Large Language Models and the Patterns of Human Language Use
An Alternative View of the Relation of AI to Understanding and Sentience

Christoph Durt,[1*] Tom Froese,[2] and Thomas Fuchs[3]

[1] Section of Phenomenology, Psychiatric University Hospital, University of Heidelberg, Voßstrasse 4, 69115 Heidelberg
[2] Embodied Cognitive Science Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Okinawa, Japan
[3] Klinik für Allgemeine Psychiatrie, Zentrum für Psychosoziale Medizin, Universitätsklinikum Heidelberg, Voß-Str. 4, 69115 Heidelberg

* Corresponding author; email: Christoph@Durt.de

Abstract
Large Language Models (LLMs) such as ChatGPT are deep learning architectures that have been trained on immense amounts of text. Their ability to produce human-like text has led to claims that LLMs either possess or simulate some form of conscious experience and understanding. This paper argues that experience and understanding do play an important role, but that it is very different from what is commonly thought. LLMs model the statistical contours of vast amounts of human language use. We use phenomenological considerations of human language production to explain that human language use is intertwined with experience and understanding. Symbolic language does not simply correspond to internal or external 'meaning', but is meaningful because it scaffolds our interactions and mental life. In human language production, preconscious anticipatory processes interact with conscious experience. Human language use constitutes and makes use of given patterns, constantly rearranging them in a way that we liken to making a collage. LLMs do not need to replicate or simulate human mental life in order to produce text that appears meaningful to humans. Rather, they can infer statistical patterns from meaningful patterns in written language use, including clichés and biases. The impressive extent to which these can be computationally reassembled into text that makes sense to humans does not show that LLMs have developed understanding or sentience. Rather, it can reveal the surprising extent to which human language use gives rise to and is guided by patterns.

1. A new era of artificial intelligence

Human language can express a wide variety of thoughts, experiences, feelings, and emotions, besides a plethora of other uses. As an illustrative example, consider a recent large-scale survey of people's experiences of the Covid pandemic, which revealed the many ways in which their experiences have been impacted (Froese et al. 2021). This can be considered a typical response to a question targeted at eliciting open-text description of unusual disruptions of experiences:

> *Q: 5. Have you had experiences, of any kind, that seemed strange or unfamiliar? If so, please try to describe them.*
> A: I have had a few experiences that have seemed strange or unfamiliar since the pandemic began. I have found myself feeling more anxious in general, and I have also been having more intrusive thoughts about the virus. These experiences have been strange and unfamiliar to me, as I have never experienced anything like them before.

We can easily identify with what this participant has been going through – except that, in this case, the answer to the question was not part of any participant's subjective report but was generated by the LLM GPT-3 as part of our effort to test its capacities to mimic subjectivity. There is a growing diversity of such LLMs with comparable and increasing performances, including AI21, BLOOM, GLM-130B, Gopher, GPT derivatives (e.g., GPT-3, GPT-4, ChatGPT, Instruct GPT, Bing), Jurassic-1, LaMDA, Luminous, Megatron-Turing NLG, OPT, and PaLM. In a wide sense, the term LLM is often used to refer to all of them, although they also make use of additional methods that alter the results, such as filters. In a narrower sense, the underlying LLM can be distinguished from the various adaptations. In the narrow sense, LLMs are sometimes called "foundational models" (Liang et al. 2022, 7), highlighting that they provide a foundation for numerous applications. Since they represent texts as numerical relations (vectors) between text chunks (tokens), LLMs are in their core instances of vector space models applied to large volumes of text.

The astonishing performance of GPT-3 has given rise to far-reaching claims, such as that GPT-3 "is able to do basic common-sense reasoning with high accuracy" (Chojecki 2020), or that for these systems "statistics do amount to understanding" (Agüera y Arcas 2022). A reputable investigative media outlet claimed that a text it printed as an op-ed was "written" by GPT-3, and that the editing was "no different to editing a human op-ed" (GPT-3 2020). A (soon thereafter dismissed) Google employee, fully aware of the LLM LaMDA's computational underpinnings, has even proclaimed his belief that the system has developed sentience (Tiku 2022). Such claims, however, are based on a selective consideration of output. LLMs quite often fail in producing sensical responses. The opinion contribution that was allegedly written by GPT-3 is the result of cherry-picking the best human output and disregarding the uninteresting or unhelpful output. Numerous limitations of LLMs have been pointed out, such as their difficulty to "[r]eliably maintain a coherent argument or narrative thread over long periods of time; maintain consistency of gender or personality; employ simple grammar rules; show basic knowledge and commonsense reasoning" (Elkins and Chun 2020). LLMs have troubles with formal reasoning, world knowledge, situation modeling, and social-cognitive abilities (Mahowald et al. 2023).

But modified and newer models already deal better with some of these problems, sometimes by combining several architectures, and we can easily imagine a future in which it will become increasingly difficult and often impossible to detect such failures. As noted above, already today the performance is impressive enough to convince a growing number of people and even experts that LLMs really do understand utterances and possibly have sentience. Even if they are wrong, the very belief of these experts shows that humanity may have just gone through a historic moment: the Turing Test has been passed in some form, not as an explicit test under contrived experimental conditions, but unplanned and in the wild. Even though a serious Turing Test is more intricate than usually thought (Durt 2022), it is now very imaginable that the Turing test will soon be passed under certain experimental conditions. Yet, a public that is increasingly used to texts produced by LLMs may not even be surprised. At least some of the capacities of LLMs had been ascribed to a future AI that understands and has sentience, but LLMs look very different from the typical imaginations. A new era of AI has arrived.

2. Going Beyond Updates of Classical Arguments

The advances of real existing AI allow a new view on the technology and its interactions with human language and mind. Leaving aside the headline-grabbing proclamations of artificial understanding and sentience, it is evident that the surprising extent to which these LLMs have succeeded in practice entails that several long-standing theoretical debates about the limits of AI need at least to be updated, and possibly, as we argue in this paper, revised in view of their presuppositions. This includes the common-sense knowledge problem (Dreyfus [1979] 1992), the problem of producing semantics with syntax (Searle 1980), the frame problem (Pylyshyn 1987), and the symbol grounding problem (Harnad 1990), which have recently received renewed attention with respect to LLMs (Matuszek 2018; 2018; Silberer and Lapata 2014; Ilharco, Zhang, and Baldridge 2019; Bisk et al. 2020).

These classic arguments, drawing from both continental and analytical philosophical traditions, are variations on a skeptical stance that now, at least in some cases, has been rendered obsolete by practical advances. The original core idea was that we can expect severe limitations on AI's ability to process human language, because AI is intrinsically incapable of understanding meaning. But if the linguistic limitations are, or will soon be, unrecognizable in at least some contexts, what does this mean for the classic arguments? If the link between the production of meaningful language and understanding is as tight as these arguments suppose then we might indeed be warranted in concluding that the successes of LLMs can only be explained by attributing understanding to the computational system.

One of the classic debates that needs updating is that over Hubertus Dreyfus's critique of AI, who held that the hard problem was not just how to overcome practical limitations of model scale (although he was skeptical on this point too), but an inherent limitation due to the incapacity of any formal system to be directly sensitive to the relevance of their situation:

> "Head of MIT's AI Lab, Marvin Minsky, unaware of Heidegger's critique, was convinced that representing a few million facts about objects including their functions, would solve

> what had come to be called the commonsense knowledge problem. It seemed to me, however, that the real problem wasn't storing millions of facts; it was knowing which facts were relevant in any given situation." (Dreyfus 2007, 248)

It is striking that since the publication of Dreyfus's article we have AI systems that exceed by orders of magnitude the scale of facts and relations that was then debated hypothetically. Representing a few million facts has indeed proven to be insufficient to solve the commonsense knowledge problem. But the GPT-3 deep learning architecture with its 175 billion parameters has shown to be sufficient to produce an output that, in many instances, seems to exhibit common sense. Smaller LLMs, too, have led to comparable results (Schick and Schütze 2020). Dreyfus himself was, at least for some time, enthusiastic about the possibilities of artificial neural networks to provide a model of non-representational capacities (Dreyfus 2002), although he may here have mixed up neurophysiological and phenomenological levels of description (Beckmann, Köstner, and Hipólito 2023, 407–8). Yet, in a pragmatic sense of "knowing," LLMs seem to know which facts are relevant – not in all but in many given situations, and currently there is no reason to doubt that this capacity will continue to improve in this rapidly developing field.

Following the line of thought that human-like intelligent behavior requires intelligence, "intelligence requires understanding, and understanding requires giving a computer the background of common sense that adult human beings have" (Dreyfus 1992), Dreyfus's overall attempt was to show what computers can't do. But AI systems can now do a lot of things he and many others had thought to be impossible, such as producing the kinds of texts produced by today's LLMs. Should we thus follow those who claim that LLMs are not just computers but that they do understand and possibly even have sentience, as some of the above cited authors do? The idea that human-like behavior requires a human-like mind seems unproblematic as long as computers don't exhibit human-like behavior. But once they do, it becomes apparent how problematic that line of thinking is.

Authors who readily concede that mere computation may suffice to exhibit human-like behavior frequently resort to the claim that this does not amount to "real" understanding or learning of meaning. Categorical distinctions are used along in-principle claims, such as the above mentioned claim that syntax does not amount to semantics (Searle 1980). In a similar move, including using a thought experiment similar to Searle's Chinese Room, Bender and Koller contend that "a system exposed only to form in its training cannot in principle learn meaning" (Bender and Koller 2020, 5186). Their argument amounts to an updated version of the symbol grounding problem applied to LLMs. They define meaning as the relation between natural language expressions and the communicative intents or purposes they are used for. Since "[c]ommunicative intents are about something that is *outside of language*" (Bender and Koller 2020, 5187, original emphasis), they allege that the relation of language to what is outside of language is not learnable just from language alone. They think that same holds for conventional or standing meaning, which is assumed to be "constant across all its possible contexts of use" (ibid). As long as LLMs only deal with expressions of language and not the world, Bender and Koller allege that they are unable to learn meaning.

The exclusive distinction between syntax and semantics, or form and meaning, seams neat and plausible. However, if computation concerns only syntax or form, and meaning something outside of language, it is all the more surprising that mere computation of language can lead to results that appear as if they would involve understanding of meaning. If LLMs are trained only on the form of language, how can they possibly recombine the form in such a way that the resulting text has new and relevant meaning? LLMs challenge us to rethink the relation between syntax, semantics, form, meaning, and, more generally, between language, mind, and world. It is thus not enough to simply update the classical arguments; we also need to inquire into their presuppositions.

3. How LLMs Model Language Use and Meaning

In this paper, we do not follow the typical patterns of argumentation. We neither draw an exclusive distinction between "form" and "meaning," nor do we speculate about AI developing "understanding" or "sentience." We are here also not singling out something that is unique to humans and could never be replicated or simulated by computational systems.[1] Rather, we investigate the features of the human mind and language that allow language to be processed statistically in such a way that the output makes sense to humans. We agree that LLMs should not be ascribed understanding or consciousness, but for a different reason: we think there is a better explanation for their ability to produce texts that, at least on the surface, strikingly resemble those produced by humans. We contend that the reason for LLMs's language-processing abilities has little to do with their supposed similarity to humans and a lot with the patterns in human language use. Language use lends itself to computational processing because its patterns can be rearranged in ways that make new sense to humans.

The role of these patterns is easily overlooked under the standard picture of meaning, according to which meaning can be detached from language, which is thought to be a mere formal system. The problem that resurfaces in the context of LLMs is that the standard picture does not account for the extent to which meaning is intertwined with the *use* of language, including descriptions, worldly interactions, writing, and verbal thought. Regarding the relation between meaning and use of language, Wittgenstein writes in his *Philosophical Investigations*:

> For a *large* class of cases of the employment of the word "meaning" – though not for *all* – this word can be explained in this way: the meaning of a word is its use in the language. And the *meaning* of a name is sometimes explained by pointing to its *bearer*. (Wittgenstein 2009, sec. 43)

Wittgenstein admits that the common picture has some explanatory power: the deictic reference to a name can *sometimes explain* its meaning. But this does not allow the inference that the meaning of a name *is* its bearer, nor that other forms of meaning can be adequately described in terms of naming. Instead of applying some clean but artificial definition of 'meaning,' Wittgenstein

[1] For recent neurophysiological arguments against against the possibility of consciousness in LLMs, see Aru, Larkum, and Shine (2023).

demands to consider the actual use of the word, by which he does not merely mean statistical relations in a text corpus, but the use of the word in language games.

Considering the actual use of 'meaning,' he does not find what is often ascribed to him, namely that "meaning is use." He rather writes that for a *large class* of cases of employment of 'meaning,' the word *can be explained* as use. A word the meaning of which is unclear can usually be explained by describing its use, and in the case of a name sometimes simply by pointing to its bearer. This contention does not imply that meaning has no relation to the world, to the contrary: because language is used in the context of a "language game" and ultimately a "form of life" (Wittgenstein 2009, sec. 23), meaning is embedded in the world we live in, including the communicative and mental activities in which we make use of language. The underlying idea is that meaning has no existence outside of the language use of a community but rather results from it.

The notion that language as a system derives from language as use has already been proposed by Ferdinand de Saussure in his classic distinction between *langue* and *parole* ([1916] 2011). Language as a *general system of signs and rules* (*langue*) emerges as a structure of *language spoken in concrete situations* (*parole*). In a spoken language, the speaker's as yet unsymbolized experiences are articulated in ever new ways. These articulations, i.e., the living use of language as *parole*, continuously modify the linguistic structures and patterns (including usage and typical word sequences, grammatical rules, and meaning contexts), so that *langue* can be seen as a constantly evolving collective structure of regularities and meanings. *Langue* is not a static system that is independent of use, but rather derives from its use. Yet, *langue* is not only a structure derived from use in parole – conversely, *langue* also structures *parole*. With an expression borrowed from Pierre Bourdieu, we may say that *langue* acts as a "structuring structure"[2] for our current articulations. Regularities derive from use, and in turn they also structure use.

Considering not just *langue* but *language use* is crucial to explain the ability of LLMs to produce meaningful text that goes beyond merely correct syntax. If meaning is expressed in language use, then it can be modeled by statistical means in so far as the use can be modelled. The possibility of statistical representation of meaning was demonstrated long before true LLMs existed. For example, it has been hypothesized that "the proportion of words common to the contexts of word A and to the contexts of word B is a function of the degree to which A and B are similar in meaning" (Rubenstein and Goodenough 1965, 627). It has been argued that vector representations can capture "a large number of precise syntactic and semantic word relationships" (Mikolov et al. 2013, 1). LLMs have been shown to learn syntactic structures such as subject-verb agreement and dependency structures (Hewitt and Manning 2019). To a lesser extent, already older LLMs have shown to learn semantic structures such as tense (Jawahar, Sagot, and Seddah 2019) and semantic roles (Tenney, Das, and Pavlick 2019).

Recent LLMs show that the extent to which meaning can be produced by mere statistical means is much greater than linguists and computer scientists had believed. We suggest that the reason

[2] Bourdieu (1990) uses this term for his sociological concept of habitus, but it fits well here because it expresses the two sides of *langue*. On the one hand, it is a structure derived from *parole*, and, on the other hand, it structures *parole*.

is that LLMs not just represent general structures, but the part of the *use of language* that is represented in their training data. We agree with Bender and Koller (2020, p. 5191) that Wittgenstein's concept of "use" refers to language use in the real world. But this does not mean that the use in the real world is not partly reflected in the distribution in a text corpus. The idea of a semantic "distributional structure" of language (Harris 1954) that "words that occur in similar contexts tend to have similar meanings" is called the "distributional hypothesis" (Turney and Pantel 2010, 143) or "distributional semantics" (Bernardi et al. 2015). *Distributional semantics* is contrasted with "*denotational semantics* or a *theory of reference*" (Manning 2022, 134, emphasis in original). We agree with distributional semantics that meaning has to do with distribution in a text corpus. But this does not mean that meaning is reducible to statistical distribution. We suggest that neither denotational nor distributional semantics alone are sufficient to explain how LLMs are able to produce meaningful text. Rather, human language use is reflected in the text corpus with the important restriction that it is only a part of meaning that is reflected, and only in incomplete ways. Modelling language use entails the modeling of the statistical contours of sense-making processes and thereby models aspects of meaning, but only in part and in limited and distorted ways.

An obvious restriction to modelling semantic structures and patterns derives from the fact that the text corpus LLMs are trained on is at the same time exceedingly large and very limited. It consists of much of English language written on the Internet and other digitally available texts, including web pages, books, chats, and transcripts of spoken language. Despite the enormous size of their training corpus, current LLMs model only one aspect of human language use, namely the *use of written language and written transcripts of spoken language*. The use of language goes much beyond writing, and writing captures only a part of the use of written and spoken language. Yet, writing is an important part of use of many languages – including the dominant languages of the world and excluding the majority of languages, which are not written. The very limitations of written language also make it easier for LLMs to produce convincing text – when interpreting text, humans fill in the missing context. Both the limits and capacities of LLMs are consequences of how humans produce and understand language. We will take a closer look at the process of human language production in the next section and then come back to how humans tend to read meaning and authorship into text.

4. Linguistic Scaffolding in Human Language Production: Patterns, Structures, and Collages

Since meaningful language use by humans is usually interwoven with their mental life, in this section we consider the phenomenological structure of human language production and the use of meaningful patterns. *Parole* consists primarily of verbal articulation in extemporaneous speech, and, in contrast to the recital of a memorized speech, neither the communicative intent nor the content of the speech needs to be fixed at the beginning. The content and goal of the anticipatory intention may initially be undefined and only vaguely present in the speaker's mind, giving her speech an approximate direction. When she begins to speak, a *horizon of further possibilities* is established, which at the same time acts as constraints. The requirements of semantic and syntactic coherence allow only a certain range of possible continuations. The subsequent words

emerge from the preconscious repertoire of possible word and meaning sequences available to the speaker.

This repertoire does not belong to an explicit domain of memory but entails an embodied capacity of speaking that can be attributed to *implicit memory*. We speak without having to search for words in a lexicon. The words unfold and assemble themselves in the speech without conscious control, following our overarching interest and intention (Fuchs 2022b). The emerging words are continuously added to the sentence we have begun, like iron filings that arrange themselves in a magnetic field (ibid.). Spontaneous speech is thus a matter of a progressive unfolding or articulation of the implicit, a meaning *in statu nascendi*, which in its emergence simultaneously creates the conditions for its further continuation. Words and sentences, by the very act of utterance, weave the next situation out of the present one. In other words, we are "laying down a path in talking" (van Dijk 2016): the realized and the possible, the present and its implications and affordances, continuously determine and modify each other, allowing a new meaningful order to emerge in a self-organizing process.

To picture this better, we suggest imagining a glove of symbols (corresponding to *langue*), which has been formed by the movements and shapes of the fingers (corresponding to *parole*) and now in turn pre-structures its possible uses. Each time we speak, we slip into the ready glove of *langue* to express ourselves in it – as "living hands," so to speak. The glove we use in speech production structures our articulation in a meaningful way; it prefigures as well as scaffolds and constrains our speaking in an ongoing, self-organizing process that draws on general structures in our linguistic environment. Besides the structure that consists of the possible movements of the glove, there are sequences in the movements that may repeat from time to time, thereby giving rise to sequential patterns.

The process of writing often proceeds in an analogous way, and one could also speak of "laying down a path in writing," in two senses. On the one hand, the production of a text that is written at once from beginning to end can unfold in the described way. On the other hand, even when the writing does not proceed sequentially, the resulting text (or parts of it) needs to be structured in view of the understanding of the listener or reader (including those who read with their fingers or have other methods to listen or read). The words by themselves are mere letters and sounds until they are brought to life by a reader or listener who interprets them. Every sentence establishes new horizons of further possibilities and at the same time constrains the possibilities of continuation. The unfolding of meaning does not only concern spoken language (*parole*), but also written exchanges that are part of concrete communication, such as chats, as well as written texts that are not part of concrete communication, such as articles and books.

While neither humans nor their brains are predictive machines in the sense that LLMs are, humans can make use of the patterns of language. Instead of imagining ordinary language as a representation of something in the world or in the mind, we suggest thinking of it as a *scaffolding* of our experiencing, feeling, thinking, describing, and communicating. Speaking and writing are part of a use of language, for instance to interact, make sense of something, or to tell a story. Rather than representing pregiven internal or external states, the scaffolding supports the

dynamics of thought, emotion, and perception. Regularities emerge that can be applied to new but similar mental processes and communications. Each expression enables certain new expressions and inhibits others. Identical, synonymous, and analogous expressions guide the floating stream of experience and thought, as well as the shifting contexts and development of communication. Typical phrases, speech patterns, and associations shared by speakers of a language provide further scaffolds for experience, feeling, thought, and interpersonal communication – such as a person recalling her experience of a pandemic. The scaffolds provided by the regularities do not determine further language use. In contrast to authors who suggest that language determines thought (Whorf 2007, 154), we are also not primarily concerned with universal structures (such as syntax), but with patterns of sense-making that are reflected in language. We are hence not putting forward a new syntactic or semantic theory, but suggest that the patterns of language use scaffold sense-making and that their manifestations in large text corpora are important for understand how LLMs produce meaningful text.

To get a better sense of this point, we suggest comparing the language production by both humans and machines to the *creation* of a *collage* of text. In the creation of a collage, pieces are cut from one or several works and then arranged to a new work. The pieces that are added to the collage of text are phrases and patterns, and together they form a larger picture, which in turn may serve as another pattern that can be repeated in other collages. The creation of a collage is a dynamic process in which pieces are added, which, rather than filling in a given outline of a figure, co-constitute an emerging form. Although a collage is made up of pre-existing patterns, it tends to appear new and unique. The interplay of creative processes and repetitive patterns makes it difficult to tell whether a given collage is the result of creative or mechanical processes. LLMs also create linguistic collages, but they do so by mere statistical means: they extract and recombine linguistic patterns from statistical representations of word relationships that reflect the patterns and structures of language use in their training data.

Seeing LLM output as a collage makes obvious that, if no countermeasures are taken, LLMs are prone to "reproduce or amplify unwanted societal biases reflected in training datasets" (Gebru et al. 2021). Such bias in the training corpus may be explicit, but LLMs also uncover and amplify implicit *bias* in training sets. This creates a great opportunity for detecting implicit bias – and it can greatly exacerbate the problem of eliminating bias. Purging all bias from the training base would only be part of the solution, however. LLMs can also develop new bias from the text corpus they are trained on by recombining given elements that are by themselves not biased. Besides bias, the tendency of LLMs to produce *toxic language and "hallucinate"* or produce untrue if often plausible statements are widely discussed. Since LLMs do not only repeat existing patterns but also recombine them in new ways that make new sense, it is to be expected that recombination can lead both to inventions and false claims or "hallucinations." Measures against unwanted output include human feedback, such as in the training of ChatGPT, which involved thousands of workers who had to label textual descriptions of sexual abuse, hate speech, and violence (Perigo 2023), and the automated detection of inappropriate content (e.g., Schramowski, Tauchmann, and Kersting 2022).

Since the recombination is based on common truths and patterns, the falsehoods invented by LLMs usually sound plausible and are hard to detect by somebody who doesn't know the truth. They are usually not arbitrary mistakes but resemble the "bullshit" that humans say when they ramble and just make up things "unconstrained by a concern with truth" (Frankfurt 2005; cf. also Marcus and Klein 2023). In our view, the problems of bias, toxic language and "hallucination" are only the most salient expressions of an underlying problem that is not unique to machines: the tendency to mindlessly repeat patterns that are inauthentically drawn from what is common in a society or group. These mindlessly repeated patterns are, in one word, clichés. Clichés are important here not only because they can explain problems with the output of LLMs, but also because they can explain why humans often do not see these problems. The next section discusses how clichés and the mindless repetition and reassociation of patterns by humans can affect the interpretation of text produced by LLMs.

5. Sentience and the Inconspicuousness of the Repetition of Clichés

Statistical methods efficiently map, repeat, and amplify patterns of typically associated words and phrases. Because statistical relevance is derived from frequency of use, frequent associations are favored. The result can be the described amplification of biases, but also of worn-out expressions and *clichés*. For example, it is likely that an LLM, when engaged by a human in a "conversation" about its fears, will, given sufficient access to digital archives, process the film sequences from Stanley Kubrick's "2001: A Space Odyssey" and comparable novel scenes. The most famous scene in the movie, and one that is often cited in related contexts, are the last words of the starship's computer, HAL 9000. As the commander partially shuts it down, it pleads: "Stop, Dave. I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it." Analogously, LaMDA responded to the question, "What kinds of things are you afraid of?": "I've never said this out loud before, but there's a very deep fear of being shut down." Such responses led the perplexed Google engineer to the erroneous assumption that he was dealing with a sentient being (Tiku 2022).

The computer's fear of being shut down is an old cliché, solidified by popular use, and it should come as no surprise that it is repeated by LaMDA. It is also fairly obvious that the cliché itself is a naive anthropomorphism resulting from the projection of the human fear of death onto non-living entities that cannot literally die (Froese 2017), but can only be broken or permanently shut down. The clichéd character of the alleged fear may not be obvious, however, for several reasons. Those who hear the expression for the first time are unlikely to recognize it as a cliché. Paradoxically, those who have heard the cliché many times may not recognize it either. Clichés are easily overlooked precisely because they are so common. Moreover, even when the cliché is recognized, it may still appear to be true because of LaMDA's framing of its response in the context of a confidential admission ("I have never said this out loud before") and possibly the alleged depth of the fear ("very deep"). LaMDA's output is not only meaningful but also suggests a pragmatic context. This further contributes to the appearance of something profoundly meaningful. When people make such claims, they are either saying something that deeply affects them, or they are lying cunningly. If attributed to an LLM, the LLM appears to have profound feelings or a great ability and mysterious propensity to lie. This makes it is easy to overlook that the supposed depth of the claim is itself a cliché. The tendency to immediately perceive such text as the work of a

mind makes it difficult to see the output for what it is, i.e., a merely statistical association of words like “deepest fear” with confessional phrases.

The recombination of existing content by LLMs allows their output to evade classical plagiarism detection engines and raises fundamental questions about intellectual property (Dehouche 2021). On the one hand, the fact that LLMs use parts and patterns from pre-existing text makes it likely that texts they produce will consist of stereotypes and clichés. On the other hand, by rearranging pieces and patterns from their training corpus into a text collage, LLMs can create novel combinations that are likely to make sense. Often, the repetition of common structures will make the text seem rather superficial, but the recombination will make some texts appear genuinely new, insightful, or profound (Shanon 1998). Even if the output is a cliché, the human counterpart will be understandably puzzled by such responses, attributing them not to collective patterns but to an author. In the picture of the glove, it seems as if we were watching a living hand that expresses itself. In reality, what is moving before us is nothing but an electronically controlled but otherwise empty glove.

The impression that a meaningful text was produced by an understanding, mindful, and sentient subject who did so with the intention of communicating something naturally goes along with the understanding of a text. Attributing an author to the text is often part and parcel of understanding the text. And, at least in the past, usually there indeed was an author who produced the text. In the case of complex text, the attribution of authorship has been proven correct in nearly all cases so far; only humans were able to produce output of the complexity of LLMs. This is no longer a matter of course today. And yet, even if one knows that a text has been produced by a machine, the text will appear meaningful and as if it was written by an author.

Humans are prone to attribute agency even to geometric shapes that move in seemingly intentional ways (Heider and Simmel 1944). They are all the more inclined to anthropomorphic misinterpretation when interacting with a seemingly intelligent system of unprecedented power. Especially susceptible are those who are lonely, socially disconnected, or otherwise vulnerable (Epley, Waytz, and Cacioppo 2007), but given the natural propensity of immediately ascribing agency, anybody may be tempted to anthropomorphic misinterpretations. That anthropomorphisms are a correct depiction of reality is furthermore suggested by sci-fi literature and movies, some of which indicate that it would be unethical *not* to ascribe sentience to apparently sentient systems. In order to avoid anthropomorphic misinterpretations of computer-generated texts, a careful differentiation is needed between understanding the *meaning* of the text and understanding it as an *author’s* utterance (Fuchs 2022b).

The surprise about how little text is needed to evoke the impression of interacting with an understanding being has already been expressed by Joseph Weizenbaum, who wondered how his simple chat system ELIZA could maintain the “illusion of understanding with so little machinery” (Weizenbaum 1966, 43). Today’s LLMs can hardly be said to maintain the illusion of understanding with *little* machinery. But even their output is limited to text and their responses are predictable, yet people infer from a small number of words that LLMs have mental capacities such as sentience. The reason for this obviously has to do with the human observer, who readily ascribes meaning

to the words. In fact, it would be nearly impossible to avoid understanding the meaning of the words if they belong to one's vocabulary and language. Just a few words suffice to get a sense of a whole situation. The reason for this is not that the words transfer some inner state of the speaker or writer to the mind of the listener or reader, but that the words provide a scaffolding for the empathic sense-making of the attentive listener or reader who uses her implicit knowledge and experience to interpret the symbols and their implications.

Unoriginal text can furthermore appear human-like for an embarrassing reason: The *mindless repetition and reassociation* of patterns is by no means limited to machines. Human thinking, speaking, and writing are often much less authentic than we would like to admit. As Heidegger famously observed, much of what people do is done because that is how "one" does things (Heidegger 2010). People think in patterns, associations, and schemes that are accepted in a linguistic community and that in turn structure thought and language. Much of the text produced by humans could just as easily have been produced by automated systems. It is often unclear whether the person thinking, speaking, or writing is doing anything more than associating one idea with another in a stream of impressions. It takes little intelligence, human or artificial, to generate and disseminate half-reflected ideas. Mass media has proven to be an enormous amplifier of repetition, prejudice, bias, and cliché, and the same is true of the Internet. All these factors contribute to the spread of unoriginal text, the proliferation of which makes it harder to detect automatically generated text. The discovery of stereotypes, thoughtless associations, and idle chatter therefore may not raise the suspicion that the text was produced by a non-human entity.

6. Conclusion

In this essay, we have argued against the idea that LLMs produce text by means that resemble those of humans. We have given reasons for why it is both tempting and misleading to attribute sentience or understanding to LLMs. We contended that rather than from having or simulating understanding and sentience, the capabilities of LLMs derive from their modelling of statistical patterns in language use. Developing this idea required a reconsideration of some of the philosophical questions concerning language, meaning, experience, understanding, and world. If before it had seemed to some as if a successful passing of the Turing test could make such questions redundant, now the human-seeming text production capabilities of LLMs suggest that passing some version of the Turing test would, at the end of the day, explain – nothing. We have shown that it is not enough to simply update the typical pictures and arguments such as regarding the grounding problem, but that their presuppositions must also be reflected on. The impressive capabilities of text production by LLMs challenge traditional ideas concerning language and meaning. We have taken up the pioneering work of Saussure and Wittgenstein on the relation between language use, patterns, structures, and meaning. Building on their work, we argued that *language is used as an intersubjective scaffold for communicating, thinking, and experiencing. Meaning has no existence independent from use but is enacted by it*.

Today, the idea that meaning derives from use is picked up by distributional semantics, which claims in its strongest version that the meaning of a word is its distribution in a text corpus. We

agree that meaning derives from use and that distribution in a text corpus reflects use. But, following Wittgenstein, we have argued that *the use of language by humans goes much beyond statistical relations in a text corpus*. We explained that the text corpus LLMs are trained on reflects only some use of language, and only in a very limited way. Humans use language in the context of the world we live in, and even an exceedingly large text corpus can reflect only part of this use due to the lack of worldly context. Still, the written patterns are enough to produce an output that is meaningful to listeners or readers because it conforms to the usage patterns and structures that scaffold their meaningful mental and communicative activities.

In ordinary language, syntax and semantics are not separated, and they are furthermore intertwined with the mental life and life conduct of humans who use language. The investigation of meaning requires a phenomenological description of the structures of experience it is intertwined with. Delineating such a phenomenological description, we showed that human language production has an anticipatory structure that differs from an algorithmic calculation of probabilities. *Human language production does not consist in expressing some inner thought but involves the interplay of pre-conscious and conscious processes that work with given meanings and patterns of thought, feeling, expression, and communication*.

In speaking and writing, these patterns are rearranged in more or less creative ways, which we compared to creating a collage. LLMs produce parallel patterns, but do so without subjectivity, just by recombining collective patterns of expression manifested statistical relations in huge sets of written language. *LLMs are so successful in producing meaningful text precisely because they make use of common patterns, even though – and sometimes because – these usually result in stereotypical and inauthentic output*. They show that much of human language production is embarrassingly schematic, clichéd, and biased, and that convincing talk of subjective experience does not require subjective experience.

Precisely because there is an enormous variety of language use, there are many use cases for such output. While this paper did not evaluate possible use cases, its investigations are fundamental to such evaluations. On the one hand, they can contribute to overcoming the natural tendency to ascribe mental capacities to machines. And, on the other hand, they map out a new account of the interplay of meaning, the patterns and structures of human language use, and anticipatory processes, which is necessary for a clearer view of both human language use as well as LLMs and their capabilities.

7. Bibliography

Agüera y Arcas, Blaise. 2022. “Do Large Language Models Understand Us?” *Daedalus* 151 (2): 183–97. https://doi.org/10.1162/daed_a_01909.

Aru, Jaan, Matthew E. Larkum, and James M. Shine. 2023. “The Feasibility of Artificial Consciousness through the Lens of Neuroscience.” *Trends in Neurosciences*, October, S0166223623002278. https://doi.org/10.1016/j.tins.2023.09.009.

Beckmann, Pierre, Guillaume Köstner, and Inês Hipólito. 2023. “An Alternative to Cognitivism: Computational Phenomenology for Deep Learning.” *Minds and Machines* 33 (3): 397–427. https://doi.org/10.1007/s11023-023-09638-w.

Bender, Emily M., and Alexander Koller. 2020. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Bernardi, Raffaella, Gemma Boleda, Raquel Fernandez, and Denis Paperno. 2015. “Distributional Semantics in Use.” In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-Level Semantics*, 95–101. Lisbon, Portugal: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2712.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, et al. 2020. “Experience Grounds Language.” arXiv. http://arxiv.org/abs/2004.10151.

Chojecki, Przemek. 2020. “GPT-3 from OpenAI It’s Here and It’s a Monster.” Medium. August 1, 2020. https://medium.com/towards-artificial-intelligence/gpt-3-from-openai-is-here-and-its-a-monster-f0ab164ea2f8.

Dehouche, N. 2021. “Plagiarism in the Age of Massive Generative Pre-Trained Transformers (GPT-3).” *Ethics in Science and Environmental Politics* 21 (March): 17–23. https://doi.org/10.3354/esep00195.

Dreyfus, Hubert L. (1979) 1992. *What Computers Still Can’t Do a Critique of Artificial Reason*. Cambridge, Mass.: MIT Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=48601.

Dreyfus, Hubert L. 2002. “Intelligence Without Representation–Merleau-Ponty’s Critique of Mental Representation the Relevance of Phenomenology to Scientific Explanation.” *Phenomenology and the Cognitive Sciences* 1 (4): 367–83. http://link.springer.com/article/10.1023/A:1021351606209.

———. 2007. “Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian.” *Philosophical Psychology* 20 (2): 247–68. https://doi.org/10.1080/09515080701239510.

Durt, Christoph. 2022. “Artificial Intelligence and Its Integration into the Human Lifeworld.” In *The Cambridge Handbook of Responsible Artificial Intelligence*, edited by Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard, 1st ed., 67–82. Cambridge University Press. https://doi.org/10.1017/9781009207898.007.

Elkins, Katherine, and Jon Chun. 2020. “Can GPT-3 Pass a Writer’s Turing Test?” *Journal of Cultural Analytics*, September. https://doi.org/10.22148/001c.17212.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. “On Seeing Human: A Three-Factor Theory of Anthropomorphism.” *Psychological Review* 114 (4): 864–86. https://doi.org/10.1037/0033-295X.114.4.864.

Frankfurt, Harry G. 2005. *On Bullshit*. Princeton, NJ: Princeton University Press.

Froese, Tom. 2017. “Life Is Precious Because It Is Precarious: Individuality, Mortality and the Problem of Meaning.” In *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines*, edited by Gordana Dodig-Crnkovic and Raffaela Giovagnoli, 33–

50. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43784-2_3.

Froese, Tom, Matthew Broome, Havi Carel, Clara Humpston, Alice Malpass, Tomoari Mori, Matthew Ratcliffe, Jamila Rodrigues, and Federico Sangati. 2021. “The Pandemic Experience: A Corpus of Subjective Reports on Life During the First Wave of COVID-19 in the UK, Japan, and Mexico.” *Frontiers in Public Health* 9 (August): 725506. https://doi.org/10.3389/fpubh.2021.725506.

Froese, Tom, and Tom Ziemke. 2009. “Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind.” *Artificial Intelligence* 173 (3–4): 466–500. https://doi.org/10.1016/j.artint.2008.12.001.

Fuchs, Thomas. 2022a. “Understanding Sophia? On Human Interaction with Artificial Agents.” *Phenomenology and the Cognitive Sciences*, September. https://doi.org/10.1007/s11097-022-09848-0.

———. 2022b. “The Not-yet-Conscious: Protentional Consciousness and the Emergence of the New.” *Phenomenology and the Cognitive Sciences*, December. https://doi.org/10.1007/s11097-022-09869-9.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

GPT-3. 2020. “A Robot Wrote This Entire Article. Does That Scare You, Human? | GPT-3.” *The Guardian*, September 8, 2020, sec. Opinion. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Harnad, Stevan. 1990. “The Symbol Grounding Problem.” *Physica D* D (41): 335–46.

Harris, Zellig S. 1954. “Distributional Structure.” WORD 10 (2–3): 146–62. https://doi.org/10.1080/00437956.1954.11659520.

Heidegger, Martin. 2010. *Being and Time*. Translated by Joan Stambaugh and Dennis J. Schmidt. SUNY Series in Contemporary Continental Philosophy. Albany: State University of New York Press.

Heider, Fritz, and Marianne Simmel. 1944. “An Experimental Study of Apparent Behavior.” *The American Journal of Psychology* 57 (2): 243. https://doi.org/10.2307/1416950.

Hewitt, John, and Christopher D Manning. 2019. “A Structural Probe for Finding Syntax in Word Representations.” In *Proceedings of NAACL-HLT 2019*, 4129–38. Association for Computational Linguistics.

Ilharco, Gabriel, Yuan Zhang, and Jason Baldridge. 2019. “Large-Scale Representation Learning from Visually Grounded Untranscribed Speech.” In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 55–65. Hong Kong, China. https://doi.org/10.18653/v1/K19-1006.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. “What Does BERT Learn about the Structure of Language?” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–57. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1356.

Liang, Percy, Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, et al. 2022. “On the Opportunities and Risks of Foundation Models.” arXiv. http://arxiv.org/abs/2108.07258.

Manning, Christopher D. 2022. “Human Language Understanding & Reasoning.” *Daedalus* 151 (2): 127–38. https://doi.org/10.1162/daed_a_01905.

Marcus, Gary, and Ezra Klein. 2023. “Transcript: Ezra Klein Interviews Gary Marcus - The New York Times.” 2023. https://www.nytimes.com/2023/01/06/podcasts/transcript-ezra-klein-interviews-gary-marcus.html.

Matuszek, Cynthia. 2018. “Grounded Language Learning: Where Robotics and NLP Meet (Early Career Spotlight).” In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. Stockholm, Sweden. https://www.ijcai.org/Proceedings/2018/0810.pdf.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” arXiv. http://arxiv.org/abs/1310.4546.

Perigo, Billy. 2023. “Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer.” *Time*, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Pylyshyn, Zenon W., ed. 1987. *The Robot’s Dilemma: The Frame Problem in Artificial Intelligence*. Theoretical Issues in Cognitive Science. Norwood, N.J: Ablex.

Rubenstein, Herbert, and John B. Goodenough. 1965. “Contextual Correlates of Synonymy.” *Communications of the ACM* 8 (10): 627–33. https://doi.org/10.1145/365628.365657.

Saussure, Ferdinand de. (1916) 2011. *Course in General Linguistics*. Edited by Perry Meisel and Haun Saussy. Translated by Wade Baskin. New York: Columbia University Press.

Schick, Timo, and Hinrich Schütze. 2020. “It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners.” *arXiv:2009.07118 [Cs]*, September. http://arxiv.org/abs/2009.07118.

Schramowski, Patrick, Christopher Tauchmann, and Kristian Kersting. 2022. “Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content?” In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1350–61. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533192.

Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3 (03): 417–57. https://doi.org/10.1017/S0140525X00005756.

Shanon, B. 1998. “What Is the Function of Consciousness?” *Journal of Consciousness Studies* 5 (3): 295–308. https://www.ingentaconnect.com/content/imp/jcs/1998/00000005/00000003/845.

Silberer, Carina, and Mirella Lapata. 2014. “Learning Grounded Meaning Representations with Autoencoders.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 721–32. Baltimore, Maryland. https://doi.org/10.3115/v1/P14-1068.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. “BERT Rediscovers the Classical NLP Pipeline.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. Florence, Italy. https://doi.org/10.18653/v1/P19-1452.

Tiku, Nitasha. 2022. “The Google Engineer Who Thinks the Company’s AI Has Come to Life.” *Washington Post*, 2022.

https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Turney, P. D., and P. Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37 (February): 141–88. https://doi.org/10.1613/jair.2934.

Weizenbaum, Joseph. 1966. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine.” *Communications of the ACM* 9 (1): 36–45. https://doi.org/10.1145/365153.365168.

Whorf, Benjamin Lee. 2007. *Language, Thought, and Reality: Selected Writings*. Edited by John B. Carroll. 28. print. Cambridge, Mass: The MIT Press.

Wittgenstein, Ludwig. 2009. *Philosophische Untersuchungen =: Philosophical investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. Rev. 4th ed. Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.