organism and our social needs. Assuming that AI is not so much a replacement as it is an enhancement of the human being, one will continue to ask how it might change the way we think in much the same way that writing and language have done so already. If we are currently transitioning – as we very much seem to be – from a 'common philosophy of grammar' to a 'common philosophy of *data*,'[39] then it behooves

---

[37] As Piantadosi and Hill demonstrate, though LLMs have no lived experience from which the meaning of many words – e.g. concrete, tangible objects – is derived, they tend to be successful nonetheless with abstract concepts, which don't have any one single referent outside of themselves, but derive their meaning primarily in relation to other words within a system ("Meaning without reference in large language models").

[38] Cf. Wei et al. 6–11.

[39] I understand this to be a form of reason which can't help but generalize in probabilities, thereby overlooking the lived quality of human life. Talal Asad conceives of such a "calculative reason" in the modern, secular nation-state in particular as a closed-system language that ignores everything outside of itself: "In an important sense, the primary language of secular

us to ask how we might yet put our chatbots to better philosophical use so as not to be drowned out by them.[40] Ironically, though understandably, the most unpredictable output of LLMs is exactly what engineers are dead set on preventing at all costs, given the very threat of destruction through bias and misinformation that could follow from it. The so-called "hallucination," for example – understood in the field of machine learning to be a made-up response not based on the trained data[41] – at once speaks to the human, all-too-humanness of our machines and the very amorality which we strive to prevent. As our artificial counterparts, chatbots seem to show us by way of negation what it means to be human after all. As with language, art and literature, we confront ourselves 'reflected' by our cultural techniques, with and against which we struggle to understand ourselves a little better.

As to how much value we are willing to place on our LLMs, if we agree with Nietzsche that anything of quality is lived, written in blood, then we already have our benchmark. No cultural technique is sentient, least of all our large language models, however much their engineers may want to convince us – and perhaps even themselves do believe – otherwise. That we are capable of suffering might be the only thing that really separates us from our machines for now.[42]

Works Cited

Abel, Günter. "Bewusstsein – Sprache – Natur. Nietzsches Philosophie des Geistes." *Nietzsche-Studien*, vol. 30, 2001, pp. 1–43.

Asad, Talal. *Secular Translations. Nation-State, Modern Self, and Calculative Reason.* Columbia University Press, 2018.

Benne, Christian, and Enrico Müller (eds.). *Ohnmacht des Subjekts – Macht der Persönlichkeit*. Schwabe, 2014.

Browning, Jacob. "Personhood and AI: Why large language models don't understand us." *AI & Society*, 12 July 2023, https://doi.org/10.1007/s00146-023-01724-y. Last accessed on 11 Feb. 2024.

Chomsky, Noam. *Cartesian Linguistics. A Chapter in the History of Rationalist Thought*, Third Edition. Cambridge University Press, 2009.

Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. "Noam Chomsky: The False Promise of ChatGPT." *The New York Times*, 8 Mar. 2023.

Crawford, Claudia. *The Beginnings of Nietzsche's Theory of Language.* De Gruyter, 1988.

Dries, Manuel. "Freedom, Resistance, Agency." *Nietzsche on Mind and Nature*, edited by Manuel Dries and P. J. E. Kail. Oxford University Press, 2015, pp. 142–162.

Hildt, Elisabeth. "Artificial Intelligence: Does Consciousness Matter?" *Frontiers in Psychology*, vol. 10, 2 July 2019, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.01535/full. Last accessed on 3 Mar. 2024.

Hinton, Geoffrey. "Will digital intelligence replace biological intelligence?" *Romanes Lecture*, 19 Feb. 2024, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjbg8uM0-CEAxVSEGIAHW4-

reason that employs numbers requires inattention to the world from which it has been abstracted" (97). Asad's findings in this and other of his works are very much in the spirit of Horkheimer and Adorno's theory of the ways in which enlightenment regresses into barbarism, and are intended to be read as a cautionary admonition to those of us living in the secular modern world.

[40] Andreas Urs Sommer suggests that having "the entire cultural inheritance in a vast digital *nunc stans*," though it certainly would have been unseemly to a 19th-century member of the educated classes, might even lead to a new kind of (barbaric!) cultural creativity – one, moreover, which Nietzsche perhaps even spearheaded ("Nietzsches kulturschöpferische Barbaren" 177, translation my own).

[41] Morally speaking: a falsehood generated by the algorithm when it is unable to make a true statement supported by the facts at hand.

[42] It is suffering, too, which serves as the basis of meaning making for Nietzsche. Cf. GM III §28, KSA 5.411–412.

DvcQwqsBegQICBAG&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DN1TEjTeQeg0&usg=AOvVaw0S6YKnutxy5Stps9zuf5Mg&opi=89978449. Last accessed on 6 Mar. 2024.

Horkheimer, Max, and Theodor W. Adorno. *Dialektik der Aufklärung. Philosophische Fragmente*. Fischer Taschenbuch Verlag, 2017.

Kittler, Friedrich. *Grammophon Film Typewriter*. Brinkmann & Bose, 1986.

Knight, Chris. "The two Chomskys." *Aeon*, 8 Dec. 2023, https://aeon.co/essays/an-anthropologist-studies-the-warring-ideas-of-noam-chomsky. Last accessed on 3 Feb. 2024.

Mitchell, Melanie, and David C. Krakauer. "The debate over understanding in AI's large language models." *PNAS*, vol. 120, no. 13, 2023, https://doi.org/10.1073/pnas.2215907120. Last accessed on 11 Feb. 2024.

Müller, Dr. Max. *Vorlesungen über die Wissenschaft der Sprache*, vol. 1. Leipzig, Verlag von Gustav Mayer, 1863.

Nagel, Thomas. *Mortal Questions*. Cambridge University Press, 2012.

Piantadosi, Steven T., and Felix Hill. "Meaning without reference in large language models." 12 Aug. 2022, https://arxiv.org/abs/2208.02957. Last accessed on 4 Mar. 2024.

Simon, Josef. *Philosophie des Zeichens*. De Gruyter, 1989.

Sommer, Andreas Urs. *Kommentar zu Nietzsches Der Fall Wagner, Götzen-Dämmerung* (= NK 6/1). De Gruyter, 2012.

Sommer, Andreas Urs. "Nietzsches kulturschöpferische Barbaren: Beobachtungen zu blonden und anderen Bestien in *Genealogie der Moral* I 11, nebst einer unwissenschaftlichen Nachschrift." *Nietzsche, das 'Barbarische' und die 'Rasse'*, edited by Sebastian Kaufmann and Markus Winkler. De Gruyter, 2022, pp. 163–180.

Uszkoreit, Jakob et al. "Transformer: A Novel Neural Network Architecture for Language Understanding." *Google Research: Blog*, 31 Aug. 2017, blog.research.google/2017/08/transformer-novel-neural-network.html. Last accessed on 30 Jan. 2024.

Vaihinger, Hans. *Die Philosophie des Als Ob. System der theoretischen, praktischen und religiösen Fiktionen der Menschheit auf Grund eines idealistischen Positivismus. Mit einem Anhang über Kant und Nietzsche*. Verlag von Reuther & Reichard, 1911.

Wei, Jason et al. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*, 2022, https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/ 69c8bf111 e0c 1 6 1d773704cb17b1c378953061a0.pdf. Last accessed on 13 Feb. 2024.

Wiener, Anna. "The Age of Chat." *The New Yorker*, 17 June 2023, https://www.newyorker.com/culture/the-weekend-essay/the-age-of-chat. Last accessed on 6 May 2024.

Williams, Bernard. "Nietzsche's Minimalist Moral Psychology." *European Journal of Philosophy*, vol. 1, no. 1, 1993, pp. 4–14.

Zavatta, Benedetta. "Die in der Sprache versteckte Mythologie und ihre Folgen fürs Denken. Einige Quellen von Nietzsche: Max Müller, Gustav Gerber und Ludwig Noiré." *Nietzsche-Studien*, vol. 38, 2009, pp. 269–298.

Zavatta, Benedetta. "Nietzschean Linguistics." *Nietzsche-Studien*, vol. 42, 2013, pp. 21–43.