Parrots All the Way Down

Controversies within AI's Conquest of Language

Jonathan Roberge & Tom Lebrun

Abstract: *Today's deployment of automated semantic models such as Google's BERT or OpenAI's GPT-3 is a remarkable challenge for the inscription of hermeneutics at the very heart of the social sciences project. Artificial intelligence is indeed conquering language. There are three important implications. First, we must take the power and possibilities of such models seriously – that is, the recent history of technological advances in deep learning and the modi operandi of these interpreting machines, particularly their two-way reading and "Transformer" architecture. Second, a better comprehension is required of the type of understanding involved – mainly how the calculation of probability, thresholds and variation, for example, vectorizes language as if to parrot it back. Our analysis takes note of the dismissal by Google of the researcher Timnit Gebru, precisely around the text "On the Danger of Stochastic Parrots", to show how the value of natural language processing (NLP) models lies in the kind of world they put forward as well as in their reference to a precise context. Finally, this should help to circumscribe the current economic, political and even ethical aporias concerning these models, including the fact that the platforms developing them overlook crucial real-world effects of the way in which they advance the extraction, commodification and instrumentalization of meaning. Ultimately, it is this close link between meaning and the displacement of power centers that becomes the central issue of Critical AI Studies.*

1. Introduction

The year 2020 was marked by yet another – substantial – crisis at Google with the departure-firing of researcher Timnit Gebru over the submission of the paper "On the Danger of Stochastic Parrots: Can Language Models Be Too Big?" (Bender et al. 2021). The story went from anecdotal to scandalous when the

company asked for the article to be withdrawn or for the names of the Google employees who contributed to it to be removed. For Jeff Dean, Google's director of AI, the matter was settled in that the work in question "didn't meet our bar for publication" (Hao 2020). However, Gebru's rebuff and the large number of supporters who rallied to her cause – in this case, more than 2,000 of the company's employees signed a letter asking for demanding greater transparency in the management of its internal affairs (Wakabayashi 2020) – had not been factored in. So, who had demanded the withdrawal of the article and why exactly? This is a question in the form of an ultimatum that remains unanswered. "Timnit wrote that if we didn't meet these demands," Dean writes, "she would leave Google […]. [W]e accept and respect her decision." The break-up was, to all intents and purposes, complete. On Twitter, the researcher expressed her dismay and called out her former boss: "@jeffdean I realize how much larger language models are worth to you now."[1]

It is precisely on this notion of worth, or value, that we want to focus here. The double meaning of the word indicates that it refers more or less distinctly to something economic as well as axiological. Indeed, the very ambiguity of the word and of its use in the tweet is what makes it emblematic of the most important issues surrounding the current deployment of natural language processing (NLP) models. While many have seen the Gebru scandal as a matter of ethics and labor relations, few have been willing to consider the question in its fullest sense and thus explore the ways in which it represents a fundamental hermeneutical issue. Yet the question deserves to be asked: have AI and the latest advances in deep learning enabled the development of (too) big, powerful and deep models? And what can such terms mean, apart from a technical point of view? Do meaning and textuality, interpretation and understanding not become (too) impoverished as a result of their automated processing?

We argue in this chapter that the emergence of models such as Google's BERT or OpenAI's GPT-3 is today a remarkable challenge for hermeneutic disciplines in general, and for the inscription of hermeneutics at the very heart of the social sciences project in particular. It is therefore not a question of denying the rise of modeling or interpreting machines – or even their power and scope – but of examining the conditions of their possibility and significance. To put it in a nutshell: the emergence of these hermeneutic machines is an opportunity to think afresh about what a critical hermeneutics can represent within

1 Timnit Gebru, Twitter, https://twitter.com/timnitGebru/status/1334345550095912961. Last access: 19 July 2021. Emphasis added.

the social sciences and how it can enter into dialogue with or serve as a basis for the development of Critical AI Studies (CAIS) (Roberge 2020; Pasquinelli/Joler 2020). Specifically, this implies taking seriously the history and modus operandi of these language models and the way in which their various problems crystallized in the form of diverging views as part of the wider implications of the Gebru–Google conflict. This in turn implies a better understanding of the type of meaning at stake; that is, above all, the type of textual world deployed – or not – and the type of reading experience that this induces. Finally, it implies circumscribing the aporias of natural language processing models that are most often decontextualized and (re)translating or reinserting them in the social, political, economic and cultural reality from which they originate, particularly in the link between platform capitalism and the ethical desiderata of today.

Our analysis is articulated in three stages which correspond to the three implications mentioned earlier: i) taking seriously, ii) understanding and iii) circumscribing the aporias of NLP models. In the first section, we try to grasp these most recent models as social constructions and socio-technical assemblages (Schwartz 1989; Woolgar 1985). BERT – or Bidirectional Encoder Representations from Transformers – was introduced by Google in 2018 and later integrated into its main search engine. It collects information from Wikipedia, for example, and reads from right to left and back again to identify multiple parallel connections and predict missing terms. OpenAI's GPT-3 – Generative Pre-Trained Transformer – is newer; with its 175 billion parameters, it is said to outperform Google's model by 400 times in "encoding" textuality and thereby opening up a huge range of writing possibilities – journalistic, IT, administrative, etc. What these two models have in common is that they are not exactly black boxes, but rather the objects of a particular historical development which is for many the object of its difficulties and limitations.

In the second section, the overall meaning of this advanced automation comes under scrutiny. What are the implications of the epistemological conception promoted through these data architectures and statistical regressions? And of both the mediation and the recipient of language in this type of connectionist and cybernetic machine? These questions prompt a certain diversion through hermeneutics – that of Paul Ricoeur will be privileged here, partly because his notion of world allows us to think of a semantics, a reference and a "Being-demanding-to-be-said" of textuality, which gives the measure of how AI models sometimes, if not often, appear "shockingly good, and completely mindless" (Heaven 2020). In other words, this world of which Ricoeur speaks is

what can allow us to rethink the link between meaning and reflexivity. The latter is understood here not just as the reader's reflexivity, but also more broadly as the rediscovered reflexivity of the real-mundane world, in the context of society, culture and political economy.

In the third and final section, we therefore seek to develop a sociological and critical understanding of the deployment of these problematic, yet perfectly practical interpreting machines – BERT and GPT-3 – that are intruding on everyday life. The value of these models is inseparable from a market of data and meaning extraction in which some thrive more than others and for which, as the Gebru case shows rather well, ethics becomes a kind of justification and even commodity.

2. The drive to automate language: an all too brief history

The last few years have seen a major evolution in natural language processing. For the first time, language models based on a so-called Transformer architecture make it possible to generate texts that are sufficiently coherent to fool their readers, without relying on a deductive and symbolic logic previously decided by a programmer (Buchanan 2005; Balpe 1991). Based on the mechanism of machine learning, and particularly deep learning, this type of computer programming proposes to imitate some of the cognitive mechanisms of the brain, notably by means of artificial neurons – in reality miniature computer programs that activate or deactivate themselves according to the result of their calculation. As with the human brain, the strength of the mechanism lies in the networking of a large number of these miniature programs. This method, for a long time on the fringes of the AI field, suddenly came back into the spotlight during the 2012 ImageNet competition, won by Geoffrey Hinton's team thanks to the combination of great computing power, a vast data set and this method, which is rightly described as connectionist (Cardon/Cointet/Mazières 2018; Domingos 2015).

The recent evolution of NLP is marked by four significant changes, corresponding to four significant publications. The first paper, published one year after Hinton's great demonstration, was "Efficient Estimation of Word Representations in Vector Space" (Mikolov et al. 2013). Written by a team from Google – Jeff Dean is one of the co-authors – the paper proposes a group of language models called Word2vec, which aims to reconstruct the linguistic context in which words are used. Word2Vec, like most machine learning technologies,

relies heavily on the principle of regression, a method of statistical analysis that allows one variable to be placed in relation to its correlations with others. Roughly summarized, the technology involves locating the variable – the "meaning" – of a word in relation to the variables – the "meanings" – of other words around it. As its name suggests, Word2Vec aims to transform words into vectors, i.e., to model the information they contain using algorithms. In practice, Word2Vec "vectorizes" words using two distinct and complementary architectures. One, called CBOW (Continuous Bag of Words), seeks to predict a word according to the five words to its right and the five words to its left. The other, called Skip-gram, does exactly the opposite and seeks to predict the words in the context according to a given word. The logic is always predictive: the model must be able to assign the "right" vector to each word. Despite its success in the early 2010s, Word2Vec is severely limited. In particular, the language model assigns only one meaning per word and only vectorizes individual words, so that the meaning of even a relatively simple sentence continues to elude it (Horn 2017; Cusin-Berche 2003). We will return to this point later.

To address these limitations, Vinyals and Le – also from Google – published a paper shortly afterwards entitled "A Neural Conversational Model" (2015). This proposed, quite simply, to apply a sequential approach to Word2Vec to model the meaning of a text by linking certain sequences with others – thus forming a longer or "networked" form of text mapping (Sutskever/Vinyals/Le 2014). With this approach, modeling can now be applied to larger sequences, including sentences: the sentences preceding and following the target sentence are thus also taken into account and the model allows for a minimum of contextual consideration.

Despite this progress, Word2Vec-type systems are still based on the approach where a word can only have one meaning. It is this limitation that the article "Deep contextualized word representations" (Peters et al. 2018) aims to overcome. The authors propose a new architecture called Embeddings from Language Models, or ELMo, within which the model can now recognize the dynamic – moving, situational – nature of word meaning. In practice, each word is assigned a coefficient or "weight" according to its influence in the sentence. A word like "bow" can now have different meanings depending on a certain context – "I broke my violin bow" and "I am sitting at the ship's bow." Above all, ELMo makes it possible for the first time to consider modeling that does not learn from the text in a purely orderly way, by offering a "reading" in three different ways: first from left to right – from beginning to end; then in reverse – from end to beginning; and finally by combining the vectorized meanings

of both types of analysis. Also, ELMo marks the real beginning of pre-trained models, allowing users to avoid having to train their models from scratch on huge data sets – an extremely expensive practice, if only in terms of time and computing power.[2]

Finally, the paper "Attention is all you need" (Vaswani et al. 2017) marks the moment when the Transformer architecture virtually seals the field's fate. Previous sequential models had difficulty retaining information about the prioritization of terms among themselves: to take the previous example, information in a simple sentence – "I broke my violin bow", "I was sitting at the ship's bow" – was difficult to retain in longer sequences – "then the bow was splashed". The Transformer architecture moves away from this approach and its many problems in terms of memory, computational speed, word position, etc., by proposing to identify the context that gives meaning to words, which are then processed in parallel. This involves the use of both an encoder and a decoder – and indeed many of them on multiple levels acting and producing feedback in a cybernetic manner. The encoder transforms information into code by giving a calculated value to a word; a decoder does exactly the opposite, transforming code into information by "calculating" a word from a value.

The point to keep in mind here is that such architectures are based on neural networks and on the "deep learning" made famous since Hinton's demonstration of 2012, and in which layers and layers of encoders and layers and layers of decoders can be arranged without too many limits other than technical ones. Above all, the truly innovative character of the Transformer architecture lies in the attention mechanism implemented. The idea is to calculate a "weighted matrix product" – in other words, a matrix score that determines the level of attention that a word should have towards other words; some might also speak, more simply, of situational dependency. An encoder can thus compute several "attention heads" that work in a bidirectional way: an attention weight is computed as input and produces an output vector. The major advantage of this in-depth bidirectionality is that it allows the information to be

2 Unlike the image recognition field – where anyone could download pre-trained face recognition models from ImageNet, for example – the field of NLP appeared before ELMo to be a unified environment, in which each research group or company had to start from scratch, with its own data and available computing power. Drawing on the example of image recognition, different types of pre-trained models emerged at the same time as ELMo, such as ULMFit or the first OpenAI Transformer system. See Ruder 2018.

processed in parallel by the different attention heads, and therefore by the different encoder layers. This results in considerably reduced training times for language models compared with sequential approaches such as Word2Vec.

It is thus these new attention mechanisms specific to Transformer architectures that are at the source of the current successes of language models, in particular Google's BERT and OPEN AI's GPT-3, as spearheads of the ongoing battle waged by the GAFAM (Google, Amazon, Facebook and Microsoft) in their quest for the mastery of artificial intelligence (Thibout 2019; Horowitz 2018). BERT is still a relatively small model compared to GPT-3, as it has been pre-trained on about 3.3 billion words and has 345 million parameters (Devlin et al. 2018). Its main objective is to end formalized keyword searches, a goal that may seem trivial at first but is central to the company's mission statement – "to organize the world's information to make it universally accessible and useful". To achieve this, Google must enable its users to express themselves in the most natural, user-friendly and dialogical way possible.[3] BERT tries to achieves this objective by focusing on the encoder part of the architecture, the part that transforms information, the written or spoken request, but also texts to be translated, for example, into code and vector as to capture their contours: who does what, where, etc. BERT, in other words, and above all, "understands" in the sense of extracting the relevant elements as rendered in more encompassing sets. Its Transformer architecture is thus very flexible and functions as an interface between the natural input language (the query) and the output (the result). It should also be noted that BERT is open-source, which is part of a corporate strategy of value creation quite specific to Google,[4] to which we will return in the last section.

---

3 Prabhakar Raghavan, vice-president of Google, explains that the ultimate goal is to respond *directly* and *intelligently* to users' needs: "Let's say you are planning to go hiking on Mount Fuji […] Do my hiking boots suffice? Today, what you do is you transcribe it into hours of interaction with Google […] Wouldn't it be a lot better if you could […] let Google figure this out and address the need behind your query? […] I want to be able to get to a point where you can take a picture of those hiking boots and ask, 'Can these be used to hike Mount Fuji?'"; see Levy 2021.

4 "With this release, anyone in the world can train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU" (Nayak 2019). See also Devlin and Chang 2018. On Google's open-source strategy (as opposed to Microsoft's, in particular), see Janakiram 2017.

GPT-3 is at the time of writing the most powerful language model, trained on around 570 gigabytes of data and composed of 175 billion parameters (Brown et al. 2020). GPT-3 explicitly aims to generate text, according to its creators. Unlike BERT, therefore, it favors the decoder part of its architecture, the part that more precisely allows code to be transformed into information, i.e., inferring missing words, completing sentences, etc. Far from being open-source, GPT-3 is currently marketed via its Application Programming Interface (API), a choice which is obviously part of a corporate strategy that aims to control the economic ecosystem on which many future companies will be based. In a blog post, OpenAI reported in March 2021 that more than 300 companies were making use of this API, a number that keeps growing. For instance, applications already available include CopyAI, which can generate slogans and product descriptions for companies, and Fable, which can model characters from novels and talk to them (Scott 2020).[5]

Because they are socio-technical assemblages, it goes without saying that these models struggle to be perfect or even to live up to the rhetoric legitimizing their use and, more generally, all that is the magic of AI (Roberge 2020; Elish/Boyd 2018). The fact is that all is not well in the best of all NLP worlds, and that upon closer inspection, its deployment is more a matter of "garbage in, garbage out" – the so-called GIGO principle, as ironically referred to by scientists working in the field (Kilkenny/Robinson 2018). With regard to input, it should be seen that while language can apparently be computationally modelled, the Transformer architecture can only achieve this from a resource that is itself a social construct: the database. This dependence of language models on their training sources is quite widely discussed (Hutchinson et al. 2020; Roberge 2018), as any given Transformer architecture remains based on the principle of regression outlined earlier, which aims to locate a variable (a word) according to its correlations with other variables (the other words in the database). This simple mathematical procedure thus constructs an approach to language based on the principle of "winner takes all". Put differently, the language model promotes the most statistically probable language constructs according to the data set on which it is trained. Also, the choice of texts on which these language models are developed participates in a certain representation of the world, whose symbolic, if not ideological, dimension is often only revealed once the models have been applied – through the racist,

5 To see the companies directly: CopyAI, https://www.copy.ai/ [last access: 18 March 2021] and Fable, https://fable-studio.com/ [last access: 4 June 2021].

misogynistic or other biases that result from them. This is one of the most striking observations in the text that led to Gebru's dismissal from Google:

> GPT-2's training data is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29. Similarly, recent surveys of Wikipedians find that only 8–15% are women or girls (Bender et al. 2021: 4).

What is thus a problem on the input side becomes a problem on the output side, with the highest number of potential and proven slippages. One of the most feared applications in this respect is what is usually referred to as astroturfing, in which a plethora of micro-speeches is automatically generated as if to simulate a mass movement accrediting such organizations, ideas, etc. (Kovic et al. 2018; Zhang/ Carpenter/Ko 2013). Indeed, the US National Intelligence Council's latest report *Global Trends 2040* lists AI-powered propaganda as one of its top ten economic and political security concerns (2021).[6] In this case, fake profiles with automatically generated content already populate social networks that are used by billions of people on a daily basis and are therefore prone to misinformation, manipulation and the promotion of hate speech (Keller et al. 2020). There are also other examples of biases embedded in BERT and GPT-3 that are related to the probabilistic ideology of these models. AI Dungeon, a computerized version of Dungeons & Dragons powered by GPT-3, made news in April 2021 for, among other things, allowing the generation of narratives featuring sexual relations involving children – a phenomenon that was obviously not foreseen by OpenAI (Simonite 2021a). In the follow-up to their *Algorithms of Oppression: How Search Engines Reinforce Racism*, Noble and others have also extensively exposed the biases that have always been built into Google environments, both in the various language models that preceded BERT and in the way that BERT is now far from solving these difficulties (Noble 2018; Bhardwaj, Majumder and Poria 2021; Hutchinson et al. 2020). "Stochastic Parrots" is part of this broader critique of NLP:

> The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP […]. However, the

6 In particular, the report explains that "[b]oth states and nonstate actors almost certainly will be able to use these tools to influence populations, including by ratcheting up cognitive manipulation and societal polarization to shape how people receive, interpret, and act on information" (National Intelligence Council 2021: 97).

> training data has been shown to have problematic characteristics […] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status (Bender et al. 2021: 4).

Despite its deleterious consequences for their authors – Gebru in particular – "Stochastic Parrots" is not particularly innovative. As Wired reports, "the paper was not intended to be a bombshell".[7] It merely explores three major issues related to the ever-growing size of language models: first, their environmental cost; second, their formal and rigid nature, which allows biases both to structure themselves and often to go unnoticed; and third, solutions that might mitigate the risks associated with their use.

In the sections that make up the core of the overall argument, the paper reminds us that the models are trained only on the form of the language and not on its substance. To use Saussurean terms, a model can only ever master the signifier of language, never the signified – an argument developed in another paper by Bender and Koller (2020). "Stochastic Parrots" uses this argument to denounce the deceptive or illusory character of the current successes of models such as BERT and GPT-3, which seem to master language when they will only ever have a statistical understanding of it:

> Text generated by an LM [language model] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. […] Contrary to how it may seem when we observe its output, an LM 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: 616–617).

3. Problematizing and understanding the world of hermeneutic machines

Even such a brief history of natural language processing should serve to show how it is very much about meaning and significance. This is essentially what

7 "The authors did not present new experimental results. Instead, they cited previous studies about ethical questions raised by large language models, including about the energy consumed […]. An academic who saw the paper after it was submitted for publication found the document 'middle of the road.'" (Simonite 2021b).

is at stake. First, it is clear that a certain hermeneutic claim of AI cannot be ignored, denied or simply dismissed out of hand. This claim is already disseminated across a wide environment: from researchers like Hinton, by declaring that models "are going to do things like common reasoning", to digital business leaders talking about their platform as a "content understanding engine" (Candela) "focus[sed] on understanding the meaning of what people share" (Zuckerberg). To take another example, the Toronto-based company Cohere, which specializes in the design of NLP models, has set itself the motto and mission of "building machines that understand the world".[8] So while all these claims should be taken seriously, this does not mean that they should be accepted without question. Second, it is clear that there is a need for a better understanding of what we are talking about here, i.e., a better grasp of both the scope and the limits of these hermeneutic machines. The intellectual effort, in other words, is still one of problematization (Romele et al. 2020; Hongladorom 2020).

Looking at the most common criticisms of this particular kind of automated "management" of language and meaning, it is possible to see how they represent variations on the theme of Clever Hans, the so-called intelligent horse from the turn of the twentieth century that appeared to find answers to arithmetic problems on a blackboard; in fact, it was only responding to its master's stimuli and indications. According to Crawford (2021: 151), for example, this represents the embodiment of our desire to anthropomorphize the non-human, as well as a certain spectacle of what intelligence is, without considering a whole set of institutional relationships and political tensions. For others, the image of Clever Hans serves to illustrate the lightness, if not the hermeneutic superficiality, of AI and its language models; as Pavlus's commentary points out, "even a simulacrum of understanding has been good enough for natural language processing" (2019).[9] That said, it is perhaps Gary Marcus in recent years who has done the most to identify the various ways in which what is deemed "deep" in all things deep learning remains only an architectural and technical property – and thus not symbolic and hermeneutic

8 Cohere website, https://cohere.ai/about. Last access: 19 July 2021.

9 This idea of "simulacrum" is understood here not so much in its postmodern and Baudrillardian sense, but more simply as the emergence of handy solutions that are accepted above all for their efficiency. This is what Floridi and Chiriatti refer to when they note how GPT-3 "represents the arrival of a new age in which we can now mass produce good and cheap semantic artifacts" (2020: 690).

(Marcus/Davis 2019, 2020). His argument is threefold. First, this type of model lacks what he calls compositionality – that is, the ability to play with complex and often plotted meanings. On this first point, Marcus is quite close to the idea of the hermeneutic circle – that of Gadamer in particular – in which the whole and the part are so much in dialogue that one can hope to arrive at a form of truth which is more than a simple methodical assemblage (Marcus 2019a; Gadamer 1996 [1960]; see also Andersen 2020). Second, Marcus insists that models like BERT or GPT-3 have "no good way to incorporate background knowledge" (2019). Categories or tools are put forward – for instance those of probability, distance, variation or threshold which have their own logic, horizontal so to speak. Of course, they calculate meanings, but without wanting or being able to draw on their historical, cultural and other richness. And, third, this is what translates into a substantial semantic issue:

> The problem is not with GPT-3 syntax (which is perfectly fluent) but with its semantics: it can produce words in perfect English, but it has only the dimmest sense of what those words mean, and no sense whatsoever about how those words relate to the world (Marcus/Davis 2020).

It is this latter notion of world that seems to be the measure here, even if obviously not without ambiguity itself. Marcus makes use of it, but defines it rather sparingly – which is also the case with everything to do with the form of language in Bender and Gebru, as seen above. How and why do words, meaning and the world appear so inseparable?

This is the type of question that is central to Paul Ricoeur's hermeneutic reflection on textuality, a reflection that can be revisited in the age of natural language processing (Ricoeur 1991a [1986]; Moore 1990; Roberge 2008). "The 'thing' of the text – is the object of hermeneutics, writes the philosopher. Now the thing of the text is the world it unfolds before itself" (1991a: 95). Something is fixed by writing that is not reducible to the intention of its author or to the social conditions of its production – behind or beyond, something by which Ricoeur seeks to guard against a certain romanticism and a certain determinism. As tautological as it may seem, the world of the text is its world, as if to signal its autonomy and objectivity, not only once but twice. On the one hand, textuality in Ricoeur's sense has an internal dynamic and structuring that are reminiscent of the compositionality discussed by Marcus. But on the other hand, and without any contradiction, all texts are always about something, namely that they all have their own reference in a world that they open up and discover? This world is not reality as such, since this would exclude all works of fiction.

No, the world in question is indeed that of the meaning unfolding in it, i.e., a certain universality in the discourse that would represent its "claim to truth" or "Being-demanding-to-be-said" (1991a [1986]: 35 and 19).

What Ricoeur is trying to do is to think of the world of textuality as mediation and suggestion, as what is, so to speak, given to interpretation. The reflection is then resolutely ontological and phenomenological – the author speaks elsewhere of the "immanent transcendence" of textuality (1984). In a text, fundamentally, what is at stake are "sensory […] and axiological values that make the world one that can be inhabited"(1991a [1986]: 11). Ontologically and phenomenologically, this means that it is also always a question of human experience, so that Ricoeur's aim is to combine or bridge different possibilities which are hardly compatible a priori: experience and reflexivity, text and action, explanation and interpretation-comprehension, as well as, more broadly, philosophy and the human and social sciences (Ricoeur 1977; 1991b).

This quick digression through the textual world can only raise the question of its destination: why does it become meaningful and for whom? The whole problem with models like BERT or GPT-3 is that they provide ethereal solutions to this issue, namely that they have infinite difficulty in constructing a meaningful world which, as a result, really means something to someone. Ricoeur saw this horizon of textuality and how, therefore, it forced a reflection on the multiple relationships – complex and ambiguous – between world and appropriation, interpretation of texts and self-understanding (Roberge, 2008). "Reading is like the execution of a musical score, he writes, it marks the realization, the enactment, of the semantic possibilities of the text" (Ricoeur 1991a [1986]: 119). Thus, what a hermeneutic theory like that of Ricoeur suggests is nothing less than the elaboration of a philosophical anthropology (see Ricoeur 1960a and 1960b; 1989). Understanding is as much effort as it is recognition: "to understand oneself is to understand oneself as one confronts the text and to receive from it the conditions of a self other than that which undertakes the reading" (1991a [1986]: 17). It is a question of a diversion through which "I find myself only by losing myself" (1991a [1986]: 88). For Ricoeur, the appropriation in question is more necessary than easy, as if hermeneutic reflection represented a call or a challenge.

However, it is this type of hermeneutic challenge that AI and natural language processing, BERT and GPT-3 in particular, refuse to take up today. One example is the discussion about the "interpretability" and "explicability" of machines that have been the subject of much ink in recent years (Biran/Cotton 2017; Gilpin et al. 2018). For the computational sciences, one of the challenges

is to move away from this (polluted) black box image by showing models in their simplicity and transparency with the avowed aim of increasing confidence in them. Dietterich illustrates this position rather well when he notes, for example, that the aim is "to translate our fuzzy notion of interpretation and understanding into concrete, measurable capabilities" (2019). Interpretability and explicability, in other words, are neologisms of a practical, if not technical and instrumental, nature, which share the logic of automation with the related terms of prediction, optimization, generalization and so on. Some, in fact, have argued that all this is conceptually confusing, to say the least; that there is "conflation" (Miller 2019) between explicability and interpretability or that the latter is "ill-defined" (Lipton 2016). Others have gone on to note that there is a kind of reassignment of the debate's parameters (Mittelstadt/Russell/Wachter 2019), and this in the double sense of translation and impoverishment. In short, the unbearable lightness of the discussion in vogue in the field of AI lies in the fact that it questions nothing or so little, whereas, quite rightly, the challenge of hermeneutics is that of an opening up, a *mise en abyme* and a problematization. For this is what it is all about: as Mittelstadt, Russell and Wachter point out, reflection within the field itself "might benefit from viewing the problem […] more broadly" (2019: 7; see also Campolo/Crawford 2020). What is understanding and interpreting in the age of natural language processing? What kind of world, subject, experience and doubt does this bring into play? Asking these questions encourages us to think of or rethink hermeneutics as part of a search for reflexivity, both individual and collective – that of a subject, but also of a society, a culture, etc.

If words, the world and experience are so inseparable, it is because this world can be said in different ways. It is not by chance that this polysemy is present in Bender and Gebru or in Marcus, and it is not by chance either that it is already present in Ricoeur. Hermeneutics is contextualizing; that is, the world is as much in the text as the text is in the world. The hermeneutical question of language automation is that of a certain pregnancy or anchoring of reflection in what can be said of reality. This can be seen, first of all, at the level of meaning; as Romele notes, "meaningfulness [should be] problematized in [its] context-dependency" (Romele/Severo/Furia 2020: 78). This can then be seen in the historicity of understanding, whereby a subject is always situated in time and space and where such a situation necessarily colors that subject's reading

of what happens.[10] Finally, and to get to the heart of the matter, this can be seen in the very object of what we are dealing with here: AI, natural language processing, BERT and GPT-3. In fact, it was even the first sociological sketches of this vast field of technologies that insisted on showing that it was "socially constituted" (Schwartz 1989; see also Woolgar 1985 in particular). This is not to say that determinism is triumphant – which, as seen above, would not satisfy a hermeneutic perspective such as Ricoeur's – but rather that there is something of a co-construction, a cross-referentiality or a resonance between contextuality and technological advances.[11]

What about this world today? Our world? What characterizes it so well that it makes possible the kind of scandal surrounding the person of Timnit Gebru and the publication of "Stochastic Parrots"? Among other fundamental things, it is clear that we are increasingly living in the midst of not only an increasing "platformization" of the web and digital culture (Helmond 2015), but also an increasing datafication of everyday life (Van Dijck 2014). When, as above, the CEO of Facebook says that his platform "focuses on understanding the meaning of what people share", this is what it is all about. Individually and collectively, it is about our data, our information, and a work that is constantly in progress, which goes as far as the way we (re)construct language, writing, reading, etc. This never fails to be problematic: it is precisely these worlds of meaning that are increasingly under the sway of an appropriation that could be described here as other or heteronomous – as we define further in the next section. At the same time, this should incite hermeneutics – very broadly – to rethink interpretation-comprehension as a sociological and critical issue by attempting, for instance, to reflect on the political economy that does not fail to go with the deployment of language processing models such as BERT or GPT-3 (Roberge 2011; 2020).

10 This is also a large part of the debate between Gadamer and Habermas about the *Vorstruktur des Verstehens*. See Roberge 2011.

11 This is also one of the basic precepts of a vast literature in Science and Technology Studies (STS) ranging as far as Holton and Boyd 2019.

4. Circumscribing the aporias: between critical hermeneutics and Critical AI Studies

The platformization-datafication of the '*hic et nunc*' world is situated in a particular context with equally particular, practical, almost down-to-earth implications and origins. It is its modus operandi that remains to be understood and interpreted. Most concretely, the recent history of AI is about a pragmatic deployment that is therefore more utilitarian than reflexive. It is a question of optimizing solutions as automated forms of action and decision-making. This applies, for example, to autonomous vehicles, cancer diagnosis by algorithmic imaging and much else, including language processing (Stilgoe 2018). What most of these applications and models share is that they are part of an engineered modus operandi, which in turn is part of what Pedro Domingos, a leading figure in the field, calls its "black art" (cited in Campolo/Crawford 2020: 7–8). To train and calibrate a model is to tinker with it; it is to "tweak it to the level of detection that is useful to you" (Amoore 2019: 6). And this is one of the reasons why these solutions are often beta and still imperfectly implemented. Choices are made that nevertheless respond to a certain logic, pressure and urgency. This once again raises perfectly concrete and practical questions: "what is being optimised, and for whom, and who gets to decide[?]" (Crawford 2021: 9). Here we must take another step forward with the author of the *Atlas of AI* when she notes the eminently political nature of all these issues. New power relations are being established prosaically, but certainly. For Crawford, what we are witnessing today is a "shifting tectonics of power in AI" (2021: 11). In a steady fashion, the control of technology gives access to controlled resources. The distribution of power is thus being reorganized more in the sense of greater aggregation than in that of greater equality or symmetry.

Politics and economics are intimately linked, of course, and in the case of AI and natural language processing, this requires a particular adaptation of contemporary capitalism. One of the fundamental reasons why the GAFAM of this world are investing in the development of interpretive machines like BERT and GPT-3 is for the competitive advantage, even dominance, that can be gained. As noted in Simonite's well-known commentary in *Wired*, there is a form of highly performative desiderata here that "makes tech giants harder to topple" (Simonite 2017). It is not that these companies show solidarity among themselves, or conspiratorial tendencies against the rest of the world, but rather that their entire innovation efforts are part of a single "cooperative struggle" (Crandall 2010). To risk an analogy: if each of them occupies a particular po-

sition on the chessboard, they all play the same game of chess that is natural language processing here. As we saw in earlier, Google's BERT is historically entwined in academic research, which ensures that an open-access and open-science model is favored. BERT, in other words, is open, even if only for strategic reasons. In fact, Google's advantage is precisely that it can bring everyone into its environment – familiar, saved in the cloud, allowing for easy transition between different devices, etc. For its part, GPT-3 follows a more direct, if not aggressive, proprietary strategy, like the Microsoft ecosystem of which it is now an integral part. As such, copyright ownership of the content generated by these language models belongs in principle to the company operating the model. It is therefore possible to see the economic challenge in which companies using services such as GPT-3 or BERT could have no more rights to what they generate, or even to the computer code from which their product operates. This issue is still in its infancy, but it is likely to be the major copyright issue of the twenty-first century.

But again, critical thinking cannot reduce everything to economic relationships. When, for example, a renowned researcher in the field such as Yoshua Bengio points out that AI models have become very valuable for GAFAM,[12] he is undoubtedly pointing to broader, if more ambiguous, possibilities. It is these possibilities that, among other important issues, will be crystallized in Gebru's case. We should remember her tweet on leaving: "@jeffdean I realize how much large language models are worth to you now." The problem with the value of natural language processing is that it is both economical and practical, on the one hand, and axial, normative and symbolic, on the other. Hence the reason for a critical hermeneutics around a political economy of meaning and significance as well as the reason for developing Critical AI Studies.

A broader and more distant reflection may point to the whole problem of "assetizing" (Birch/Muniesa 2020) not only data, but also language models and language as such. Optimizing-reducing, enriching-appreciating, common-particular, this "becoming-resource" of language is one of those uncertain couples whose meaning emerges in the gap separating it from appropriation, as described above. This meaning is no longer so much reflexive as extractivist. Following the argumentative line of Birch and Muniesa as well as of Crawford and others, there is a justification and a belief currently being implemented that "everything is data and is there for the taking" (Crawford 2021: 93) –

12 Yoshua Bengio, https://www.technovation.org/blogs/an-interview-with-yoshua-bengio/. Last access: 28 June 2022.

something Shaev et al. summarize perfectly by speaking of "platform meaning extraction" (2019). The general business model works to become a model of the world; that is, it sets up a new normality of which it is both the guarantor and the main beneficiary. The AI myth continues unchallenged, except that, once again, it is the task of a critical hermeneutics to ask questions and show how everything from AI to BERT or GPT-3 is a construction and contingency for which other possibilities are imaginable.

Consider, for example, the ethical turn that the debate around the Gebru affair has sometimes, if not often, taken. In fact, it is common sense to link AI and ethics – as if the hype of the one could not go without the hype of the other and as if, in this cross-discussion, there was not a whole industry, both public and private, of discourse production (Jobin/Ienca/Vayena 2019; Roberge 2020). However, this association is never self-evident and always rather problematic. Authors such as Mittelstadt have shown, for example, how the major principles put forward on the international scene were quite rightly vague and formal, as well as representing "a reason not to pursue new regulation" (2019: 501; see also Wagner 2018). In the same vein, Elish and Boyd have emphasized the normative and political aspects that go hand in hand with such "ability to manufacture legitimacy" within fashionable ethical discourses (2018). And this is what the Gebru test or crisis exposes. When it comes to evaluating or amending itself, Google remains judge and jury. What the company wants to say – or make clear – diverges from what it needs to do. Hao's commentary points out: "As Google underscored in its treatment of Gebru […], the few companies rich enough to train and maintain large language model investments have a heavy financial interest in declining to examine them carefully" (2021: 2).[13] An important part of "Stochastic Parrots" is the discussion of discrimination and bias – gender, race, etc. – that is not just aimed at Google. This is almost worse, as it signals that the problem is more fundamental – a structural one. The article speaks of "real harm" and a simultaneous immediate and insatiable need for accountability, as if this is where the very meaning of criticism becomes eminently practical.

13 An "insider" account of this same idea is found in Lemoine: "Google has moved from being the company whose motto is 'Don't be evil' to being the company whose motto is 'if you don't like it there's the door.' Business interests kept clashing with moral values and time and time again the people speaking truth to power were shown the door" (2021: 4).

The fate of criticism as a result of the Gebru affair is of great interest to a perspective like ours. Critical hermeneutics and Critical AI Studies are in fact intimately linked to these exercises of reflexivity *in situ*, to discourse on discourse and to the development of a political economy of meaning. Its stakes are perfectly summarized by Hanna and Whittaker (2020):

> Gebru's firing suggests this dynamic is at work once again. Powerful companies like Google have the ability to co-opt, minimize, or silence criticisms of their own large-scale AI systems – systems that are at the core of their profit motives […]. The handful of people who are benefiting from AI's proliferation are shaping the academic and public understanding of these systems, while those most likely to be harmed are shut out of knowledge creation and influence.

Perhaps it only remains to be added that what is at stake here is the possibility of a critical culture. What can still be discussed in the automation of language? What can still be discussed about it? These questions should not be closed. Gebru, for her part, is annoyed, but basically she is right: "'Responsible AI' at Google = promote those good at ethics washing & ensuring the marginalization of those already marginalized. I'm telling you after all this they have zero shame."[14]

5. Conclusion

Artificial intelligence is now well and truly conquering language; this is as much a form of zeitgeist as it is of technological development. Models such as BERT and GPT-3 are becoming powerful interpreting machines, to say the least, which, of course, is not without its share of claims. It was the purpose of the first section of this text to take these models seriously. The recent history of natural language processing is partly linked to advances in deep learning and the way in which this type of architecture and networking, based on a principle of statistical regression, now allows parallel processing of a large quantity of data that a model does not need to "understand" in order to calculate efficiently. The strength of Transformers lies in their flexibility: the relationship between letters and numbers, words and codes, or sentences and vectors is

14 Timnit Gebru, tweet, https://twitter.com/timnitGebru/status/1391111917968707585. Last access: 20 July 2021.

thus played and replayed in a continuous flow. The models adapt to platforms and digital culture, which at least partly obliterates some of their weaknesses. As we have seen, these are not exactly hidden, but nonetheless struggle to emerge. When Gebru and company, for example, begin this discussion, it is mostly done through a questioning of the upstream and downstream, i.e., of the biases in the constitution of the databases and in the impacts on populations. The hermeneutic core of the problem remains more or less intact, which is to say nothing of the reception of the researcher's proposal by the industry.

It is, then, a matter of better problematizing in order to better understand. As the second main section has tried to show, the issue of natural language processing is fundamentally semantic. Following the example of Marcus mentioned above, GPT-3 has "no sense whatsoever about how […] worlds related to the word." This is not simple, of course, since this notion of world is sufficiently rich and encompassing to be polysemous. And it is here that a diversion through hermeneutics – that of Ricoeur in particular – is fruitful, insofar as a world can be that of a text as a truth value and a relationship to appropriation, as well as that of a context, namely our world through and for history, culture, etc. An important part of the merit of the Ricoeurian position is its ability to hold these two possibilities together, as if it were not necessary to choose, but to reflect on their innumerable interactions. Reality and interpretation revive each other, as do signification and criticism. In the present discussion, this makes it possible to update hermeneutics to reflect on AI and the way it appropriates something of us through the automation of both data and language. The whole problem is that we need to disentangle a new normality that is inseparably technological, cultural, social, economic and political. Circumscribing the aporias – as the final section of the chapter seeks to do – means showing multiple variations of power, inequalities and their justifications, ethical or otherwise. Most fundamentally, the meaning of AI and natural language processing is to be an extraction of meaning and significance. And that is probably where the choice lies. Gebru has chosen. Her rebuff may be personal and not perfectly calibrated, but it has the great advantage of assuming its political charge by indicating that a critique is always possible, a fortiori when it draws its source from experience and echoes the very idea of society.

Bibliography

Amoore, Louise. 2019. Doubt and the algorithm: on the partial accounts of machine learning. *Theory, Culture & Society* 36(6):147–169.

Andersen, Jack. 2020. Understanding and interpreting algorithms: toward a hermeneutics of algorithms. *Media, Culture & Society* 42(7-8):1479–1494.

Balpe, Jean-Pierre. 1991. Macro-structures et micro-univers dans la génération automatique de textes à orientation littéraire. In *L'imagination informatique de la littérature, Colloque de Cerisy,* Eds. Bernard Magné and Jean-Pierre Balpe, 128–149. Presses Universitaires de Vincennes.

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bhardwaj, Rishabh, Navonil. Majumder and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognitive Computation:* 1–11.

Biran, Or and Courtenay V. Cotton. 2017. Explanation and justification in machine learning: a survey. IJCAI-17 Workshop on Explainable AI (XAI).

Birch, Kean and Fabian Muniesa (Eds). 2020. *Assetization: Turning Things into Assets in Technoscientific Capitalism*. MIT Press.

Brown, Tom B., Benjamin Mann, Nick Ryder et al. 2020. Language models are few-shot learners. arXiv preprint, 1–75.

Buchanan, Bruce G. 2005. A (very) brief history of artificial intelligence. *AI Magazine* 26(4):53–60.

Campolo, Alexander and Kate Crawford. 2020. Enchanted determinism: power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society* 6:1–19.

Cardon, Dominique, Jean-Philippe Cointet and Antoine Mazières. 2018. La revanche des neurones. *Réseaux* 5:173–220.

Crandall, Jordan. 2010. The Geospatialization of Calculative Operations: Tracking, Sensing and Megacities. *Theory, Culture & Society* 27(6):68–90.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven, Conn.: Yale University Press.

Cusin-Berche, Fabienne. 2003. *Les mots et leurs contextes.* Paris : Presses Sorbonne nouvelle.

Devlin, Jacob and Ming-Wei Chang. 2018. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. Last access: 15 June 2021.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 1–16.

Dietterich, Thomas G. 2019. What does it mean for a machine to 'understand'? https://medium.com/@tdietterich/what-does-it-mean-for-a-machine-to-understand-555485f3ad40. Last access: 21 July 2021.

Domingos, Pedro. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Duesenberry, James S. 1949, *Income, Saving, and the Theory of Consumer Behavior.* Cambridge, Mass.: Harvard University Press.

Elish, Madeleine C. and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85(1):57–80.

Floridi, Luciano and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30(4):681–694.

Gadamer, Hans-Georg. 1996 [1960]). *Vérité et méthode. Les grandes lignes d'une herméneutique philosophique.* Paris: Seuil.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Baiwa, Michael Specter and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89.

Hanna, Alex and Meredith Whittaker. 2020. "Timnit Gebru's Exit from Google Exposes a Crisis in AI," Wired. https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/. Last access: 28 June 2022.

Hao, Karen. 2020. We read the paper that forced Timnit Gebru out of Google. Here's what it says. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/. Last access: 28 June 2022.

Hao, Karen. 2021. The race to understand the exhilarating, dangerous world of language AI. https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/. Last access: 28 June 2022.

Heaven, Will D. 2020. OpenAI's new Language Generator GPT-3 is shockingly good, and completely mindless. *MIT Technological Review,* July.

Helmond, Anne. 2015. The platformization of the Web: making Web data platform ready. *Social Media + Society* 1(2):1–11.

Holton Robert and Ross Boyd. 2019. 'Where are the people? What are they doing? Why are they doing it?' (Mindell). Situating artificial intelligence within a socio-technical framework. *Journal of Sociology* 7(2):179–195.

Hongladarom, Soraj. 2020, Machine hermeneutics, postphenomenology, and facial recognition technology. *AI & Society*, 1–8.

Horn, Franziska. 2017. Context encoders as a simple but powerful extension of word2vec. arXiv preprint, 1–5.

Horowitz, Michael C. 2018, Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 2018:1–22.

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton et al. 2020. Social biases in NLP models as barriers for persons with disabilities. arXiv preprint, 1–5.

Janakiram, M.S.V. 2017. How Google Turned Open Source Into a Key Differentiator for Its Cloud Platform. https://www.forbes.com/sites/janakiramms v/2017/07/09/how-google-turned-open-source-into-a-key-differentiator -for-its-cloud-platform/?sh=7a52302e646f .Last access: 15 June 2021.

Jobin, Aanna, Marcello Ienca and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Keller, Franziska B., David Schoch, Sebastian Stier and JungHwan Yang. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication* 37(2):256–280.

Kilkenny, Monique F. and Kerin M. Robinson. 2018. Data quality: 'Garbage in–garbage out'. *Health Information Management Journal* 47(3):103–15.

Kovic, Marko, Adrian Rauchfleisch, Marc Sele et al. 2018. Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences* 18(1): 69–85.

Lemoine, Blake 2021. The History of Ethical AI at Google. https://cajundiscord ian.medium.com/the-history-of-ethical-ai-at-google-d2f997985233. Last access: 21 July 2021.

Levy, Stephen. 2021. Prabhakar Raghavan Isn't CEO of Google – He Just Runs the Place. https://www.wired.com/story/prabhakar-raghavan-isnt-ceo-of -google-he-just-runs-the-place/. Last access: 15 June 2021.

Lipton, Zachary C. 2016. The mythos of model interpretability. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), 1–9.

Marcus, Gary and Ernest Davis. 2019a. If computers are so smart, how come they can't read? https://www.wired.com/story/adaptation-if-computers-a re-so-smart-how-come-they-cant-read/. Last access: 21 July 2021.

Marcus, Gary and Ernest Davis. 2019b. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Marcus, Gary and Ernest Davis. 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. Last access: 21 July 2021.

Mikolov, Tomas, Kai Chen, Greg Corrado et al. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781.

Miller, Tim. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Mittelstadt, Brent, Chris Russell and Sandra Wachter. 2019, Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288.

Mittelstadt, Brent. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1:501–507.

Moore, Henriatta. 1990. Paul Ricoeur: Action, Meaning and Text. In *Reading Material Culture. Structuralism, Hermeneutics and Post-Structuralism*, Ed. Christopher Tilley. Oxford: Basil Blackwell.

National Intelligence Council. 2021. Global Trends 2040: A More Contested World.

Nayak, Pandu. 2019. Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/. Last access: 4 June 2021.

Noble, Safiya U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Pasquinelli, Matteo and Vladan Joler. 2020. The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism. *AI and Society*, 1–18.

Pavlus, John. 2019. Machines beat humans on a reading test. But do they understand? *Quanta Magazine* [online]. https://www.quantamagazine.org/machines-beat-humans-on-a-reading-test-but-do-they-understand-20191017/. Last access: 21 July 2021.

Peters, Matthew E., Mark Neumann, Mohit Iyyer et al. 2018. Deep contextualized word representations. arXiv preprint, 1–15.

Ricœur, Paul. 1960a. L'antinomie humaine et le problème de l'anthropologie philosophique. *Il Pensiero* 5(3) : 283–290.

Ricœur, Paul. 1960b. *L'homme faillible*. Paris : Aubier.

Ricœur, Paul. 1977. Phenomenology and the social sciences. *The Annals of Phenomenological Sociology* 2:145–159.

Ricœur, Paul. 1984. *Temps et récit. La configuration dans le récit de fiction*. Vol. II, Paris : Seuil.

Ricœur, Paul. 1989. L'homme comme sujet de philosophie. *Anzeiger der philosophisch-historischen Klasse der Österreichischen Akademie der Wissenschaften* 126:73–86.

Ricœur, Paul. 1991a. *From Text to Action: Essays in Hermeneutics. II*, Trans. K. Blamey and J. B. Thompson. Evanston, Ill.: Northwestern University Press. [First published as Ricoeur. 1986. *Du texte à l'action. Essais d'herméneutique II*. Paris: Seuil.]

Ricœur, Paul. 1991b. L'herméneutique et les sciences sociales. In *Théorie du droit et science,* Ed. P. Amselek, 15–25. Paris : Presses universitaires de France.

Roberge, Jonathan and Michael Castelle. 2020. Toward an End-to-End Sociology of 21st-Century Machine Learning. In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, Eds. Jonathan. Roberge and Michael Castelle, 1–29. New York: Palgrave Macmillan.

Roberge, Jonathan, Marius Senneville and Kevin Morin. 2020. How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data and Society* 7(1). https://journals.sagepub.com/doi/full/10.1177/20539517209 19968.

Roberge, Jonathan. 2008. *Paul Ricœur, la culture et les sciences humaines.* Collection Sociologie contemporaine. Québec : Presses de l'Université Laval.

Roberge, Jonathan. 2011. What is critical hermeneutics? *Thesis Eleven* 106(1): 5–22.

Romele, Alberto, Marta Severo and Paolo Furia. 2020. Digital hermeneutics: from interpreting with machines to interpretational machines. *AI & Society* 35:73–86.

Ruder, Sebastian. 2018. NLP's ImageNet moment has arrived. https://ruder.io /nlp-imagenet/. Last access: 19 July 2021.

Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, Mass.: Harvard University Press.

Schwartz, H. Andrew and Dirk Hovy. 2019. Predictive biases in natural language processing models: a conceptual framework and overview. arXiv preprint, arXiv:1912.11078.

Schwartz, Ronald D. 1989. Artificial intelligence as a sociological phenomenon. *Canadian Journal of Sociology/Cahiers canadiens de sociologie* 14(2):179–202.

Scott, Kevin. 2020, Microsoft teams up with OpenAI to exclusively license GPT-3 language model. https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/. Last access: 4 June 2021.

Simonite, Tom. 2017. AI and 'Enormous Data' could make tech giants harder to topple. *Wired*. https://www.wired.com/story/ai-and-enormous-data-could-make-tech-giants-harder-to-topple/. Last access: 21 July 2021.

Simonite, Tom. 2021a. It began as an AI-fueled dungeon game. It got much darker. *Wired*. https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/. Last access: 4 June 2021.

Simonite, Tom. 2021b. What really happened when Google ousted Timnit Gebru. *Wired*, https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/. Last access: 21 July 2021.

Stilgoe, Jack. 2018. Machine learning, social learning and the governance of self-driving cars. S*ocial Studies of Science* 48(1): 25–56.

Sutskever, Ilya., Oriol Vinyals and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. arXiv preprint, 1–9.

Thibout, Charles. 2019. La compétition mondiale de l'intelligence artificielle. *Pouvoirs* 3:131–142.

Van Dijck, José. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12 (2): 197–208.

Vaswani, Ashish, Noam Shazeer, Niki Parmar et al. 2017. Attention is all you need. arXiv preprint, 1–5.

Vincent, James. 2021. Google is poisoning its reputation with AI researchers. https://www.theverge.com/2021/4/13/22370158/google-ai-ethics-timnit-gebru-margaret-mitchell-firing-reputation. Last access: 10 July 2021.

Vinyals, Oriol and Quoc Le. 2015. A neural conversational model, arXiv preprint, 1–8.

Wagner, Ben. 2018. Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping? In *Being Profiled: Cogitas Ergo Sum,* Ed. Mireille Hildebrandt, 1–7. Amsterdam University Press.

Wakabayashi, Daisuke. 2020. Google chief apologizes for A.I. researcher's dismissal. https://www.nytimes.com/2020/12/09/technology/timnit-gebru-google-pichai.html. Last access: 19 July 2021.

Woolgar, Steve. 1985. Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology* 19(4):557–572.

Zhang, Jerry, Darrell Carpenter and Myung S. Ko. 2013. Online Astroturfing: A Theoretical Perspective. *Proceedings of the Nineteenth Americas Conference on Information Systems*, 1–7.