https://doi.org/10.48417/technolang.2024.02.04
Research article

Do Language Models Communicate? Communicative Intent and Reference from a Derridean Perspective

Rebeca Perez Leon (✉)
Facultad de Estudios Superiores Acatlán, Universidad Nacional Autónoma de México, Avenida Jardines de San Mateo s/n, Sta. Cruz Acatlan, Naucalpan de Juárez, 53150 Estado de México, Mexico
drrpleon@gmail.com

Abstract

This paper assesses the arguments of Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Margaret Mitchell in the influential article "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" These arguments disputed that Language Models (LM) can communicate and understand. In particular, I discuss the argument that LMs cannot communicate because their linguistic productions lack communicative intent and are not based on the real world or a model of the real world, which the authors regard as conditions for the possibility of communication and understanding. I argue that the authors' view of communication and understanding is too restrictive and cannot account for vast instances of communication, not only human-to-human communication but also communications between humans and other entities. More concretely, I maintain that communicative intent is a possible but not necessary condition for communication and understanding, as it is oftentimes absent or unreliable. Communication need not be grounded in the real world in the sense of needing to refer to objects or state of affairs in the real world, because communication can very well be about hypothetical or unreal worlds and object. Drawing on Derrida's philosophy, I elaborate alternative concepts of communication as the transmission of an operation of demotivation and overwhelming of interpretations with differential forces, and of understanding as the best guess or best interpretation. Based on these concepts, the paper argues that LMs could be said to communicate and understand.

Keywords: Language Model; Stochastic Parrot; Communication; NLU; ChatGPT; Derrida

Citation: Perez Leon, R. (2024). Do Language Models Communicate? Communicative Intent and Reference from a Derridean Perspective. *Technology and Language, 5*(2), 40-56. https://doi.org/10.48417/technolang.2024.02.04

УДК 1: 004.8
https://doi.org/10.48417/technolang.2024.02.04
Научная статья

Общаются ли языковые модели? Коммуникативное намерение и референт с точки зрения Дерриды

Ребека Перес Леон (✉)
Национальный автономный университет Мексики, Авенида-Гардинес-де-Сан-Матео, Наукальпан-де-Хуарес, 53150 Мехико, Мексика
drrpleon@gmail.com

Аннотация
В этой статье оцениваются аргументы Эмили М. Бендер, Тимнит Гебру, Анджелины Макмиллан-Мейджор и Маргарет Митчелл в влиятельной статье “Об опасностях стохастических попугаев: Могут ли языковые модели быть слишком большими?” Эти аргументы ставили под сомнение тот факт, что языковые модели (LM) могут общаться и понимать. В частности, я обсуждаю аргумент о том, что языковые модели не могут быть коммуникативными, потому что их лингвистические произведения лишены коммуникативной направленности и не основаны на реальном мире или модели реального мира, которые авторы рассматривают как условия возможности общения и понимания. Я утверждаю, что авторский взгляд на коммуникацию и понимание является слишком ограничительным и не может охватить обширные случаи коммуникации, не только коммуникации между людьми, но и коммуникации между людьми и другими сущностями. Более конкретно, я утверждаю, что коммуникативное намерение является возможным, но не необходимым условием для общения и понимания, поскольку оно часто отсутствует или ненадежно. Коммуникация не обязательно должна быть основана на реальном мире в том смысле, что она должна ссылаться на объекты или положение дел в реальном мире, потому что коммуникация вполне может касаться гипотетических или нереальных миров и объектов. Опираясь на философию Дерриды, я разрабатываю альтернативные концепции коммуникации как передачи операции демотивации и подавления интерпретаций различными силами, а также понимания как наилучшей догадки или наилучшей интерпретации. Основываясь на этих концепциях, в статье утверждается, что можно сказать, что языковые модели, передают информацию и понимают.

Ключевые слова: Языковая модель; Общение; NLU; ChatGPT; Большие Языковые модели

Для цитирования: Perez Leon, R. (2024). Do Language Models Communicate? Communicative Intent and Referent from a Derridean Perspective // Technology and Language. № 5(2). P. 40-56. https://doi.org/10.48417/technolang.2024.02.04

INTRODUCTION

In 2021, Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Margaret Mitchell published a paper titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" highlighting some of the ethical, political, environmental, financial and social problems of training LMs with enormous amounts of indiscriminate data and using them for numerous purposes. Among these various problems, the authors identified the mischaracterization AI developers make of LMs based on the performance of "LMs [… in] tasks intended to test for natural language understanding (NLU)" (Bender et al., 2021, p. 615). Basically, developers have tested LMs in different evaluations intended to measure "language understanding and/or commonsense reasoning" (p. 615) such as the General Language Understanding Evaluation (GLUE), the Stanford Question Answering Datasets (SQuAD) and the Situations with Adversarial Generations (SWAG). Significantly, LMs such as BERT have obtained remarkably high scores in these tests leading developers to characterise them as "'[…] language understanding systems'" (615) and their operation as "machine comprehension" (Bender and Koller, 202, p. 5185). The authors, however, emphatically claim that "no actual language understanding is taking place in LM-driven approaches to these tasks" (Bender et al., 2021, p. 615). Their claims have resulted in a lively discussion, fired up by the appearance of ChatGPT in late 2023 fired up and spilling over to disciplines outside AI and machine learning,[1] making the paper a remarkably influential criticism of LMs with the added feat of having coined the term 'stochastic parrot'[2] to critically refer to LMs more generally.

In a previous paper, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (Bender and Koller, 2020), Emily M. Bender and Alexander Koller clarify the key concepts and arguments at the basis of the emphatic claim that LMs cannot understand language. Within the context of communicative exchanges in particular, Bender and Koller affirm that LMs cannot communicate because their linguistic productions lack communicative intent and reference to the real world, and cannot understand because they are trained to manipulate the form but not the meaning of language. In this paper, I first unpack and clarify Bender et al.'s concepts of communication and understanding in communication. Drawing on Jacques Derrida's philosophy, I then raise a number of questions regarding the necessity of communicative intent and reference to the real world underpinning their concept of communication, and the idea of understanding as retrieving meanings. Next, taking recourse again to Derrida's deconstruction I advance a concept of communication that retains the idea of transporting, transmitting, and production of signs while dispensing with the necessity of communicative intent and reference to the real world. This less restrictive concept does allow answering the question of whether LMs communicate and understand in the affirmative.

[1] A quick look at the statistics of this paper on Google Scholar shows that it has been cited over 3500 times in disciplines ranging from legal studies to education studies, from linguistics to environmental studies.

[2] The term 'stochastic parrot' was the AI-related word of the year 2023: https://americandialect.org/wp-content/uploads/2024/01/2023-Word-of-the-Year-PRESS-RELEASE.pdf

LANGUAGE MODELS AS STOCHASTIC PARROTS

The authors of "On the Dangers" outline three requirements of "human language use" without which communication could not take place: a) it happens between individuals who hold beliefs and other mental states –that is, they have attitudes towards propositions, for example, holding them true or not, desiring they were the case or not, etc.; b) they "share common ground and are mutually aware of that sharing (and its extent)" (Bender et al., 2021, p. 616); c) they use language to "convey" a "communicative intent" (p. 616). Condition (a) presupposes fully fledged language users, who feel at home in the business of linguistic exchange, and have had many such exchanges whereby they have formed a more or less coherent system of beliefs and other mental states. Condition (b) suggests that those partaking in the communicative exchange share a common world, namely "the real world the speaker and listener inhabit together" "against" which they can test the "truth" of their "interpretations" (Bender and Koller, 2020, p. 5187). The authors maintain, furthermore, that those participants in the communicative exchange are aware that they share a real world and the extent to which they share it. Finally, condition c) presupposes that individuals engage in communicative exchanges "for a purpose," namely, "in order to achieve some *communicative intent*" (p. 5187). Communicative intent can be defined generally as wanting one's speech or writing to do something, for example, inform, request, make another laugh, release frustration or anger, among many others. People's intentions can be achieved through numerous means, and one of them is to use language in either spoken or written form. Speakers and writers choose some particular strings of words, then, that they think will allow them to communicate what they intend to communicate. In this picture, the linguistic articulation of the communicative intent carries 'meaning' – what the speaker means to say. That is, what makes a particular string of words meaningful is that the string was chosen with the expectation that it will do what the speaker or writer intends to do.

Understanding in communication, in turn, consists in "the process of retrieving [communicative intents] given [some strings of words]" (Bender and Koller, 202, p. 5187). Such retrieval requires the "ability to recognise interlocutors' beliefs […] and intentions […] within context" (Bender et al., 2021, p. 616). In the process of understanding, the speaker/writer and interpreter are busy assigning words with meanings and attributing beliefs and other mental states to each other, and correcting these assignments and attributions until they both appear to 'get the other' and behave as expected in response. "Human-human communication", they continue, "is a jointly constructed activity" (p. 616). This holds true not only for spoken language where speaker and interpreted are co-present, but also for written language where even "if we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us and use this in interpreting their words" (p. 616).

Based on this normative framework, Bender et al. answer the question of whether LMs can understand language and communicate, which, given its content, inevitably leads to the foregone conclusion that they do not. Firstly, LMs' linguistic production is not meaningful because the strings of words they produce convey no communicative intent. LMs, the authors state, "only have success in tasks that can be approached by

manipulating linguistic form" (Bender et al., 2021, p. 610) understood as "any observable realisation of language: marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators" (Bender and Koller, 2020, p. 5187). The authors oppose form to meaning, which expresses itself in the linguistic articulation of a communicative intent. Not having communicative intent means that LMs do not *want* to do something with the strings of words they produce; rather, their linguistic outcomes are based on certain probabilistic operations. To clarify this, the authors explain how LMs produce the strings of words they do. To start with, the term *language model* (LM) "refer[s] to systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context" (Bender et al., 2021, p. 611). N-gram models' predictions were simpler because they were restricted by number of characters, directionality and horizontality of cues: "traditional n-gram LMs can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer)" (p. 616). Coming after n-gram models, transformer models do not have previous restrictions, and are able to articulate language with impressive naturalness, "produc[ing] text that is seemingly not only fluent but also coherent even over paragraphs" (p. 616). LMs' predictions are based on the data they are 'trained' with, allowing them "to perform apparently meaning-manipulation tasks such as summarisation, question answering, and the like" (p. 612) with notable success as they have excelled in numerous language understanding and common sense reasoning tests (p. 615). These successes, the authors insist, are technical, probabilistic rather than communicative successes. Bender and Köller affirm, "far from doing the 'reasoning ostensibly required to complete the tasks, they [are] instead simply more effective at leveraging artefacts in the data" (Bender and Koller, 2020, p. 5186). In other words, they just shuffle data without comprehending what it is that they are shuffling or that they are shuffling it. This is shown in the fact that when they are trained with deliberately opposing data that contradict or negate some of the data they already have, their performance "falls to significantly below chance" (p. 5186).

Secondly, LMs' linguistic productions are "not grounded on […] any model of the world, or any model of the reader's state of mind" (Bender et al., 2021, p. 616). For these authors, not having a world seems to suggest that these language models "have never observed" or otherwise interacted in any way with "the real world" (Bender and Koller, 2020, p. 5188). Having never interacted with the world, LMs cannot engage in any kind of meaningful linguistic behaviour with others at all because they do not have beliefs about the world nor can they attribute beliefs to speakers much less 'model' another's 'state of mind.' The authors conclude, "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" (Bender et al., 2021, p. 616-617). There is no reference to meaning because, as said, the strings of words they produce are "not grounded in communicative intent, any model of the world, of any model of the reader's state of mind" (p. 616).

This conclusion might appear obvious, and it does seem so to the authors of "On the Dangers." Yet, when we attend to how individuals interact with LMs, the authors find that it is not all that obvious. Looking at the side of human beings in their exchanges with LMs, the authors identify human beings' "tendency" to regard LMs' linguistic production as meaningful, as carrying communicative intents, which further contributes to the misrepresentation of LMs as natural language understanding models[3]: "the tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking synthetic text as meaningful" (Bender et al., 2021, p. 611). It is the "tendency" to attribute communicative intentions to others, the first of which is the intention of being understood which LMs lack but human beings misattribute to LMs. But this tendency goes much further, for by attributing communicative intent, human beings are attributing mental states to LMs along with a model of the world, and, what is more, a similarity to themselves, which is something human beings do in their communicative exchanges generally and insofar as they engage in communicative exchanges. The authors regard these attributions as "an illusion arising from our singular human understanding of language" (p. 616). Despite this blatant illusion, human beings, the authors continue, are "very willing… to attribute communicative intent even if the originator of the signal is not an entity that could have communicative intent" (Bender and Koller, 2020, p. 5187). Their 'willingness', however, is unwarranted.

RECONSIDERING COMMUNICATIVE INTENT AND REFERENCE AS CONDITIONS OF COMMUNICATION AND UNDERSTANDING

After this brief clarification of the concepts and arguments presented in "On the Dangers," it should be clear that this framework is evidently applicable only to human-to-human linguistic exchanges, specifically involving adult humans. It fails to account for entities capable of producing strings of words but lacking beliefs, intentions and interactions with the human world such as certain animals and machines. This concept of communication is overly restrictive not only because it excludes entities other than humans, but also because it only captures certain instances of human-to-human communication. In the following discussion, I problematise the premises and concepts of the authors' arguments from the perspective of Jacques Derrida's philosophy, particularly focusing on the two main conditions of communication and understanding in communication: communicative intent and model of the world. Then, I articulate alternative concepts of communication and understanding in order to revisit the question of whether LMs and chatGPT communicate.

a) Communicative intent

The conceptual and argumentative framework presented in the previous section is explicitly based on the Gricean model of communication (Bender and Koller, 2020, p. 5187), where intentions play a decisive role in the entire communicative process, from the selection of strings of words or noises to how the communicative process itself unfolds

[3] LMs' developers are also responsible for this confusion, for they tend to describe LMs as 'understanding'. Cf. Bender and Koller, 5185-5186.

and whether the communication can be deemed successful. A version of this view is largely accepted by philosophers of language in the analytic tradition from Searle to Davidson, albeit with some important differences (Searle, 1977; Davidson, 1992). For Bender et al., communicative intent or communicative intention presupposes a person deliberately initiating a communicative exchange with the purpose of doing something or affecting the interpreter in a particular way. This purpose is their communicative intent, and it makes the strings of words or noises in a linguistic exchange meaningful insofar as these strings were selected as a means to achieve the desired effect on the interpreter.

Derrida notes, however, that no such person imbuing a string of words or noises with communicative intent is necessary for signs to be meaningful and have the expected effects on interpreters. More generally, no communicative intent is necessary for signs to be meaningful. Signs should be able to function, that is, to be interpretable in meaningful ways, even if the producer of the signs is not present or has never existed. "A mark," he says, is "a sort of machine which is productive in turn" and "must continue to 'act' and to be readable even when what is called the author […] no longer answers for what he has written […] be it because of a temporary absence, because he is dead, or because he has not employed […] the plenitude of his desire to say what he means" (Derrida, 1988, p. 8). This is possible, Derrida continues, because, in order to be such, signs have to be repeatable, that is, any sign can be weaned from its putative or otherwise context of production and placed in an altogether different context without hindering its possibility of being meaningful.

In his response to Derrida's "Signature, Event, Context," Searle takes up and discusses this argument. Although he agrees with the repeatability of signs, he disagrees with the conclusion Derrida draws from it (Searle, 1998, p. 201-202). He concedes that even if it were the case that no producer is necessary for a meaningful communication, understanding and interpreting signs would still require assuming a communicative intent "because a *meaningful sentence is just a standing possibility of the corresponding (intentional) speech act*" (Searle, 1998, p. 202). Understanding, he says, is just "knowing what linguistic act its utterance would be a performance of,", even if there was no actual utterance. In these cases, strings of words or noises are meaningful and can, thus, be presumed to have communicative intent if they follow the rules of language. "To understand it [a speech act], it is necessary to know that anyone who said it and meant it would be performing that speech act *determined* by the rules of language that give the sentence its meaning in the first place" (Searly, 1998, p. 202, my emphasis). In other words, Searle maintains that it is through the rules of language that a hypothetical communicative intent can be articulated. For example, when a chat bot produces the lines 'Provide your name, email address and order details' in an automated way, there is no actual individual imbuing these phrases with communicative intent. However, the use of the imperative directs the interpreter to the rules governing imperatives, from which an intention can be inferred. Specifically, the intention in this example is to be authoritative and to prompt the interpreter to comply by providing the requested information. It is rules of language – or conventions, as Searle sometimes calls them – that confer intent or purpose to spoken or written marks, thereby rendering them meaningful. It is through these rules that a communicative intent can be discerned. In contrast to the view of

communicative intent advanced by the authors of "On the Dangers," Searle's concept of communicative intent does not necessarily involve an actual person intending to convey a meaning by selecting the strings of words or sounds that are most likely to articulate that intent. Instead, it is regarded as a "strategy of understanding" – a helpful presupposition that facilitates the interpretation and comprehension of written and spoken marks (Searle, 1998, p. 202).

Searle's way of sidestepping Derrida's argument dispenses with the necessity of a producer of communication without sacrificing meaningfulness and communicative intent. However, inadvertently, his position seems to align more closely with Derrida's and further from Bender et al.'s, as both Searle and Derrida agree that actual individuals or, more broadly, human beings linguistically articulating communicative intents are not essential for communication to occur. This agreement challenges Bender et al.'s first condition of communication, which states that communication occurs between individuals possessing beliefs and other mental states. Derrida would also agree with Searle's assertion that intention is a presupposition orienting interpretation and understanding rather than the necessary key to interpretation and understanding when he says, "the category of intention will not disappear, it will have its place, but from that place it will no longer be able to govern the entire scene and system of utterance" (Derrida, 1988, p. 18) Bender et al. would actually perceive Searle's way of salvaging communicative intent when applied to LMs as part of the problem, for the mischaracterisation of LMs as language understanding systems is fostered partly by presupposing and attributing communicative intent where there is none.

Searle's position shifts our focus towards the interpreter rather than the speaker or initiator of communication. From the interpreter's perspective, whose first task is deciphering the communication, communicative intent, as mentioned earlier, is possible but not necessary, and functions as a presupposition assisting in interpreting linguistic productions. The crux of Searle's viewpoint lies in the rules of language, particularly grammar as logical syntax, whose correct usage aids in articulating an intention which may or may not have been actual. This view is, however, fairly easy to question, for the idea that language use requires knowing or even applying grammar correctly is unwarranted. Language acquisition and use are primarily practical and occur without the explicit need to learn grammar rules. Furthermore, using language 'incorrectly' from a grammatical standpoint does not necessarily hinder communication as it is common to successfully interpret speech containing grammatical errors.

Derrida delves into this topic, discussing not just isolated grammatical errors, but agrammaticality – instances where there is no longer "'logical' language" (Derrida, 1988, p. 11) Even these cases, Derrida affirms, need not compromise communication. The reason is that since cases of agrammaticality "[f]or instance, 'the green is either' or 'abracadabra'", he says, "do not constitute their context by themselves, nothing prevents them from functioning in another context as signifying marks" (p. 12). This underscores Derrida's initial argument that in order to function and continue to be interpretable and meaningful, signs have to be capable of separation from their context of inscription. Thus, 'the green is either' could be inscribed in semantic or real contexts where it could be meaningful. Derrida's argument here challenges the necessity of rules of language for

communication and understanding in communication. This does not imply that language rules are never useful in interpretation; rather, they are not a sufficient and necessary condition of communication and understanding.

Bender et al. could strengthen their position by drawing on Davidson's reflections on language use and communicative exchanges. In Davidson's work, they could find an alternative defence of intentions in communication based on another, arguably more important, function of intentions. He affirms that "[t]he necessary presence of intentions would be significant, since it would give content to an attribution of error by allowing for the possibility of discrepancy between intention and accomplishment" (Davidson, 1992, p. 259). Speakers initiate communicative exchanges not necessarily with a single intention as they likely aim to convey something to provoke a response or behaviour, which in turn may lead to further outcomes. However, we can narrow down this array of intentions to the primary and fundamental intention of being understood. It is possible though that this intention is not fulfilled. For various reasons, such as the speaker incorrectly assuming that the listener would grasp crucial cues, misinterpreting the context, or misjudging the listener's knowledge or willingness to interpret in certain ways, the chosen words or sounds may fail to achieve the intended communicative effect. In such situations, intentions serve to highlight the discrepancy between intention and actual outcome of the communication, ranging from explicit acknowledgement of misunderstanding to the disparity between the behaviour the speaker anticipated and the actual response from the listener. Davidson argues that without communicative intent, there would be no means of spotting errors, i.e., unsuccessful communications. Or, more consequentially, there would be no mistaken interpretations as any interpretation would probably be good enough.

We could ask, however, is it not rather common that we cannot test our interpretations against the speaker's/writer's intentions? In Davidson's picture, the speaker and interpreter are facing one another, which vastly facilitates testing the interpretation. However, it does not guarantee that an error, if there is one, will be spotted because it might be the case that the speaker, as Derrida affirms, "has not employed his absolutely actual and present intention or attention, the plenitude of his desire to say what he means" (Derrida, 1988, p. 8), or has conflicting intentions or is not fully conscious of his intentions and cannot respond for what he has said or written. Testing interpretations becomes even more challenging in various cases of communication, such as reading the newspaper, listening to someone's voice message, interpreting the work of a long deceased author, perusing personal journal entries of years past, and the list could go on. In these cases, interpreters cannot double check their interpretation against the intentions of the speaker/writer. If intention were the norm of interpretation, we would have to acknowledge that misinterpretation is highly possible, and probably factually common, for it is simply not an element that can be relied upon in all cases of communication, either because it is factually absent or because the communication was not intended with full attention, among many other possible reasons. This does not mean that intention, if there is one, can never serve as a guiding post orienting interpretation and aiding in identifying errors of interpretation. At times, it may serve this purpose. As suggested

before, however, it cannot be regarded as a necessary condition for communication and understanding in communication.

In this section, I have discussed intentions as a condition for communication on various grounds: because communication is to have a purpose, because it serves as a 'strategy of understanding' or presupposition orienting understanding, and because it functions as a norm sanctioning interpretation, which closely aligns with Bender et al.'s definition of understanding as 'retrieving communicative intent'. If we consider, in relation to the first of these grounds, that the primary purpose or intention of communication is to be understood, then all of these grounds emphasise the role of intentions for the possibility of understanding, that is, for the possibility of a successful interpretation of what is communicated. Here, understanding means successful interpretation where intended meaning matches interpreted meaning. Drawing from Derrida's views, however, I argued that this matching cannot be deemed necessary because the process of matching presupposes an intention which may or may not be present and reliable. Thus, it would be useful to rethink the concept of understanding in communication in such a way that it is not defined necessarily by this matching. More concretely, it should be a concept of understanding that does not necessarily consist in the fulfilment of intention insofar as communicative intent may or may not be present and reliable, and that may rely on semantic and real contexts without regarding them as fixing interpretation. Such a concept of understanding could probably be characterised as our *best guess* about what the communication is about and aims to cause. It is a *guess* because there is no single factor that interpretation can rely upon in all cases and with full certainty. But it is the *best* guess because factors such as possible intentions, possible contrasting the interpretation with the speaker's intention, rules of language, semantic and real context, previous experiences of communication, etc., can sometimes assist to a greater or lesser degree in orienting interpretation, and which can be appealed to to justify one interpretation over another.

This concept of understanding is less restrictive than Bender et al.'s and, I want to say, more immediately applicable to beings other than human beings such as animals and LMs. Focusing on LMs, Bender et al. state that LMs can neither mean something nor retrieve meaning, that is, they can neither communicate nor understand because their linguistic outputs are the result of probabilistic operations indicating the likelihood –not the meaningfulness- that certain strings of words follow the input. Yet, when understanding is defined not as meaning retrieval, matching of intended and interpreted meaning, or strict grammar rule following, but rather as the best guess, namely, as the most probable interpretation, could not LMs be said to understand in this sense? Before discussing this question, I would like to move on to the second condition of communication and understanding at the basis of Bender et al.'s emphatic claim that LMs can neither communicate nor understand in communication.

b) The real world or a model of the world

The second condition necessary for communication and understanding in communication, as discussed by Bender et al. in "On the Dangers" is what they term 'a model of the world.' Bender and Koller do not talk about the model of the real world but rather about the real world, so I will treat them interchangeably for the time being. The

'real world' or 'the model of the real world' has two interrelated functions in Bender and Koller. Firstly, it is that which communications are about, and, secondly, they are that against which the truth of communications can be tested, but they focus exclusively on the former. The bulk of my criticisms rest on the fact that Bender et. al. and Bender and Koller provide insufficient reasons to restrict the scope of communicational topics to objects in the real world. In addition, I argue that LMs can be said to have a model of the world.

In "On the Dangers", the model of the world is vaguely referred to as the "common ground" which speakers/writers and interpreters share (Bender et al., 2021, p. 616). In "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data" Bender and Koller clarify the notion of 'real world' and spell out in more detail how it features in communicative exchanges. It first appears in relation to communicative intents: the authors maintain that "communicative intents are about something that is *outside language*" (Bender and Koller, 2020, p. 5187, italics in the original). The examples given include "*Open the window!* [and] *When was Malala Yousafzai born?*" (p. 5187, italics in the original). Communicative intents "can also be about abstract worlds, e.g. bank accounts, computer file systems" (p. 5187). These examples are quite evidently about something non-linguistic, in particular windows, Malala Yousafzai, bank accounts and computer file systems, and in this sense they could be said to be 'outside' language. Bender and Koller do not mean it in this sense, though. Rather, they claim that in these examples, "the communicative intent is grounded in the real world the speaker and listener inhabit together" (p. 5187). They do not explain what 'grounding' here means, but they do state that this grounding is at the basis of the role 'the real world' or 'the model of the real world' performs in communicative intents. It is strange, however, that immediately following this assertion, the authors state that communicative intents can also be about "a purely hypothetical world in the speaker's mind" (p. 5187), because a purely hypothetical world in the speaker's mind does not seem to be grounded in the real world if 'grounding' is meant to signify a relation of reference whereby strings of marks are about objects in the real world. Setting this aside for the moment, the 'real world' or the 'model of the real world' features in yet another instance, namely, linguistic systems.

Linguistic systems "provid[e] a relation […] which contains pairs (*e*, *s*) of expressions *e* and their conventional meanings *s*" (Bender and Koller, 2020, p. 5187). Conventional meanings seem to be standardised meanings, and the authors defined them as "what is constant across all of its possible contexts of use" (p. 5187) and "an abstract object that represents the communicative potential of a form, given the linguistic system it is drawn from" (p. 5187). Linguistic systems also relate to the real world, which is what seems to be described as 'outside language.' They say, linguistic systems "connec[t] language to objects outside of language" (p. 5187). So, in communication, the speaker is 'grounding' her/his speech on the 'real world' or the 'model of the real world' from two angles: the speaker has a communicative intent grounded in the real world, which is conveyed through the use of expressions of the linguistic system that itself is also grounded in the real world. Now, the listener shares the real world or the model of the real world with the speaker, and also largely (p. 5187 n 6) shares the linguistic system

with the speaker. This twofold sharing helps the listener to retrieve the communicative intent of the communication.

In order to illustrate the roles the real world or the model of the real world plays in communicative exchanges, Bender and Koller describe a particular scenario. In this scenario, two English speakers are stranded on two separate isolated islands, but luckily, previous inhabitants of these islands left behind telegraphs that they can use to communicate, and they start to use it routinely to have all kinds of conversations. Unbeknownst to them, an exceptionally smart octopus living underwater finds "a way to tap into the underwater cable and listen in on A and B's conversations" (Bender and Koller, 2020, p. 5188). The octopus "is very good at detecting statistical patterns" (p. 5188), so it manages to identify numerous patterns in how A and B speak. One day, the octopus poses as B and responds to A's messages. The question the authors ask is whether the octopus can "successfully pose as B without making A suspicious?" (p. 5188). The answer, they continue, depends on what the conversation is about: if A's "utterances […] have a primarily social function, and do not need to be grounded in the particulars of the interlocutors' actual physical situation nor anything else specific about the real world" (p. 5188), then the octopus might actually manage to pull it off because in this case "it is sufficient to produce text that is internally coherent" (p. 5188). If, on the other hand, A's conversation refers to something in the world, then the octopus is likely to struggle to produce speech that is meaningful. For example, "A […] is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself" (p. 5189). This task, the authors continue, "requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that [the octopus] would fail" (p. 5189). The reason for the failure is that the octopus "has never observed these objects, and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives" (p. 5188). The octopus has no model of the world or experience of the real world A experiences, so its responses will be meaningless.

A couple of questions could be raised here. Firstly, the reason the octopus statistician can only produce meaningless responses to A is that it has not experienced or observed the world. From this, we can conclude that observation and experience are crucial for the possibility of constructing a model of the world or grounding communications in the real world. This is, however, evidently not always the case. A blind person can know a lot about blue skies and be able to talk at length about them without ever having seen blue skies. Certainly, this person could be said to have experienced blue skies vicariously through others, but this person has certainly not *observed* blue skies as such. So, observation cannot be regarded as necessary for the possibility of having a model of the world or talking about something. Derrida discusses this point in relation to Husserl's first consideration of the absence of the referent. He says, "An utterance whose object is not impossible but only possible can very well be made and understood without its real object (its referent) being present, either to the person who produced the statement or to the one who receives it" (Derrida, 1988, p. 10). For example, a person could say "'The sky is blue'" and this utterance would be intelligible and interpretable even if neither the speaker nor the interpreter see the sky, if

the speaker is mistaken or is lying. This is clearly not always the case, "but the structure of possibility of this utterance includes the capability to be formed and to function as a reference that is empty or cut off from its referent" (Derrida, 1988, p. 10-11). Without this possibility, Derrida contends, signs would not function and be readable and interpretable.

Secondly, Bender and Koller maintain that communicative intents, as well as linguistic systems, are grounded in the real world, in the sense that words articulating intents and forming expressions refer to individualisable objects in the world. Bender and Koller are quick to make a disclaimer to the effect that this relation of reference is not a relation of grounding truth. They state, "we should be careful not to confuse communicative intent with ground truth about the world, as speakers can of course be mistaken, be intentionally dissembling, etc." (Bender and Koller, 2020, p. 5187). If the relation of reference is not a relation intended to ground the truth of statements, then it could be thought of as a relation of 'aboutness' in the sense that the objects usually defined as constituting the world can be objects of descriptions, topics of conversation, say. But if that is the case, then it is unclear why Bender et al. limit topics of conversation to objects in the real world, especially because their octopus story clearly oversteps that limit insofar as there are no octopuses versed in statistics that we know of and they have no referent in the real world. Derrida's discussion of Husserl's second consideration of the absence of the referent clarifies why limiting meaningfulness to what can be referred to in the real world is not justified (Derrida, 1988, p. 11): Husserl analyses the "absence of the signified" (in three instances: a) signs can be manipulated and intelligible without them referring to anything, for example, "mathematical symbolism" (p. 11); b) "[c]ertain utterances can have meaning although they are deprived of *objective* signification" (p. 11). The example Husserl gives is 'the circle is square'. This phrase, Derrida continues, "has sufficient meaning at least for me to judge it false or contradictory" (p. 11); c) "what Husserl calls *Sinnlosigkeit* or agrammaticality. For instance, 'the green is either' or 'abracadabra'" (p. 11). In these examples "there is no more 'logical' language" (Derrida, 1988, p. 11). However, these phrases can very well be placed in contexts where they will be meaningful. These considerations show that communications do not have to be about the real world.

Now, as noted earlier, Bender et al. talk about 'the model of the real world' whereas Bender and Koller about 'the real world,' and there are reasons to think that these expressions are not interchangeable. LMs and chatGPT can be said to lack a world and a relation with the real world in the sense that they cannot sit down on a chair or buy a train ticket, and so are unable to form beliefs, memories, desires, etc., about the world. Yet, LMs and chatGPT have huge amounts of data about the world and, in that sense, they could be said to have a model of the world. Insofar as the data they are trained with is not tested for coherence or truth, it is probable that their models of the world are not particularly coherent. However, neither are those of human beings, which can at best be described as largely coherent rather than as fully coherent.

RECONCEPTUALISING COMMUNICATION, UNDERSTANDING AND THE MODEL OF THE REAL WORLD

A point continuously made by the authors of "On the Dangers" is that human beings are mistaken in treating their interactions with LMs as communicative exchanges. They regard the "tendency" to treat LMs' linguistic productions as meaningful as an "illusion" (Bender et al., 2021, p. 616), and a "deception" (Bender and Koller, 2020, p. 5189). The norm against which they make these claims is the view of communication we have been discussing, which requires utterances to be grounded in communicative intents and the real world, and involves numerous cognitive processes like attributing beliefs, retrieving intentions, assigning meanings to words, testing their correctness, etc., which LMs clearly cannot do. It is worth asking, however, whether individuals' description of their exchanges with LMs can provide additional reason to justify a transformation of the concepts of communication and understanding to make room for the novel forms of interactions individuals are having with these new technologies.

In an opinion piece published in *Globe and Mail*, Derek Ruths states, "The way we interact with ChatGPT is virtually identical to the way we communicate with people every day" (Ruths, 2023). His examples are significant: "on a screen, with a small text box, viewing a scrolling window of dialogue[, t]he standard ChatGPT interface looks like WhatsApp, SMS, Apple Messages, and every other messaging app" (Ruths, 2023). Indeed, the interfaces of messaging apps are basically indistinguishable from ChatGPT's interface which explains why exchanges with ChatGPT feel 'virtually identical' to many of the communications people have with other human beings through these apps. For younger generations, this kind of virtual exchanges, rather than face-to-face interactions, is actually the norm not only because they are digital natives, but also because they grew up during the COVID-19 pandemic where online communication was the predominant mode of communication. For these younger generations, interacting with ChatGPT does not 'feel' significantly different to the online interactions they have with their friends and relatives. And it is certainly not only younger generations that 'feel' this way. It is not only the similarity of ChatGPT's interface to those of other apps that contribute to these feelings of similarity. There are also signs which individuals would normally interpret as indicating that ChatGPT is performing some cognitive activity: "ChatGPT even generates little thought typing bubbles while it's working up its response" (Ruths, 2023). Those bubbles also appear when a real person is typing a message, regardless of whether the person is actually doing any thinking.

In this article, I have offered reasons to support changing the concept of communication. In particular, I argued that communicative intent should not be taken as a condition for the possibility of communication and understanding in communication because it is not always present or reliable in vast cases of communication, even in face-to-face communication. This does not exclude intent *tout court* as it will be useful to articulate communications and orient interpretation in some cases. A similar argument was made regarding the real world and its grounding function. It was argued that communication can very well be about hypothetical and illusory objects, without

hampering meaningful interpretations. The article has also hinted at some possible ways forward, which I briefly discuss in this last section.

We are seeking a concept of communication that does not necessarily require communicative intent and is not necessarily grounded in the real world. Derrida finds in Austin's theory of speech acts a concept of communication that can meet these conditions with some modifications. He says that in Austin's analysis, communication acts "do not designate the transference or passage of a thought-content, but, in some way, the communication of an original movement (to be defined within a *general theory of action*), an operation and the production of an effect" (Derrida, 1988, p. 13). The performative utterance, Derrida continues, "would be tantamount to communicating a force through the impetus (*impulsion*) of a mark.". Moreover, "the performative does not have its referent … outside of itself or, in any event, before and in front of itself. It does not describe something that exists outside of language and prior to it. It produces or transforms a situation, it effects" (Derrida, 1988, p. 13). The performative utterance is exemplary because, in contrast to constative utterances, it does not have to transmit a meaning, refer to something in the world or assert some truth about objects in the world or state of affairs in the world.

The original movement or operation Derrida talks about in the quotation is elaborated on in the first chapter of *Of Grammatology*, where Derrida recounts the transformation of the philosophical concept of language. Initially defined as a mediating tool between subject and world, alternatively expressing the meaning intended by the former or representing objects in the latter, language is reconceptualised as an operation of "demotivation" (Derrida, 1976, p. 51) and "overwhelming" (p. 7) production of signs with various degrees of repeatability (differential force). As an operation of demotivation, the repetition of signs erodes any intention or reference (if there was one) constraining interpretation. Demotivation and overwhelming do not render signs unintelligible or inscrutable. Instead, by eroding the limit imposed by an intention or an object in the world, the operation of demotivation makes signs 'overwhelming' in that possibilities of interpretation (or possibilities of understanding in communication) increase potentially to infinity. This proliferation of possibilities alters the world by expanding the realm of possible interpretations. While some interpretations may possess more 'force' than others at times, meaning they are more likely to become binding or authoritative, the overwhelming nature of signs persists as other interpretations remain possible even if they have less force, less chance of becoming binding or authoritative at certain times.

The differential force of interpretations derives from such factors as the possible intention, the semantic context, the real context, the rules of language, past communicative experiences, among others. It was said that none of these elements is necessary for the possibility of communication and understanding, but that none of them is excluded either, for they are all possible factors that might contribute to tipping the interpretative balance in one direction or another. Earlier, I proposed that understanding could be defined as 'the best guess.' The notion of 'guess' is appropriate here because the demotivated and overwhelming character of signs challenges the idea that there is only one interpretation, *the* correct interpretation that can be easily retrieved by following a sure method. Without one single interpretation, the aforementioned factors can incline

interpreters toward one interpretation, but they don't guarantee it. Thus, the most interpreters can offer is a guess, their best guess, given the factors contributing to the force of their interpretation. Defining understanding as the best guess would allow us to regard LMs' outputs as understanding, for their outputs are also a guess based on likelihood rather than meaning. Finally, it was suggested that although LMs lack experience of the real world, they could be said to have a model of the world. In their case and in contrast to human beings, their model of the world is not a system of beliefs related inferentially. Instead, it comprises the information, the data about the world with which they have been trained. LMs' outputs are based on this information, which, just like human beings' models of the world, is more or less coherent.

If these concepts of communication, understanding in communication, and model of the world are accepted, then it would be possible to reconsider the question of whether LMs communicate and understand in communication, and provide an affirmative answer.

CONCLUSION

This paper discusses Bender et al.'s and Bender and Koller's concepts and arguments underlying their negative answer to the question of whether LMs communicate and understand, in particular the concept of communicative intent and model of the world or real world. I argued that communicative intent cannot be regarded as a condition for the possibility of communication because it is absent or unreliable in vast cases of communication. I also argued that communication and understanding need not be grounded in the real world if that implies limiting communication to what can be referred to in the real world. Having discarded the necessity, although not the possibility, of these two factors, alternative concepts of communication and understanding were elaborated. Drawing from Derrida's discussion of Austin's speech act theory and his own philosophy, the alternative concepts of communication and understanding retain the idea of communication as transmitting and affecting. Yet, what is transmitted is an operation of demotivation and overwhelming of possible interpretations with differential forces. Understanding was defined as the best guess rather than as retrieving *the* correct interpretation, which is aided by a model of the world defined as more or less coherent information. Armed with these concepts, the paper advances an affirmative answer to the question of whether LMs communicate and understand.

REFERENCES

Bender, E., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). Association of Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). *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* (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Davidson, D. (1992). The Second Person. *Midwest Studies in Philosophy*, *17*(1), 255-267. https://doi.org/10.1111/j.1475-4975.1992.tb00154.x

Derrida, J. (1988). Signature, Event, Context. In *Limited Inc* (pp. 1-24). Northwestern University Press.

Derrida, J. (1976). *Of Grammatology*. The John Hopkins University Press.

Ruths, D. (2023, May 19). ChatGPT is Blurring the Lines between what it Means to Communicate with a Machine and a Human. *The Globe and Mail*. https://www.theglobeandmail.com/opinion/article-chatgpt-is-blurring-the-lines-between-what-it-means-to-communicate/

Searle, J. (1997). Reiterating the Differences: A Reply to Derrida. *Glyph Review*, *2*, 198-208

СВЕДЕНИЯ ОБ АВТОРЕ / THE AUTHOR

Ребека Перес Леон, drrpleon@gmail.com
ORCID 0000-0003-4912-4912

Rebeca Perez Leon, drrpleon@gmail.com,
ORCID 0000-0003-4912-4912

Статья поступила 3 марта 2024
одобрена после рецензирования 1 июня 2024
принята к публикации 10 июня 2024

Received: 3 March 2024
Revised: 1 June 2024
Accepted: 10 June 2024