Non-Human Words: On GPT-3 as a Philosophical Laboratory

Tobias Rees

In this essay, I investigate the effect of OpenAI's GPT-3 on the modern concept of the human (as alone capable of reason and language) and of machines (as devoid of reason and language). I show how GPT-3 and other transformer-based language models give rise to a new, structuralist concept of language, implicit in which is a new understanding of human and machine that unfolds far beyond the reach of the categories we have inherited from the past. I try to make compelling the argument that AI companies like OpenAI, Google, Facebook, or Microsoft effectively are philosophical laboratories (insofar as they disrupt the old concepts/ontologies we live by) and I ask what it would mean to build AI products from the perspective of the philosophical disruptions they provoke: can we liberate AI from the concept of the human we inherited from the past?

In May 2020, OpenAI released GPT-3 (Generative Pre-trained Transformer 3), an artificial intelligence system based on deep learning techniques that can generate text. GPT-3's interface invites a user to provide the AI system with a bit of text and then, based on the prompt, GPT-3 writes. GPT-3 can write short stories, novels, reportages, scientific papers, code, and mathematical formulas. It can write in different styles and imitate the style of the text prompt. It can also answer content-based questions (that is, it learns the content of texts and can articulate this content). It can translate text from almost any language into almost any other; and it can provide summaries of lengthy passages.

The quality of GPT-3's output is remarkable, often impressive. As many critiques have pointed out, GPT-3 makes silly errors that no human would ever make. And yet GPT-3's translations often exceed translations done by humans and capture even subtle differentiations and wordplays; the summaries are almost always concise; and the text it generates on the basis of prompts is most often surprisingly consistent: GPT-3 can mimic the style of an author to such a degree that is nearly impossible to determine whether the text was written by a given author or by GPT-3.

How can we relate to GPT-3? Or to the many other, often equally powerful large language models (LLMs) built over the last few years: Google's BERT, LaMDA, and Wordcraft; Microsoft's Megatron-Turing Natural Language Generation;

https://doi.org/10.1162/DAED_a_01908

Inspur's YUAN 1.0; Huawei's PanGu-Alpha; Naver's HyperCLOVA; or Sberbank's various Russian models, most notably ruRoBERTa-large?

I have come to think of the development of GPT-3 and its kin as a far-reaching, epoch-making philosophical event: the silent, hardly noticed undoing of the up-until-now exclusive link between humans and words.

The consequences of this undoing are sweeping: the entire modern world – the modern experience of what it is to be human, as well as the modern understanding of reality – is grounded in the idea that we humans are the only talking thing in a world of mute things.

No longer.

Philosophical Stakes

At the beginning of the seventeenth century, a remarkable transformation in our understanding of language took place. Up until that time, the comprehensions of the world as described by Plato and Aristotle had largely remained in place. Most humans still experienced themselves, in accordance with the writings of the Greek philosophers, to be living in a God-given nature-cosmos in which everything – including the human thing – had a well-defined role.

Nature – a metaphysical ground – was all there was.

The particular role of humans in this nature-cosmos was defined by their having language. The assumption was that at the beginning of time, humans received a spark of the divine *logos* that gave things their eternal essence or names, of which the visible world was a mere reflection. This divine gift not only enabled humans to communicate with one another, it also gave them access, via contemplation (a practice that consists in applying logos to itself), to the true names of things and thus to the eternal order of the real.

Around 1600, the ancient, medieval nature-cosmos began to break open. Within a few short decades, the comprehension of reality – the structure of experience of what it is to be human – underwent a remarkably far-reaching change. And at the center of this change was language.

If until then language was a divine gift that enabled humans to know the eternal essence/names of things, then now language became the human unique power to name things and to thereby order and know them and bring them under human control. If language had hitherto defined the role of humans *in* the nature-cosmos, then language was now what set them apart from what was increasingly considered to be *mere* nature: nature was no longer understood and experienced as a divine cosmos but as the *other* of the human, as the nonhuman realm of animals and plants, as mere matter organized in mechanical principles.

The exemplary place where this new concept of language – of humans – is articulated is René Descartes's *Discourse on the Method*, published anonymously in 1630.

> For it is a very remarkable thing that there are no humans, not even the insane, so dull and stupid that they cannot put words together in a manner to convey their thoughts. On the contrary, there is no other animal however perfect and fortunately situated it may be, that can do the same. And this is not because they lack the organs, for we see that magpies and parrots can pronounce words as well as we can, and nevertheless cannot speak as we do, that is, in showing that they think what they are saying. On the other hand, even those humans born deaf and dumb, lacking the organs which others make use of in speaking . . . usually invent for themselves some signs by which they make themselves understood. And this proves not merely animals have less reason than men but that they have none at all. . . . We ought not to confound speech with natural movements which betray passions and may be imitated by machines as well as be manifested by animals. . . . They have no reason at all; it is just nature which acts in them according to the disposition of their organs, just as a clock, which is only composed of wheels and weights.

According to Descartes, language is a power only we humans possess, a power that sets us apart, in a qualitative, unbridgeable way from everything else there is, notably from animals and machines. It is the fact that we have language, for Descartes a proxy for reason (logos), that we humans are more than mere matter extended in space: we are subjects, capable of thought and knowledge.

It is difficult to exaggerate the importance of *Discourse on the Method* for the birth of the modern age. It was more than just an argument: it was an obituary for the medieval nature-cosmos and the birth certificate of a new era: modernity, or the age of human exceptionalism.

It articulated a new structure of experience, which remained relatively stable for the subsequent four hundred years:

Here the human, there the world.

Here humans, subjects in a world of objects, thinking and talking things in a world of mere and mute things, there nature and machines.

Here freedom, knowledge, reason, politics, there nothing but necessity and mechanism.

Here language, there silence.

Enter GPT-3.

If machines could talk and write, if they had words too, then that would make untenable the clear-cut distinction between human and non-human things (animals and machines) that has defined the modern Western experience of self and the world ever since the early seventeenth century. If language were no longer exclusive to humans, then comprehension of reality that silently structures the modern understanding and experiencing of the world would no longer hold. The logical presupposition on which that structure was dependent – that only humans have words – would be false.

Arguably, a machine with words is something our classical modern ontology cannot accommodate: it cannot be subsumed under our modern understanding of what it is to be human – or of what machines are – without disrupting it.

Or am I overstating the importance of GPT-3?

Critique (Meaning)

I understand that there are those who judge me to be naive. I am thinking of the many critics who have rejected, often with vehemence, the idea that GPT-3 really has words. When I worked through these critics, I found myself struck by the recognition that, no matter how diverse their background, they almost all offer a version of a single argument, which roughly goes like this: no matter how good GPT-3 appears to be at using words, it does not have *true* language; it is just a technical system made up of data, statistics, and predictions.

If one asks the critics what *true* here refers to, the common answer is understanding meaning.[1] What though does *meaning*, what does *understanding*, refer to? Why, and in what sense, does GPT-3 or other LLMs not have it?

I found the most insightful and articulate voice among the critics to be linguist Emily Bender. In a recent podcast, discussing her critique of LLMs generally, she explained her position in terms of an analogy with machine vision:

> Vision. There is something both about perception, so how does the eye and the ocular nerve . . . what happens when the light hits the retina and then what happens in the brain that's processing it to create maybe some sort of representation that just has to do with the visual stimulus? But then that gets connected to categories of things. So vision is not just about physics and perception and stuff like that. It is also about categories and ontologies and how we understand our world.[2]

Bender argues that vision is made up of two intertwined aspects. On the one hand is the physical and chemical reality of the act of seeing, the proper domain of the natural sciences. On the other is what she calls the "categories and ontologies and how we understand our world."

This latter aspect is of its own kind and lays beyond the physical realities of nature and thus beyond the reach of the natural sciences: it is the proper realm of the human, constituted by our capacity to invent meaning, to organize objects that surround us by assigning them meaning, a process that produces "our world."

And language?

In analogy to her description of vision, Bender understands language as a combination of a formal, quasi-mechanical aspect that can be studied by science, and a domain that lies beyond the reach of the natural sciences or engineering: the domain of meaning. As she put it in a paper published with computational linguist Alexander Koller:

> We take form to be any observable realization of language: marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators. We take meaning to be the relation between the form and something external to language.[3]

It is from this vantage point that she criticizes LLMs: they are trained on form and hence will fail when it comes to meaning. As she says in the podcast:

> So, what do we mean by meaning? . . . [With] a language model . . . the only input data it has is the form of language . . . that's all it sees. And meaning is a relationship between those linguistic forms and something outside of language.

According to Bender, the intersubjective, intentional production and negotiation of that language is a quality unique to humans. Non-humans have "*a priori* no way to learn meaning." Whenever we think otherwise – whenever we assume that animals or machines have that ability too – we are mistaken. "Our singular human understanding" may trick us into believing that animals or LLMs have language and hence meaning too. But they do not. As a recent paper Bender cowrote puts it:

> Contrary to how it may seem when we observe its output, a language model is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.[4]

Here the human, singular subject in a world of objects, there physics, chemistry, nerves, stimuli, machines, algorithms, parrots.

Here the human, there everything else.

If I add up the remarks offered by Bender – and by most other critics of GPT-3 – I arrive at remarkable ontological descriptions of what humans are, of what the role of humans in the world is: Being human unfolds outside the realm of nature and the natural sciences, outside the realm of all the things that are reducible to the mechanical or to physics and chemistry. The way in which we humans manage being outside of nature – being irreducible to mere mechanism or instinct – is the intentional invention of meaning: we are intentional subjects who can make things into objects of thought by assigning and negotiating meaning. Inventing meaning is the human way of being in the world: it enables us to organize things, allows us to jointly produce worlds. Generally, meaning is arbitrary. Things do not have meaning in themselves; they must be endowed with meaning, and only humans can do that.

Because only humans *can* have language or words, any claim that machines have language is, ultimately, an ontological error insofar as it assigns qualities of one ontological area (humans) to another one (animals or machines). This ontological error is problematic insofar as it compromises what humans can be: it

reduces us to machines or to mere mechanism. To defend humans against machines – machines like GPT-3 – is thus to defend an ontological, moral, and somehow timeless order.

In short, at the core of the suggestion that GPT-3 does not have understanding or meaning is an ontological claim about what language *is*, a claim grounded in definite statements about what humans are (subjects, things with words) and also about what animals and machines are (objects, things without words).

The force of Bender's critique, which I take to be exemplary of most critics of GPT-3, depends on whether this ontological claim holds. Does it?

One way of addressing this question is to ask: When and under what circumstances did the idea that language is about meaning, and that only existentially situated subjects can have words, first emerge? What sets this concept apart from prior conceptualizations? What shifts and transformations in a conceptual understanding of the world – of humans and of language – had to occur for the ontology defended by the critics of GPT-3 to become possible?

A Brief History of Words (and Humans)

In rough strokes, there have been three epochs in the history of how humans understand language and experience the capacity to speak: I call them ontology (words and being), representation (words and things), and existence (words and meaning).

Words and being. Most ancient and medieval authors took it for granted that visible reality is a mere reflection of invisible ideas generated by a divine logos: the things we see or can touch were considered imprecise, steadily changing derivatives of something unchanging and eternal. The path toward understanding, thus, was hardly a study of the visible, of haptic, empirical things. On the contrary, the only way to comprehend how reality is organized was a contemplation of the invisible.

One privileged form contemplation took was a careful analysis of language. The reason for this was the conviction that humans had language (logos) only insofar as they had received a spark of the divine logos – a divine logos that also organized reality: intrinsic in language was thus a path toward the real. All that was necessary was for humans to direct their thinking to the structure of language and thought (logos).

As Aristotle puts it in his *Peri Hermeneias*:

> Spoken words are the symbols of mental experience and written words are the symbols of spoken words. Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images.

The sounds humans conventionalize into nouns or verbs differ, Aristotle argues, but the structure of language is – must be – the same for all humans. After all, it is a direct reflection of the divine logos.

Aristotle's assumption that language is the path to understanding being – that there is a direct correlation between words and things – builds directly on Plato's theory of ideas, and remained the unchallenged reference for well over a thousand years. Things only began to change around 1300.

Words and things. The parting ways of words and things was a gradual and cumulative process. Its earliest indicator was the emergence of nominalism in the early fourteenth century, in the works of Peter Abelard and William von Ockham. Independently from one another, the two clerics wondered if perhaps words are not in reality arbitrary conventions invented by humans, rather than the way to true being.

At least in retrospect, the importance of nominalism was that it seemed to imply that things could perhaps exist independent from words. For Aristotle and Plato, the *really* real was immaterial. In the aftermath of Abelard and von Ockham, this began to change. Reality was increasingly defined in empirical terms:

Call it a sweeping shift in the experience of what reality is – a shift from the invisible to the visible, from the abstract to the concrete.

One effect of this shift in the comprehension of reality was the emergence of a new problem: If things were independent of words, if reality could not be understood in terms of an abstract reflection about language, then how is knowledge possible? How can humans get to know the natural world that surrounds them? Can they?

The effort to answer this question amounted to the invention of a whole new comprehension of both reality and knowledge, in physical rather than in metaphysical, in empirical rather than in contemplative terms.

The two most prominent authors of this form-giving, new-age-defining invention were Thomas Hobbes on the one hand and René Descartes on the other. As if coordinating their efforts across a distance, they in parallel argued that what set humans apart from mere nature (now understood as the realm of animals and plants) was their capacity for empirical knowledge – and insisted that the key to this new kind of knowledge was in fact language. Though in sharp contrast to their scholastic contemporaries, they no longer thought of language in terms of a divine logos but rather as a human-unique tool for naming and ordering reality. The French and the English philosopher bid their farewell to the idea that language is the major path to being and instead rethought it in terms of representation. To quote Hobbes:

> By the advantage of names it is that we are capable of science, which beasts for want of them, are not; nor man without. . . . A name is a word taken at pleasure to serve for a mark, which may raise in our mind a thought like to some thought we had before, and

> which, being pronounced to others, may be a sign to them of what thought the speaker had, or had not, before in his mind.

For Hobbes, language was arbitrary, and precisely because it was arbitrary, it was a powerful tool for naming things and for building a systematic representation of the outside world. Through language (representation) we store, organize, and examine our experiences or ideas.

I would like to bring into focus the quite radical conceptual difference between the early modern and the ancient concept of language: what separates the former from the latter is hardly progress. As if all that was needed was to think a little harder and a little longer, and then one would suddenly recognize that language is not the path to understanding being. In fact, the ancients thought pretty hard and pretty long. Their research was as rigorous and careful as could be. Rather, what separates Plato or Aristotle from Descartes or Hobbes or Locke is a series of sweeping conceptual transformations that led to a whole new experience and understanding of reality: and this new understanding of reality was simply unthinkable from within the concept – the epistemic – space available to ancients.

Words and meaning. It is difficult today to appreciate that humans who lived before 1700 did not think of themselves as individuals. Before that time, the truth about a given human being was sought in that which they have in common with types: Choleric? Melancholic? Sanguine? It was only in the course of the eighteenth century that the idea emerged that what defined someone is that in which they differ from anyone else: their individuality.

Historians have explained the gradual ascendance of individuality with the collapse of feudalism, provoked by both the Enlightenment and a nascent industrial revolution. The Enlightenment, the argument goes, steadily undermined the religious institutions and their grip over life, and the industrial revolution provoked a younger generation to leave the countryside for the city, trading a life in the family compound for an ultimately individual pursuit. The coming together of these two developments was an early version of the lonely crowd: individuals cut loose from their families and their villages, alienated from the beliefs they had grown up with.

One of the outcomes of these developments – call it the incidental rise of individualism and city life – was the sudden appearance, seemingly out of nowhere, of reflections about the subjective, inner experiences in the form of diaries, autobiographies, and letters.

This discovery of interiority and subjectivity is a fascinating chapter in the history of humanity. Prior to the second half of the eighteenth century, documentations of subjectivity written up for their own sake are practically absent: Clerics may have written highly stylized accounts of conversion experiences or confessions. But deeply individual or circumstantial reflections about the ups and downs

of everyday human life – from boredom to disease, fear, love or death – are nowhere to be found.

By the end of the nineteenth century, the rise of individualism, the discovery of subjectivity, and the fading of the grip religious institutions previously had over life gave rise to the birth of a new branch of philosophy: existentialism. Surprising as it may sound, conceptualizations of what it is to be human in terms of *existence*, in terms of being thrown in a meaningless world, alone, with questions but without answers, cannot be found before the late nineteenth century.

And language?

The emergence of subjectivity ultimately resulted in a whole new understanding of language. The form-giving author of this new understanding was Ernst Cassirer.

Beginning shortly after the turn of the century, Cassirer set out to cut loose modern philosophy from the epistemological project that until then had defined it, and sought instead to ground it in terms of existence. His point of departure was Kant. Kant's "Copernican revolution" suggested that human experience – and hence knowledge – is contingent on a set of categories. As Kant saw it, these categories are transcendental or independent of experience. Put in a formula, they are the condition of the possibility of experience, not the outcome of experience. According to Cassirer, Kant got it both right and fundamentally wrong. He got it right insofar as humans are indeed subjects whose minds can only operate with the help of categories. But he got it all wrong because these categories are not transcendental epistemological principles. They are symbols. They are arbitrary meanings invented and stabilized by humans:

What Cassirer offered was a radically new concept of the human and of language.

Of the human:

The basic condition of the human was no longer what it had been from Descartes and Hobbes onward: the capacity to know. And the basic question of philosophy was no longer what it had been from Descartes via Hume to Kant: can humans know? How? Instead, the basic condition of humans became now their existential condition. Humans are simultaneously defined by their finding themselves thrown into a meaningless world and their singular capacity to invent meaning. Call it word-making.

Of language:

At the center of this new conceptualization of what humans are is language. Language now ceases being primarily about representation, a tool in the process of producing knowledge, and instead comes into view as a means to produce and assign and negotiate meaning. Call it world-making. In short, there was a shift from understanding the subject as capable of knowledge to comprehending the subject as capable of inventing meaning through language.

Though no matter how much Cassirer reversed modern philosophy, in one key respect the existence-meaning configuration did not break with the subject-knowledge configuration of the early modern period: human exceptionalism. Humans were still singular and exceptional. They, and they alone, have words, can think, wonder, make meaning. Here subjects longing for meaning, producing meaning, there the world of objects, nature and technology, meaninglessness.

I summarize my tour de force: The concept of humans and of language upheld by the critics of GPT-3 is neither timeless nor universal. Their claims about what language *is* are of recent origin, little more than a century old. They are a historically situated, specific mode of knowing and thinking that first surfaced in the early twentieth century and that became possible only through a set of conceptual ruptures and shifts that had occurred during the eighteenth and nineteenth centuries.

Two far-reaching consequences follow.

The first is that in prior times, the conceptualization of humans in terms of existence, and of language in terms of meaning, would have made no sense because these prior times had different structures of experience and understanding of reality (reality was organized by quite radically different ontologies).

The second is that there is no timeless truth to the concept of the human and language upheld by critics of GPT-3. It is a historically contingent concept. To claim otherwise would mean to miss the historicity of the presuppositions on which the plausibility of the argument is dependent.

To me, the importance of GPT-3 is that it opens up a whole new way of thinking about language – and about humans and machines – that exceeds the logical potential of argument that the critics uphold. GPT-3, that is, provides us with the opportunity to think and experience otherwise, in ways that are so new/different that they cannot be accommodated by how we have thought/experienced thus far.

Once this newness is in the world, the old, I think, can no longer be saved. Though what is this newness?

Structuralism, Experimental

I think of GPT-3 as engineering in terms of structuralism.

The idea of structuralism – a term coined by Russian-American linguist Roman Jakobson in the 1920s – goes back to a distinction between *langue* and *parole* originally offered a few years earlier by the Swiss linguist Ferdinand de Saussure.

De Saussure observed that most humans tend to think of language in terms of the act of speaking (*parole*). From this perspective, language is grounded in a human subject and in a subject's intentions to communicate information. Alternatively, he argued, we can think of language as an arbitrary system that exists somewhat independent of speakers and can be analyzed independent of who speaks (*langue*).

One may object, he conceded, that language does not really exist independent of the individual: that is, situated human subjects and their experiences. However, it is hard to disagree with the simple observation that we humans are born into language: into a system that predates any speaker and, in fact, determines the space of possibility from within which a subject can speak.

To support his argument in favor of a structural approach, de Saussure offered his famous distinction between signifier (*signifié*) and signified (*signifient*). It is often assumed, falsely, to suggest that there is no causal relation between signifier and the signified, that meaning is arbitrarily assigned to things. Though that point was already made seven hundred years earlier, by the nominalists. Rather, de Saussure's point was that the relation between signifier and signified was subject to a set of law-like principles that are independent from the subject (the meaning intended or experienced by a speaker) as well as from the object (actual meaning that is experienced or the actual thing to which meaning is assigned).

In his words, "language is a system of signs that expresses ideas."

Put differently, language is a freestanding arbitrary system organized by an inner combinatorial logic. If one wishes to understand this system, one must discover the structure of its logic. De Saussure, effectively, separated language from the human.

There is much to be said about the history of structuralism post de Saussure. However, for my purposes here, it is perhaps sufficient to highlight that every thinker that came after the Swiss linguist, from Jakobson (who developed Saussure's original ideas into a consistent research program) to Claude Lévi-Strauss (who moved Jakobson's method outside of linguistics and into cultural anthropology) to Michel Foucault (who developed a quasi-structuralist understanding of history that does not ground in an intentional subject), ultimately has built on the two key insights already provided by de Saussure: 1) the possibility to understand language, culture, or history as a structure organized by a combinatorial logics that 2) can be – must be – understood independent of the human subject.

GPT-3, wittingly or not, is an heir to structuralism. Both in terms of the concept of language that structuralism produced and in terms of the antisubject philosophy that it gave rise to. GPT-3 is a machine learning (ML) system that assigns arbitrary numerical values to words and then, after analyzing large amounts of texts, calculates the likelihood that one particular word will follow another. This analysis is done by a neural network, each layer of which analyzes a different aspect of the samples it was provided with: meanings of words, relations of words, sentence structures, and so on. It can be used for translation from one language to another, for predicting what words are likely to come next in a series, and for writing coherent text all by itself.

GPT-3, then, is arguably a structural analysis of *and a structuralist production of* language. It stands in direct continuity with the work of de Saussure: language comes into view here as a logical system to which the speaker is merely incidental.

There are, however, two powerful differences between de Saussure and the structuralists. The first is that the incidental thing that speaks is not a human; it is a machine.

All prior structuralists were at home in the human sciences and analyzed what they themselves considered human-specific phenomena: language, culture, history, thought. They may have embraced cybernetics, they may have conducted a formal, computer-based analysis of speech or art or kinship systems. And yet their focus was on things human, not on machines. GPT-3, in short, extends structuralism beyond the human.

The second, in some ways even more far-reaching, difference is that the structuralism that informs LLMs like GPT-3 is not a theoretical analysis of something. Quite to the contrary, it is a practical way of building things. If up until the early 2010s the term *structuralism* referred to a way of analyzing, of decoding, of relating to language, then now it refers to the actual practice of building machines "that have words."

The work of OpenAI and others like it, from Google to Microsoft, is an engineering-based structuralism that experimentally tests the core premises of structuralism: That language is a system and that the thing that speaks is incidental. It endows machines with a structuralist equipment – a formal, logical analysis of language as a system – in order to let machines participate in language.

What are the implications of GPT-3 for the classical modern concept of the human, of nature, and of machines?

GPT-3 provokes a conceptual reconfiguration that is similar in scale to the ones that have occurred in the 1630s (Descartes, Hobbes) and around 1900 (Cassirer). Call it a philosophical event of sweeping proportions:

Machine learning engineers in companies like OpenAI, Google, Facebook, or Microsoft have experimentally established a concept of language at the center of which does not need to be the human, either as a knowing thing or as an existential subject. According to this new concept, language is a system organized by an internal combinatorial logic that is independent from whomever speaks (human or machine). Indeed, they have shown, in however rudimentary a way, that if a machine discovers this combinatorial logic, it can produce and participate in language (have words). By doing so, they have effectively undermined and rendered untenable the idea that only humans have language – or words.

What is more, they have undermined the key logical assumptions that organized the modern Western experience and understanding of reality: the idea that humans have what animals and machines do not have, language and logos.

The effect of this undermining is that the epoch of modernity – call it the epoch of the human – comes to an end and a new, little understood one begins: machines with words not only undermine the old, they also create something new and different. That is, LLMs not only undermine the presuppositions on which

the seventeenth- and the late-nineteenth-century concept of the human/language were contingent, they also exceed them and open new possibilities of thinking about the human or machines.

In fact, the new concept of language – the structuralist concept of language – that they make practically available makes possible a whole new ontology.

What is this new ontology? Here is a rough, tentative sketch, based on my current understanding.

By undoing the formerly exclusive link between language and humans, GPT-3 created the condition of the possibility of elaborating a much more general concept of language: as long as language needed human subjects, only humans could have language. But once language is understood as a communication system, then there is in principle nothing that separates human language from the language of animals or microbes or machines.

A bit as if language becomes a general theme and human language a variation among many other possible variations.

I think here of the many ML-based studies of whale and dolphin communication, but also of Irene Pepperberg's study of Alex the parrot (pace Descartes and Bender).[5] I think of quorum sensing and the communication – the language – that connects trees and mycelial networks. And I think of GPT-3, BERT, YUAN, PanGu, and RU.

I hasten to add that this does not mean these variations are all the same. Of course they are not. Human language is in some fundamental way different from, say, the clicking sounds of sperm whales. But these differences can now come into view as variations of a theme called language.

What is most fascinating is that the long list of variations runs diagonal to the old ontology that defined modernity, the clear-cut distinction between human things, natural things, and technical things, thereby rendering them useless.

The power of this new concept of language that emerges from GPT-3 is that it disrupts human exceptionalism: it opens up a world where humans are physical things among physical things (that can be living or non-living, organism or machine, natural or artificial) in a physical world. The potential is tremendously exciting.

Beyond Words

Each month, humans publish about seventy million posts on WordPress, arguably the dominant online content management system. If we estimate that an average article is eight hundred words long, then humans produce about fifty-six billion words a month, or 1.8 billion words a day on WordPress. GPT-3 is producing 4.5 billion words a day, more than *twice* what humans on WordPress are doing collectively. And that is just GPT-3; there are the other LLMs.[6]

The implications of this are huge. We are exposed to a flood of non-human words. What to do about this flood of words that do not ground in subjective ex-

perience, an intent to communicate, a care for truth, an ambition to inspire? How to relate to them, how to navigate them?

Or are these the wrong questions for the new age of non-human words? How do we ask these questions without either defending the old concept of the human or naively embracing machines?

And this is just words.

LLMs like GPT-3 have recently been called "foundational models."[7] The suggestion is that the infrastructure that made LLMs possible – the combination of enormously large data sets, pretrained transformer models, and significant amounts of compute – is likely to be the basis for all future AI. Or at least the basis for the first general purpose AI technologies that can be applied to a series of downstream tasks.

Language, almost certainly, is just a first field of application, a first radical transformation of the human provoked by experimental structuralism. That is, we are likely to see the transformation of aspects previously thought of as exclusive human qualities – intelligence, thought, language, creativity – into general themes: into series of which humans are but one entry.

What will it mean to be surrounded by a multitude of non-human forms of intelligence? What is the alternative to building large-scale collaborations between philosophers and technologists that ground in engineering as well as an acute awareness of the philosophical stakes of building LLMs and other foundational models?

It is naive to think we can simply navigate – or regulate – the new world that surrounds us with the help of the old concepts. And it is equally naive to assume engineers can do a good job at building the new epoch without making these philosophical questions part of the building itself: for articulating new concepts is not a theoretical but a practical challenge; it is at stake in the experiments happening in (the West at) places like OpenAI, Google, Microsoft, Facebook, and Amazon.

Indeed, as I see it, companies like OpenAI, Google, Facebook, and Microsoft have effectively become philosophical laboratories: they are sites that produce powerful ruptures of the categories that define the spaces of possibility from within which we (still) think. At present, these philosophical ruptures occur in an unplanned, almost accidental way – because the philosophical is not usually a part of R&D or product development. My ambition is to change that: What would it take to build AI in order to intentionally disrupt some of the old (limiting or harmful or anachronistic) categories we live by? Or, perhaps less provocative, what would it mean to build thinking and talking machines from the perspective of the ruptures they inevitably provoke?

AUTHOR'S NOTE

I am deeply grateful to Nina Begus for our many conversations about AI and language in general and about GPT-3 in particular.

ABOUT THE AUTHOR

Tobias Rees is the Founder and CEO of Transformations of the Human (tofth.org). Prior to founding ToftH, he served as Reid Hoffman Professor of Humanities at Parsons/The New School and was a Director at the Los Angeles–based Berggruen Institute. He is a Fellow of the Canadian Institute for Advanced Research (CIFAR) and the author of three books, most recently *After Ethnos* (2018).

ENDNOTES

[1] Emily Bender and Alexander Koller, "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* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[2] "Is Linguistics Missing from NLP Research? w/ Emily M. Bender–#376," The TWIML AI Podcast (formerly This Week in Machine Learning & Artificial Intelligence), May 18, 2020.

[3] Bender and Koller, "Climbing towards NLU."

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021).

[5] Irene Maxine Pepperberg, *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots* (Cambridge, Mass.: Harvard University Press, 2002).

[6] Jason Dorrier, "OpenAI's GPT-3 Algorithm Is Now Producing Billions of Words a Day," SingularityHub, April 4, 2021, https://singularityhub.com/2021/04/04/openais-gpt-3-algorithm-is-now-producing-billions-of-words-a-day/; and "A Live Look at Activity Across WordPress.com," https://wordpress.com/activity/.

[7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.