filosofia unisinos | Unisinos Journal of Philosophy

ISSN 1984-8234

Filosofia Unisinos
Unisinos Journal of Philosophy
25(1): 1-14, 2024 | e25112

Unisinos – doi: 10.4013/fsu.2024.251.12

Article

AI beyond a new academic hype: an interdisciplinary theoretical analytical experiment (computational, linguistic and ethical) of an AI tool

IA para além de uma nova moda acadêmica: um experimento teórico-analítico interdisciplinar (computacional, linguístico e ética) de uma ferramenta de IA

Murilo Mariano Vilaça[1]
https://orcid.org/0000-0001-9720-5552
Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública, Programa de Pós-Graduação em Bioética, Ética Aplicada e Saúde Coletiva (PPGBios), Programa de Pós-Graduação em Saúde Pública (PPG-SP), Rio de Janeiro, RJ. Brasil. Email: murilo.vilaca@fiocruz.br

Isabella Lopes Pederneira[2]
https://orcid.org/0000-0002-5884-8472
Universidade Federal do Rio de Janeiro - UFRJ, Programa de Pós-Graduação em Linguística, Rio de Janeiro, RJ, Brasil. Email: isabellapederneira@letras.ufrj.br

Mariza Ferro
https://orcid.org/0000-0003-0191-582X
Universidade Federal Fluminense - UFF, Programa de Pós-Graduação em Computação, Niterói, RJ, Brasil. Email: mariza@ic.uff.br

[1] Bolsista da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ - JCNE). GN: E-26/201.377/2021. Bolsista APQ do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq): PROGRAMA PRÓ-HUMANIDADES, GN: 421523/2022-0. Bolsista de Produtividade em Pesquisa (PQ) do CNPq, GN: 315804/2023-8. Agradeço à FAPERJ e ao CNPq pelo apoio.

[2] Bolsista da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ - JCNE), GN: E-26/200.118/2023.

ABSTRACT

Artificial Intelligence (AI) is the newest technological hype. Although it involves a great diversity of technologies, the disseminated image of AI has been a generic, super-powerful, extremely negative one, at the edge of dystopian, especially in works of science fiction. Recently, following the launch of ChatGPT, one finds an explosion of journalistic and academic publications, some of which reinforce this public imagination. In academia, AI's risks remain highlighted and even apocalyptic scenarios related to it are taken into consideration. Ethical concerns, then, take on eminently negative connotations, and the much more realistic possible benefits of AIs are under-focused. We highlight in this paper the importance of going beyond the hype and developing multi/interdisciplinary, delimited, realistic, and thoughtful approaches. Employing the collective intelligence embodied in the interaction between researchers from three fields —Computer Science, Linguistics, and Philosophy— we approach a free machine translation tool: Google Translate. Our goal is to show that well-defined, multidisciplinary, technically supported approaches that do not adhere to sensationalist discourse lead us to more epistemically consistent and more thoughtful critical-normative reflections, which would be crucial for advancing the debate on AI.

Keywords: artificial intelligence, machine translation, thoughtful approaches, ethics.

RESUMO

A Inteligência Artificial (AI) é a mais nova hype tecnológica. Embora envolva grande diversidade de tecnologias, uma imagem genérica, superpoderosa, fortemente negativa e, no limite, distópica da IA tem sido disseminada, especialmente por obras de ficção científica. Recentemente, sobretudo após o lançamento do ChatGPT, houve uma explosão de publicações jornalísticas e acadêmicas, que, em parte, reforçam esse imaginário. Nas abordagens acadêmicas, os riscos seguem em destaque e, inclusive, cenários apocalípticos relacionados à AI são levados em consideração. A preocupação ética, então, assume um sentido eminentemente negativo e os possíveis benefícios das AIs, muito mais realistas, são subfocalizados. Neste artigo, destacamos a importância de ir além da hype, desenvolvendo abordagens multi/interdisciplinares, delimitadas, realistas e ponderadas. Recorrendo à inteligência coletiva expressa na interação entre pesquisadores de três áreas (ciência da computação, linguística e filosofia), abordamos uma ferramenta de tradução automática (Machine Translation) gratuita: Google Translate. Nosso objetivo é mostrar que abordagens bem delimitadas, multidisciplinares, tecnicamente respaldadas e que não adotem o discurso sensacionalista nos conduzem a reflexões epistemicamente mais consistentes e crítico-normativas mais ponderadas, o que seria fundamental para o avanço do debate sobre AI.

Palavras-chave: inteligência artificial, tradução automática, abordagens ponderadas, ética.

1 Introduction: identifying some issues

One of the central features of our time, the deliberate process of accelerating technological progress, brings some challenges. One of them is directed at our reflexive-critical exercise, it refers to how we, academics, can bridge the gap between the speed of technological advancement and the time demanded to produce consistent reflection upon it. The gap has always been a challenge for academics, but it has gained a new layer of complexity in the face of the phenomenon of technological hype. In a scholarly environment, this phenomenon leads to the challenge of keeping up with the frequent shift of attention produced by *the constant transition from one 'academic technological fever' to the next*. To put it simply, after a new technological subject arrives and generates great mobilization and a publication frenzy, the arrival of the new topic seems to cover up previous ones, consequently shifting academics' attention. Sometimes, this occurs before the debate has properly matured. (Seifert, Fautz, 2021).

The sudden appeal generated by an innovation in a determined techno-scientific field; the rapid dissemination of discourse (often inflated and tabloid-like) via mass media; the way a technology is predominantly addressed in fictional works influencing public debate; and a sudden journalistic and academic frenzy of publications, are some elements of a *hyped scenario*, where one finds some pitfalls.

Artificial Intelligence is the latest technological hype, academia included. Under the influence of catastrophic/apocalyptic fictional narratives, on the one hand, and sensationalist/inflated discourses (both journalistic and academic) —whose entertaining and mobilizing-engaging effects put under focus the *psychological accessibility of the public as a mere consumer of a product with an immediate return* (Habermas, 2003)—, on the other, the rigid distinction between fact and fiction becomes blurred, influencing —worrisomely, we would say— the public debate and the academic approaches on the topic (Geraci, 2010a). Such a problem has been identified in several academic-technological debates. In them, works of literary or cinematic fiction (some of a religious nature) and hyperbolic speech set a tone where confusing rhetoric masks itself as an argument (Buchanan, 2011; Vilaça, 2021; Rueda, 2023).

In AI's specific case, academics are among those that share the apocalyptic imagination with some wide-ranging intensity and persistence, revealing our permeability to the influences of a questionable mix of fictional narrative and religion, which includes prophecies about AI's future and its impact on humanity (Geraci, 2006; 2007; 2008; 2010a; 2010b; 2022; DiCarlo, 2016; Shermer, 2017; Kearney, Wojcik, Babu, 2019; Lemay, Basnet, Doleck, 2020; Noy, Uher, 2022; Paulus Jr., 2023; Robertson, Maccarone, 2023). AI as the bringer of the 'end of days' is, so to speak, the apex of the tone dominating a part of this new hype, which is deeply —and questionably— marked by binary opposition, the presence of polarized positions (Buchanan, 2011), and by the so-called dilemmatic fallacy, according to which a complex problem is reduced to two diametrically opposed options (solutions) (Junges, 2019). The result is a hope or threat, promise or peril, hype or reality, good or evil, utopia or apocalypse, sensationalistic academic literature. The proliferation of such literature sometimes dominates the debate. However, it does so without producing worthy epistemic contributions to its advancement, since restricts itself to a *pro versus con* rhetoric for unnecessarily long periods, leaving little space for level-headed and reasonable stances (Buchanan, 2011; Seifert, Fautz, 2021; Vilaça, Lavazza, 2022).

Ethics ends up colonized by the polarization discourse in the scenario as delineated (a sketch made in general terms and unrepresentative of the totality of the debate). Ethical people would be the ones assuming a stance committed to what Hofmann (2019) calls *status quo bias*, as they believe themselves to be ethical when they focus almost exclusively on the associated risks linked to changes resulting from the technology. They take upon themselves the role of protectors of certain values, the ones responsible for the maintenance of the status quo, acting in a precautionary manner. In the extreme opposite, one finds the group influenced by what Hofmann (2019) labels *progress bias*, the ones adhering to an exaggerated and unjustified scientific and technological optimism.

One of the possible harmful effects of this would be a questionable bias in AI's ethical debate. The debate would be mainly focused on its risks, giving rise to what we call *risk ethics*[3]. The reason may have

[3] Although benefits are sometimes mentioned, the emphasis on risks is also evidenced by the significantly different number of occurrences of the terms 'risk' and 'benefit'. In searches carried out in three databases (VHL Portal, Web of Science, and PubMed) – using, respectively, the strategies (artificial intelligence) AND (benefit) and (artificial intelligence) AND (risk) (applied to the summary item); (artificial intelligence[Title/Abstract]) AND (benefit[Title/Abstract]) and (artificial intelligence[Title/Abstract]) AND (risk[Title/Abstract]); (AB=(artificial intelligence)) AND AB=(risk) and (AB=(artificial intelligence)) AND AB=(benefit) – significant discrepancies were found in the number of results. By way of illustration, in the PubMed database, the results for the term "risk" was 4954, while for "benefit" it was 1381. Using the same database, when applying the search strategies ((artificial intelligence[Title/Abstract]) AND (risk[Title/Abstract])) NOT (benefit[Title/Abstract]) and ((artificial intelligence[Title/Abstract]) AND (benefit[Title/Abstract])) NOT (risk[Title/Abstract]), the results were, respectively, 4720 and 1081. Although slightly less discrepant, there was a relevant difference in results in the other databases as well.

something to do with theoretical-conceptual-methodological gaps in theoretical and applied ethics. But, we believe this is only half of the story. The focus on threats (ultimate threats and all), on risks (of existential or catastrophic natures), and on our duty to avoid them (giving only a 'negative' delineation to ethical approaches) makes explicit not only a partial understanding of the ethics. Adding to that, it seems to be a consequence of the lack of understanding of technological dynamics, it is a misunderstanding of epistemic and technical aspects.

Having that in mind, our goal is to show the need for approaches guiding us beyond AI's hype. Such approaches are not polarized, they are well-delimited, with consistent technical support, and they are accompanied by level-headed normative critical analyses. In this direction, in this paper we reject polarization. Focusing on a specific AI, we develop a multidisciplinary approach and we indicate the relevance of a delimited and level-headed ethical analysis.

Bringing together researches from three fields (Computer Science, Linguistics, and Philosophy), we approach a free Machine Translation (MT) tool, Google Translate (GT). Our goal is to show that well-delimited, multidisciplinary, technically supported approaches that avoid sensationalist discourse can lead to more epistemically consistent level-headed normative-critical reflections.

2 Automatic translation with machine learning: a computational account

Since the launching of ChaGPT in November of 2022, a lot has been said about Large Language Models (LLMs), their impacts on academic life, and the impacts on ethical discussions associated with them. Despite the advancements achieved in AI models in recent years, these language models are not new. Dating back to 2010, automatic translators started becoming accessible, allowing users to enter texts in a specific language in a search engine (the source language) and receive the translation in another language (the target language).

The evolution in Natural Language Processing (NLP) research, and access to more powerful hardware, has enabled the improvement in recent years of the quality and effectiveness of automatic translators, mainly via the employment of artificial neural network models. One significant step worthy of mention was the result of the research made by Google in 2013 into a new way of representing texts (Mikolov *et al*., 2013), revolutionizing the paradigms for developing automatic translators.

Currently, it is possible to point a smartphone camera and view translated text in real-time, it is also possible to copy the content of the camera from the screen. The benefits go beyond written language, they also include the possibility of converting spoken language to another language, using spoken language translators. All of these translation solutions permeate everyday society, facilitating numerous commercial, social, and political activities, as well as facilitating more inclusive access to information. In the academic world, among other benefits, the models also allowed countless students access to important learning resources.

NLP is the field of research in Computer Science responsible for seeking solutions requiring the computational treatment of a language, written or spoken. NLP is tied to AI, but also to computational linguistics. Currently, the leading paradigm in NLP, i.e., the one defining how the knowledge about a language will be expressed, is the neural paradigm. In it, the main technology used to give your programs 'intelligence' is associated with Machine Learning (ML) algorithms, especially with deep neural networks (Deep Learning).

What characterizes ML is providing large amounts of data about what must be learned to the algorithms created to 'learn' it, in other words, large amounts of data about a concept or a task (Nunes *et al*., 2023) (for instance, summarizing, translating or generating new texts, in the famous generative version).

When developing applications for automatic translation, large sets of text[4] are the knowledge source used to 'teach' neural networks how to create a model. An artificial neural network is an ML model inspired by the structure and functioning of the human brain. A neural network is made up of individual processing units, called artificial neurons, they receive one or more inputs, and then they perform a mathematical operation on them, producing an output. Artificial neurons are interconnected, they form layers, and each connection has a weight that adjusts the influence of the output from one neuron on the input of another one. During training, neural networks adjust the weights of the connections to minimize the error between expected and desired outputs (Aggarwal, 2018). In the case of translating using the neural paradigm, several layers of artificial neurons are used to learn how to translate a source sentence into a target sentence.

Building an automatic language translator anchored on neural networks, as is the case of Google Translate, is a complex process. It requires a high-quality training data set. For deep neural networks to succeed, the volume of input text needs to be quite significant, on the magnitude of billions of words, only then it allows the generation of a good performance model, meaning, a good model of the whole language.

Next, it is required to pre-process the input text with tokenization[5] and coding of the text. The text is transferred to a language that is understood by computational algorithms, while humans understand a word as a sequence of letters, a computer does not understand symbols, letters, or words, it understands only numerical representations. Thus, in this step, words (or subwords) are converted into numerical representations using, for example, a word embedding technique, such as Word2Vec (Goldberg, Levy, 2014).

Distributional semantics is the most used representation approach in NLP, since the launch of the Word2Vec project by Google in 2013 (Mikolov *et al.*, 2013). In this approach, words or subwords are represented through vectors of real values, known as embeddings, which encode the meaning of words based on their distribution in texts (Harris, 1954). Each word is represented as a point in a multidimensional vector space, constructed from the distribution of its neighboring words. This type of approach can map morphological, syntactic, and semantic characteristics to a vector space (Castilho, Caseli, 2023), and, through algebraic operations on the vectors, it can discover complex similarities between words. The insight is that similar linguistic context between words tends to also mean similar or approximated meanings. For example, the word "big" is similar to "bigger", and the same for "small" and "smallest" (Mikolov *et al.*, 2013). By using bilingual embeddings, it is possible to find the similarity between languages, for example, finding the similarity between words in English and Portuguese (for instance: my – meu). For this reason. embeddings are used as a form of language representation in neural models of machine translation (Castilho, Caseli, 2023), as they are capable of encoding subtle, but important, information about the relationships between words. For example, if a model learns something about the relationship between Paris and France (e.g. they speak the same language), there is a good chance that the same will occur with Berlin and Germany and with Rome and Italy. Notice that something obvious to human intelligence was only possible after years of study and after Google (Mikolov *et al.*, 2013) analyzed millions of documents collected from Google News to discover which words tend to appear in similar sentences.

However, the word embedding scheme is not able to capture an important fact about natural language; words often have multiple meanings. In Portuguese, typical examples are *manga de camisa* (part of a piece of clothing), and *pé de manga* (the tree of the mango fruit). Once more, solving ambiguities like this is simple for human intelligence, context can be used as grounds for interpretation. But computationally, the task is not simple, as there are no straightforward or deterministic rules for doing it in AI.

[4] Also called *corpus.*

[5] Technical term used in NLP meaning the separation of words in the pre-processing phase of machine learning.

Because of this, another important technique enters the scene, helping current automatic translators to be as precise (namely, contextualized embeddings). The two main methods for generating this type of embedding are with the recurrent neural network (Peters *et al.*, 2018) and Transformer classes of architectures (Vaswani *et al.*, 2017).

The neural network's architecture is defined in the next step of development, in which the construction of the neural model is carried out. The architecture describes how layers, neurons, and connections are organized to perform learning-related tasks. Furthermore, the architecture of a neural network includes hyperparameters such as the learning rate, and the batch size, referring to the number of training examples being used in each iteration of the network's training, and the number of training periods. The architecture used by Google Translate is the state of the art in machine translation. It is based on the Transformer architecture (Vaswani *et al.*, 2017), which is a neural network that uses attention mechanisms to process and translate text.

After these steps, the neural network model is trained, which may involve countless hyperparameter adjustments to find the best combination of weights and other measurements until a sufficiently accurate model is obtained. However, an important innovation of this particular neural paradigm is that, by using attention and memory mechanisms, the neural networks do not need explicitly labeled data (input-texts). Instead, they learn by trying to predict the next word in common passages of text, making it so any written material is suitable for training these models.

One can observe through the technical concepts put forward that, even if the automatic translation is made by AI, it does not follow that those technologies are intelligent, they only allow such applications to act more intelligently. The current models are able to generate surprising results in terms of translation and coherence, many times generating text indistinguishable from text produced by humans. Nevertheless, it is important to observe that those texts are essentially a made-up sequence of words, the models form statistically probable sentences in a specific idiom. These sentences are generated from textual input via models, but the models themselves have no understanding of the generated content. Instead, they learn during training common linguistic patterns derived from the data present in the corpus, subsequently reproducing these patterns in new outputs, as observed by Bender *et al.* (2021), similar to stochastic parrots. Furthermore, once one understands how texts are converted to numbers, that vector calculations are applied to them, and that neural networks are tangles of mathematical equations, it becomes clear why these approaches are called black boxes. Since they do not allow inquiry into how a result or conclusion was obtained, it also becomes clear why they are pointed out as having negative ethical implications. To add insult to injury, one can consider that large-scale language model training datasets often consist of a vast amount of linguistic data collected from the internet, and given the imbalance in internet access, it is likely that the training data is not representative. For that reason, it does not take into account the cultural and linguistic diversity that exists (Castilho, Caseli, 2023), reminding us of another ethical implication, the famous biases. All of those steps for the development of a language model, described here very briefly, require many hours of training and immense computational power. So, they also consume a lot of electrical energy before the final application is ready to the users (Ferro *et al.*, 2023).

But everything said so far also allows us to better consider the ethical implications of this type of technology, as well as its positive impact on people's lives. While many negative implications are highlighted, no different for the case of NLP applications, overall the balance is positive in favor of AI. In addition to our own multidisciplinary reflection presented throughout the paper, there is evidence in the literature that AI has the potential to impact the achievement of the 17 Sustainable Development Goals (SDGs) (United Nations, 2015), much more positively than negatively (Unesco, 2021). Such works (Vinuesa *et al.*, 2020; Saetra, 2021; Theodorou, Nieves, Dignum, 2022) are dedicated to analyzing and evaluating these effects based on the benefits and harms that AI produces for the 169 goals represented in the SDGs. The balance is positive in favor of AI. It helps to achieve 134 goals, while it has the potential to inhibit 59 of them.

3 A linguistic analysis of Google Translate

In the Generative Grammar Theory (Chomsky, 1969), it is assumed that the human brain is modular and one of the modules of human (and only human) cognition is Language. Therefore, this results in the idea of an innate linguistic component originating from the sphere of human biology. Comparatively, within linguistic studies, the term "Grammar" is recognized as a polysemic word, especially in theoretical debates. Therefore, in the context of this paper, it is important to understand that it refers to a set of rules internalized by the native speaker in the Language Acquisition phase (Pinker, 1984). This phase of Language Acquisition is also known as the Critical Period of Language Acquisition, which lasts from zero to approximately six years of age (Pinker, 1984). The period is one in which the brain is an open window to new possibilities. Based on primary data from the mother tongue, the native speaker can recognize the rules of a particular language. In other words, a set of non-conscious rules analogous to a computer program that builds an infinite number of sentences from a finite number of words (Pinker, 2004).

To better understand how this works, it is important to start from the acknowledgment of Universal Grammar, the very core mechanism that enables us to develop a particular language with so much skill in such a short time. Chomsky (1988) calls this Plato's Problem: how it is possible to know so much in so little time. The solution more consistent with this reality is assuming a Faculty of Language constituted by the characteristics exposed above: task-specificity and innateness. A question that follows from this abstraction is how it is that Language's functioning mechanisms work in such a manner. Following this assumption, what is understood as Language is a module constituted by three other submodels. One of them is the only one that generates linguistic materials, which we named Syntax. Generative Grammar arises exactly from this derivational sphere. Such a central module sends abstract material to the other Language submodules that do not generate linguistic material, rather, they interpret it: Phonology and Semantics. One must bear in mind that the abstract material needs to have sound and meaning, so it also has a communicative function. The syntactic module is universal, while the interpretative ones are parametric, that is, they concern themselves with particular rules of a specific language, such as Portuguese, English, Paumarí, or LIBRAS (Brazilian sign language). The theory gives rise to a dichotomy: Competence and Performance. If Competence is our unconscious knowledge, performance is the counterpart in our daily linguistic practice.

According to Chomsky, since the internal grammar creates algorithms that are unaware of the language: "each expression consists of two complexes of features at language's "interfaces," the phonetic interface and the semantic" (Chomsky, 2009, p. 42). Assuming these to be the basis for the development of natural languages, would it be possible to make a comparison with AI, considering that programs like ChaGPT turn out to fulfill tasks that previously could only be achieved by human beings? What are the limits and implications of this process in the development of technologies in parallel to what is known about the human Language?

It is common knowledge that AI does manipulate a combination of huge volumes of digital data and intelligent algorithms. Respectively, they allow the system to read and interpret both patterns and information, learning automatically. Almost exactly what human babies do in the Language Acquisition phase. As a parallelism, we can consider some comparative factors between human language and Artificial Intelligence, even though the computer is faster, it only performs tasks for which it was programmed, while the brain analyzes new and unknown situations, and past experiences, being able to react to them. Abstraction is also strictly human. For example, we can imagine things, even those that have not been experienced. To imagine a pineapple on the moon is perfectly possible for a human, but it is a difficult task for a computer, even though it recognizes the compositionally of the parts of the proposition (a feature related to the creativity present in human language).

Humans lose to computers in information processing speed. While it takes us a few milliseconds to process information, the computer takes less than a nanosecond. However, if one thinks that, in terms of memory, humans would lose by great margins, one would be wrong. A typical PC has 1 TB (Terabyte) of hard drive storage. Our brain, according to an estimate by the Salk Institute for Biological Studies, in the USA, would have the capacity to store up to 1,024 TB (Ossola, 2016).

In regards to the limits of Human Language, as it was said above, the Language Acquisition phase itself is available until 12 years of age at the most, with the best performance up to 6 years of age; so after that, one has already an obstacle to the Language acquisition with their native speaker skills. After the whole phase, Language Acquisition is compromised and we move on to what we know as Learning. Although we don't even remember any difficulty we had in acquiring our mother tongue as children, anyone who has tried to learn a language after the age of 12 knows that the task is no longer so simple. If we are learning English and we are native speakers of Brazilian Portuguese, for example, we have to learn that order, nucleus, and modifier change, so *menina* (girl) *bonita* (beautiful) becomes beautiful girl, not girl beautiful. When inflecting verbs in their tenses, we gradually realize that in English the tense is not always marked on the verb, sometimes the tense appears in separate future or conditional particles, *will* and *would*, respectively. We are also learning that, although in Portuguese we can leave the space for the subject empty without major repercussions, the same does not apply in English, which does not accept the subject position to be a phonological void. Therefore, if the path in terms of difficulty for a native English-speaking child and a Brazilian child is the same when acquiring their respective native languages, the exact same task for an adult is much more arduous and conscious.

Having passed the limit of what is called Acquisition, we move on to learning a second language or a foreign language. There are limits concerning human beings, some imposed by biology itself and others arising from social issues, such as access to courses, ideological barriers, lack of time and/or interest. All of this can result in some impediments and difficulties in being able to access some practical opportunities. The problem of the limits between Acquisition and Learning, in principle, does not affect AI, which would provide the latter with an advantage over humans. Especially, when considering the ability to store and identify patterns after the Critical Period of Language Acquisition. For a machine, it doesn't matter how old it is when exposed to data from a language. What matters most is the quantity and quality of data the machine was exposed to.

The specific linguistic study brought forward here was carried out using Google Translate, taking into account the specificities of their AI model and its ethical repercussions. Such specificity concerns itself precisely with what was outlined in the previous paragraph. It is visible how much Google Translate from Portuguese to English has improved considerably, but the opposite – English to Portuguese – has seen a less significant leap. The answer is on the quantity and quality of user data. Many more people need to translate texts into their native languages from English, which means that the English language has a much larger database, resulting in greater precision in the program's task.

Everything so far indicates that we can highlight some positive points of using this AI, on average faster learning than humans (because it is through humans that data is provided and adjusted), including when taking into account second language data, and greater than human data storage capacity.

We can also conjecture positive consequences of GT and similar models for the practical life of humans, as those observed in Education, taking into account the access it provides to texts in foreign languages (especially English). It promotes the possibility of initiating exchanges, even if mediated by AI. Socially, the task is much more available to certain groups than others. Among other impediments. one has to consider the price of courses, and the need to conciliate with other work, home life, and other studies.

Since our objective is to thoughtfully analyze the debate, it involves showing positive points, but also pointing out some negative ones, such as the fact that AI does not have critical capacity, preventing it from identifying errors on its own. In this regard, we could consider that some humans also have this limitation, but not the experts. Another worthy fact that deserves to be highlighted concerns the fact that human language has algorithms for the functioning of its grammatical architecture, but also has tools to deal with creative phenomena providing economy to the system; those include word polysemy, structural ambiguity, creation of new words, structures that are phonologically null (but that have semantic interpretation), and reserved syntactic space. Polysemy is economical in the sense that it is enough for the syntactic system to generate just one construction, while the semantic interpretative system is responsible for the different contextualized meanings. The same goes for structural ambiguity. The creation of new words proves that the human linguistic system operates with a few rules that can generate infinite and creative combinations. Finally, the syntax-phonology interface also shows the functioning of a model capable of accounting for any and all generated linguistic constructs. How does Google Translate handle generated linguistic components such as those mentioned? Given its operational limitations, can we also create conjectures about some of its negative consequences, such as the fact that it does not emerge from the Acquisition and Learning dichotomy? Does the fact that there is no natural language, i.e., an acquired language, generate consequences for the artificial system?

We went to Google Translate to ascertain certain guesses, aiming to consult some specific points regarding translations. We were always starting from Portuguese to English, as it has the best performance. We start by translating idiomatic expressions from Portuguese to English. But before we continue, it is important to say that the idiomatic expressions were constructions whose semantic interpretations are not derived from the regular calculus of its parts, such as *João comeu uma maçã*. By knowing that *João* is a person's proper name, by knowing the meaning of *comeu* (i.e., has eaten/ate), and also by knowing the meaning of *maçã* (apple), we know what *João comeu uma maçã* means. There are no surprises. When trying to translate it into English using Google Translate, the result was expected: John ate an apple, a compositional translation, the result of calculating the relation between the parts.

Languages do not just have constructions whose semantic interpretations always come from calculations resulting in compositional and well-behaved meanings. In Portuguese, we can say that a person "jumped over the fence" (*pulou a cerca*), but the meaning is not just that a person jumped over an artifact that separated the space between a house and the street, it can also mean that the person betrayed their partner. There are well-known books, such as Millôr Fernandes' (1988) *The Cow Went to the Swamp*, in which one plays with the translation of idiomatic expressions from Portuguese into English. There are also photo books with idiomatic expressions such as "*engolir sapo*" (literally meaning to swallow a frog), "*chorar pelo leite derramado*" (to cry over spilled milk), "*pendurar as chuteiras*" (to hang up your soccer shoes) (Ballardin, Zocchio, 1999). These creative works cause laughter and amazement precisely because we know that such expressions should not be interpreted literally, for the most part. The ones that have a compositional/literal counterpart are always in a syntactic and/or pragmatic context that does not allow for doubt.

When we put the sentence *Joana chutou o balde e acordou tarde* into Google Translate, the tool suggests the following translation: Joana kicked the bucket and woke up late. At first glance, one cannot point out a specific problem, since the literal translation is exactly the one provided by Google Translate. Obviously, human Language needs to have a higher number of compositional data when compared to idiomatic data, to achieve a high level of performance. However, such a translation creates some problems, especially for those with little English language skills. The first problem arises from the fact that kicking the bucket has two interpretations in Portuguese, one literal, which consists of someone having kicked a bucket, and another non-literal/idiomatic/irregular, in which "kicking the bucket" means "having no regard for the consequences". If faced with just *Joana chutou o balde*, the system could have simply disregarded idiomatic reading and opted to guarantee the more frequent compositional reading. But the entire sentence has a coordination imposing an idiomatic reading that was not considered by the system (Joana had woken up late because she had/chose to have no regard for the consequences). In other words, the syntactic structure, primordial and automatic in humans, does not seem to be a relevant factor for the AI, resulting in an error in the system's interpretation. The system generated syntactically grammatical information, but without a valid interpretation, as it did not consider either syntax or pragmatics.

We can detect a second problem when translating *chutou o balde* as "kick the bucket", the system was not aware of the fact that this is also an idiomatic expression in English, which generates a third problem, because the meaning of it the expression is different in English, it means "to die". Following this problem, the chain of issues increases, and this leads also to a fourth problem, the so-called Encyclopedic or real-world knowledge problem: people bear in mind while reading it that a dead person cannot wake up late.

Another interesting search was carried out on Google Translate with lexical items/polysemous words, words that have more than one plausible meaning. We tested the translations with the verb *bater* in the sentence. Researched the translation of *Joana bateu a porta do carro*, GT correctly translated it as Joana slammed the car door. When we changed the sentence to *Joana bateu à porta*, again the artificial translation tool identified the polysemy and changed the verb to Joana knocked on the door. However, it is possible to say in Brazilian Portuguese that *Joana bateu um prato de comida*. In this case, we have as a result a non-compositional meaning for the verb *bater*. The verb, in this context of verbal complementation, means that Joana ate a large and/or entire edible content from the plate, which is the container. However, Google Translate gives the following translation: Joana hit a plate of food. It is true that "hit" is a plausible translation for the verb *bater*, but when used in contexts such as *Joana bateu no menino*, a sentence meaning "Joana hit the boy", as in was smacking a male child.

Regarding a specific parameter of Portuguese when compared to English, the null subject, Google Translate has been increasingly efficient in relation to it. For example, nowadays, if we enter "*comprei um sapato ontem*". The system identifies the lack of subject and correctly inserts it in the English translation, taking into account that the subject must always be phonologically expressed in this target language, the translation correct being "I bought a shoe yesterday". The tool also identifies null subjects in embedded/subordinate sentences, like *Paulo disse que vai viajar*. GT gives the correct translation "Paul said he is going to travel". It is only when we use verbs of natural or impersonal phenomena that the GT has a worse performance. With the prompt *ontem choveu*, it translates correctly: yesterday it rained. But if one changes the order of the constituents to *chuveu ontem*, the translator assumes it is a question, even without the corresponding punctuation, so it translates as "Did it rain yesterday".

Perhaps the most "egregious" issues are the ones regarding technical words. For example, if one tries to translate "features" from English to Portuguese for a text in formal linguistics, even though it is expected the word "*traços*" or the expression "*feixe de traços*", it is common for the word "*características*" to appear. But if we test the same word reversing the translation order, the program also makes an error, as "*feixe de traços*" is translated as "trace beam".

In summary, we have highlighted some limitations of the Artificial Intelligence systems operating on linguistic apparatus, as their model, human Language, operates on a greater number of regular constructs. In most cases, the artificial system has the potential to generate translations acceptable from the sentence's grammaticality point of view. A more naive reader could think that testing sentences might be more positively viable for the artificial system. For this reason, the frequency of mistakes in compositional constructions has been decreasing over time. However, texts are made up of sentences, whether simple or compound periods, like the coordination tested above.

4 Toward thoughtful ethical approaches

The ethics of machine translation involves multiple emphases and approaches, on which they apply possible uses of MT in different environments. They also apply to the respective issues raised, as well as to the implications for the various people involved/interested. For example, its use in the L2 classroom context can be considered a problem to be dealt with (Ducar, Schocket, 2018). The presence of biases should be the subject of attention (Prates, Avelar, Lamb, 2020). The (in)accuracy of the translation of technical terms, as well as their implications, should not be ignored (Patil and Davies, 2014). Differences in performance, when comparing languages and textual genres, continue to remain an important factor to be considered (Almahasees, Meqdadi, Albudairi, 2021; Adiel, 2021). There are also the implications of the technology for professional translators (Kenny, 2011). In other words, there are a variety of poignant ethical issues, many of which are reasonable to take our time with, but there is no reason to throw out the baby with the bathwater.

Given the analysis carried out in the previous sections, there is no need to fear MT/GT: it will not replace humans in linguistic processing. There is also no reason to overestimate it. The technology will continue to make mistakes (just like humans do), but it will continue to learn and improve (idem). But also, there is no reason to underestimate it. Being an AI-based tool, it can be quite useful, and in the case of its usage in the academic environment, it would be irrational to discard it. Although many papers point out the limits of the tool in terms of *acceptability*, specifically in regard to the grammatical quality of the results (which can vary considerably depending on the source languages and target languages involved), the benefits of its use in academia seem to outweigh the problems. This is especially the case in contexts where the percentage of the population proficient in English —the predominant language in science— is quite small (the Brazilian case, for example). The perception of the benefits of something may vary from individual to individual, by virtue of subjective factors. Among a plethora of individual factors, personal perception of benefits related to acceptability, usability, and satisfaction can vary considerably depending on age, education, knowledge of languages, and cognitive aspects related to learning. In addition, it can also vary according to the level of demand to which a given user is subjected (Kasperé *et al.*, 2023). Let's compare three hypothetical groups, formed by individuals speaking the same language (Brazilian Portuguese, for example). Let's assume they have high, medium, and low proficiency in the English language (the GT is more efficient). We can infer that the importance of the tool may vary from completely unnecessary to extremely fundamental. As a tool is not beneficial if and only if it is equally beneficial for everyone, therefore, there is no reason to discard it.

It is also worth highlighting that GT is a free access tool, which tends to an accelerated improvement of its performance. In theory, the optimization of its results will increase its benefits for all users (Groves, Mundt, 2015). In terms of hybrid intelligence, i.e., the benefits of the union between human and artificial intelligence, MT/GT sounds promising, so it seems ethically justifiable to continue using it, as well as investing collectively in its improvement.

In assuming the stance that, from an ethical point of view, identifying and mitigating technological risks is equally as important as identifying and promoting benefits from the technologies, it is necessary to invest in a less unbalanced ethical debate on AI. In this sense, both *status quo bias* and *progress bias* must be overcome, in the name of empirically supported and ethically considered approaches.

In the end, we hope to have shown that well-delimited, multidisciplinary, technically consistent approaches that do not adhere to hyperbolic and sensationalist discourse are able to conduct us to a more precise and clear understanding about present and future AI applications. Such approaches can help us identify, in a level-headed way, the AI's actual and potential benefits and risks. This makes it possible to deal with AI in an adequate and realistic way, making the important debate about this type of technology move beyond just another hype that tends to be surpassed by another soon.

References

ADIEL, M. A. E. 2021. Automatic Translation of Arabic Classic Poetry; Case-Study of *Google Translation*. *International Journal of Humanities Social Sciences and Education*, **8**(8): 81-95.

AGGARWAL, C. C. 2018. *Neural Networks and Deep Learning: A Textbook*. Berlin, Springer, 506 p.

ALMAHASEES, Z.; MEQDADI, S.; ALBUDAIRI, Y. 2021. *Journal of Language and Linguistic Studies*, **17**(4): 2065-2080.

BALLARDIN, E.; ZOCCHIO, M. 1999. Pequeno Dicionário de Expressões Idiomáticas. São Paulo: Editora Salesiano, 159 p.

BENDER, E. M. *et al.* 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *In:* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery. Available at: https://doi.org/10.1145/3442188.3445922.

BUCHANAN, A. 2011. *Beyond Humanity? The Ethics of Biomedical Enhancement*. Oxford, Oxford University Press, 286 p.

CASTILHO, S.; CASELI, H. M. 2023. Tradução Automática - Abordagens e Avaliação. *In:* H. M. CASELI; M. G. V. NUNES (orgs.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. Available at: https://brasileiraspln.com/livro-pln/1a-edicao/parte8/cap18/cap18.html

CHOMSKY, N. 1969. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press, 251 p.

CHOMSKY, N. 1988. *Language and Problems of Knowledge –The Managua Lectures*. Cambridge: MIT Press, 216 p.

CHOMSKY, N. 2009. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Cambridge, UK: Cambridge University Press, 158 p.

DICARLO, C. 2016. How to Avoid a Robotic Apocalypse: A Consideration on the Future Developments of AI, Emergent Consciousness, and the Frankenstein Effect. *IEEE Technology and Society Magazine*, **35**(4): 56-61.

DUCAR, C.; SCHOCKET, D.H. 2018. Machine Translation and the L2 Classroom: Pedagogical Solutions for Making Peace with Google Translate. *Foreign Language Annals*, **51**(4): 779-795.

FERNANDES, M. 2001. *A vaca foi pro brejo = The cow went to the swamp*. Rio de Janeiro: Record, 128 p.

FERRO, M. *et al.* 2023. Towards a Sustainable Artificial Intelligence: A Case Study of Energy Efficiency in Decision Tree Algorithms. *Concurrency and Computation: Practice and Experience*, **35**(17): e6815.

GERACI, R. M. 2006. Spiritual Robots: Religion and Our Scientific View of the Natural World. *Theology and Science*, **4**(3): 229-246.

GERACI, R. M. 2007. Robots and the Sacred in Science and Science Fiction: Theological Implications of Artificial Intelligence. *Zygon*, **42**(4): 961-980.

GERACI, R. M. 2008. Apocalyptic AI: Religion and the Promise of Artificial Intelligence. *Journal of the American Academy of Religion*, **76**(1): 138-166.

GERACI, R. M. 2010a. The Popular Appeal of Apocalyptic AI. *Zygon*, **45**(4): 1003-1020.

GERACI, R. M. 2010b. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford: Oxford University Press, 248 p.

GERACI, R. M. 2022. *Futures of Artificial Intelligence: Perspectives from India and the U.S.* Oxford: Oxford University Press, 240 p.

GOLDBERG, Y.; LEVY, O. 2014. "word2vec Explained: Deriving Mikolov *et al.*'s Negative-Sampling Word-Embedding Method". Available at: https://arxiv.org/pdf/1402.3722.pdf.

GROVES, M.; MUNDT, K. 2015. Friend or Foe? Google Translate in Language for Academic Purposes. *English for Specific Purposes*, **37**: 112-121.

HABERMAS, J. 2003. *Mudança estrutural da esfera pública: investigações quanto a uma categoria da sociedade burguesa*. Rio de Janeiro: Tempo Brasileiro, 568 p.

HARRIS, Z. 1954. Distributional Structure. *Word*, **10**(2-3): 146–162.

HOFMANN, B. 2019. Progress Bias versus Status Quo Bias in the Ethics of Emerging Science and Technology. *Bioethics*, **34**(3): 252-263.

JUNGES, J. R. 2019. Falácia dilemática nas discussões da bioética. *Revista de Bioética*, **27**(2): 196-203.

KASPERÉ, R. *et al.* 2023. Is Machine Translation a Dim Technology for Its Users? Na Eye Tracking Study. *Frontiers in Psychology*, **14**: 1076379.

KEARNEY, E.; WOJCIK, A.; BABU, D. 2019. Artificial Intelligence in Genetic Services Delivery: Utopia or Apocalypse? *Journal of Genetic Counseling*, **29**(1): 8-17.

KENNY, D. 2011. The Ethics of Machine Translation. *In:* New Zealand Society of Translators and Interpreters Annual Conference 2011, 4-5 June 2011. Available at: https://core.ac.uk/download/pdf/11311284.pdf.

LEMAY, D. J.; BASNET, R. B.; DOLECK, T. 2020. Fearing the Robot Apocalypse: Correlates of AI Anxiety. *International Journal of Learning Analytics and Artificial Intelligence for Education*, **2**(2): 24-33.

MIKOLOV, T. *et al.* 2013. Efficient Estimation of Word Representations in Vector Space. *In:* Proceedings of Workshop at ICLR.

NOY, I.; UHER, T. 2022. Four New Horsemen of na Apocalypse? Solar Flares, Super-volcanoes, Pandemics, and Artificial Intelligence. *In:* I. NOY; S. MANAGI (eds.), *Economics of Disasters and Climate Change*, Switzerland, Springer, p. 393-416.

NUNES, M. G. V., *et al.* 2023. Questões éticas em IA e PLN. *In:* H. M. CASELI; M. G. V. NUNES (orgs.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. Available at: https://brasileiraspln.com/livro-pln/1a-edicao/parte10/cap24/cap24.html.

OSSOLA, A. 2016. *The Human Brain Could Store 10 Times More Memories Than Previously Thought*. Available at: https://www.popsci.com/human-brain-could-store-10-times-more-memories-than-previously-thought/.

PATIL, S.; DAVIES, P. 2014. Use of Google Translate in Medical Communications: Evaluation of Accuracy. *BMJ*, **349**: g7392.

PAULUS JR., M. J. 2023. *Artificial Intelligence and the Apocalyptic Imagination: Artificial Agency and Human Hope*. Eugene, OR: Cascade Books, 162 p.

PETERS, M. E. *et al.* 2018. Deep Contextualized Word Representations. *In:* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, **1** (Long Papers). Anais...Association for Computational Linguistics. Available at: https://doi.org/10.18653/v1/n18-1202.

PINKER, S. 1984. *Language Learnability and Language Development*. Cambridge MA: Harvard University Press, 474 p.

PINKER, S. 2004. *O instinto da linguagem: como a mente cria a linguagem*. São Paulo, Martins Fontes, 640 p.

PRATES, M. O. R.; AVELAR, P. H.; LAMB, L. C. 2020. Assessing Gender Bias in Machine Translation: A Case Study with Google Translate. *Neural Computing and Applications*, **32**: 6363-6381.

ROBERTSON, A.; MACCARONE, M. 2023. AI Narratives and Unequal Conditions. Analyzing the Discourse of Liminal Expert Voices in Discursive Communicative Spaces. *Telecommunications Policy*, **47**(5): 102462.

RUEDA, J. 2023. Problems with Dystopian Representations in Genetic Futurism. *Nature*, **55**: 1081.

SAETRA, H. S. 2021. AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System. *Sustainability*, 13(4): 1738.

SEIFERT, F.; FAUTZ, C. 2021. Hype After Hype: From Bio to Nano to AI. *NanoEthics*, **15**: 143-148.

SHERMER, M. 2017. Apocalypse AI. *Scientific American*, **316**(3): 77.

THEODOROU, A.; NIEVES, J. C.; DIGNUM, V. 2022. Good AI for Good: How AI Strategies of the Nordic Countries Address the Sustainable Development Goals. Available at: https://arxiv.org/pdf/2210.09010.pdf.

UNESCO. 2021. *Preliminary Report on the First Draft of the Recommendation on the Ethics of Artificial Intelligence*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000374266.locale=en.

UNITED NATIONS. 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*.

VASWANI, A. *et al.* 2017. Attention is All you Need. *In:* Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, December 4-9. Long Beach, CA, USA.

VILAÇA, M. M. 2021. Contra a perfeição, o melhoramento humano ou pela dádiva? Uma análise dos argumentos de Michael Sandel sobre a engenharia genética. *Síntese*, **48**(152): 779-805.

VILAÇA, M. M.; LAVAZZA, A. 2022. Not Too Risky. How to Take a Reasonable Stance on Human Enhancement. *Filosofia Unisinos*, **23**(3): 1-16.

VINUESA, R. *et al.* 2020. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, **11**(1): 233.

Submetido em 06 de novembro de 2023.

Aceito em 11 de janeiro de 2024.