Sociologie et sociétés

BERT, GPT-3, Timnit Gebru et nous
L'intelligence artificielle à la conquête du langage
BERT, GPT-3, Timnit Gebru and us
How artificial intelligence subsumes language
BERT, GPT-3, Timnit Gebru y nosotros
La inteligencia artificial a la conquista del lenguaje

Jonathan Roberge et Tom Lebrun

Volume 53, numéro 1-2, printemps–automne 2021

Une sociologie herméneutique ?
Hermeneutic sociology?

URI : https://id.erudit.org/iderudit/1097750ar
DOI : https://doi.org/10.7202/1097750ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0038-030X (imprimé)
1492-1375 (numérique)

Découvrir la revue

Citer cet article

Roberge, J. & Lebrun, T. (2021). BERT, GPT-3, Timnit Gebru et nous : l'intelligence artificielle à la conquête du langage. *Sociologie et sociétés*, *53*(1-2), 235–257. https://doi.org/10.7202/1097750ar

Résumé de l'article

Le déploiement aujourd'hui de modèles sémantiques automatisés tels le BERT de Google ou le GPT-3 d'OpenAI se montre comme un remarquable défi pour l'inscription de l'herméneutique au coeur même du projet des sciences sociales. L'intelligence artificielle est bel et bien à la conquête du langage. Cela implique d'abord qu'il faille prendre au sérieux les possibilités et la puissance de tels modèles, en se penchant sur l'histoire récente des avancées technologiques en apprentissage profond et les *modi operandi* de ces machines interprétantes. Cela implique ensuite de s'attarder au type de compréhension mis en jeu, à savoir principalement comment le calcul de probabilité, de variation et de seuil par exemple vient vectoriser le langage pour le restituer à la manière d'un perroquet. L'article aborde le renvoi par Google de la chercheure Timnit Gebru suivant la parution de « On the Danger of Stochastic Parrots » pour montrer comment la valeur du traitement automatisé du langage tient tant au monde qu'il met de l'avant qu'à sa référence à un contexte précis. Cela, enfin, doit permettre de circonscrire les apories économiques, politiques et éthiques autour de ces modèles, notamment le fait que les plateformes les développant font l'impasse sur la manière dont ils procèdent par extraction et instrumentalisation du sens. À terme, c'est ce lien étroit entre signification et déplacement des centres de pouvoir qui devient l'enjeu central des *Critical AI Studies*.

BERT, GPT-3, Timnit Gebru et nous

L'intelligence artificielle à la conquête du langage

JONATHAN ROBERGE
Institut national de recherche scientifique
jonathan.roberge@inrs.ca

TOM LEBRUN
Université Laval
tom.lebrun@fd.ulaval.ca

L'année 2020 fut marquée par une énième, mais pour le moins substantielle, crise chez Google avec le départ-licenciement de la chercheure Timnit Gebru. D'anecdotique, l'histoire entourant la soumission de l'article « On the Danger of Stochastic Parrots : Can Language Models Be Too big ? » (Bender, Gebru, McMillan-Major *et al.*, 2020) devient scandale au moment où la compagnie demande soit le retrait dudit article, soit l'oblitération des noms des employés de Google y ayant contribué. Pour Jeff Dean, son directeur de l'IA, l'affaire était entendue en ceci que le travail en question « didn't meet our bar for publication » (cité dans Hao, 2020). Or, c'était sans compter sur la rebuffade de Gebru et les très nombreux appuis ralliés à sa cause — en l'occurrence plus de 2 000 employés de la compagnie ont signé une lettre demandant davantage de transparence dans la gestion de ses affaires internes (Wakabayashi, 2020). Qui donc avait sommé le retrait de l'article et pourquoi au juste ? Questions en forme d'ultimatum, mais questions sans trop de réponses, néanmoins. « Timnit wrote that if we didn't meet these demands », écrit encore Dean, « she would leave Google […] we accept and respect her decision » (*ibid.*). La rupture était à toute fin pratique consommée. Sur Twitter, la chercheure exprimait son désarroi et

Sociologie et sociétés, vol. LIII, nº 1-2, printemps-automne 2021, p. 235-257

interpellait son ancien patron : « @jeffdean I realize how much large language models are *worth* to you now »[1].

C'est très précisément à cette notion de *worth* ou de valeur à laquelle nous voulons nous attarder dans le présent article. Le double sens du mot indique qu'il renvoie plus ou moins distinctement à quelque chose d'économique *et* d'axiologique. L'ambiguïté même du mot et de son utilisation dans le gazouilli, pour le dire autrement, est ce qui le rend emblématique des enjeux les plus importants autour du déploiement aujourd'hui des modèles de traitement automatique du langage naturel (TALN, ou *Natural Langage Processing* en anglais, NLP). Si plusieurs ont vu dans le scandale Gebru un problème éthique et de relation de travail, moins nombreux sont ceux qui ont voulu prendre la question en son sens entier et ainsi explorer de quelles manières elle représente un enjeu herméneutique fondamental. Parce que la question mérite d'être posée : l'IA et les dernières avancées en apprentissage profond (*deep learning*) ont-ils permis l'élaboration de modèles (trop) gros, performants et profonds ? Et que peuvent vouloir dire ces derniers termes, hormis d'un point de vue technique ? Est-ce que la signification et la textualité, l'interprétation et la compréhension, ne sortent pas (trop) appauvries de leur traitement automatisé ? Le propos ici est à dire que l'émergence de modèles tels le BERT de Google ou le GPT-3 d'OpenAI se montre aujourd'hui comme un remarquable *défi* pour les disciplines herméneutiques en général, et pour l'inscription de l'herméneutique au cœur même du projet des sciences sociales en particulier. Il ne s'agit pas alors de nier la montée en puissance de modélisations ou de machines interprétantes — ou même leur portée — mais d'en questionner les conditions de possibilité et la signification. Pour le dire d'un trait, l'émergence de ces machines herméneutiques est l'occasion de penser à nouveaux frais ce que peut représenter une herméneutique critique au sein des sciences sociales et comment cette dernière peut venir dialoguer ou servir d'assise au développement des *Critical AI Studies* (CAIS) (Roberge et Castelle, 2021 ; Pasquinelli et Joler, 2020). Concrètement, cela suppose d'abord de prendre au sérieux l'histoire et le *modus operandi* de ces modèles langagiers et comment leurs différents problèmes ont commencé à se cristalliser au travers de l'affaire Gebru. Cela suppose ensuite de mieux comprendre le type de signification mis en jeu, c'est-à-dire surtout le type de monde du texte qu'ils déploient — ou non — et le type d'expérience de lecture que cela induit. Cela suppose, enfin, de circonscrire les apories des modèles le plus souvent décontextualisés et de justement les (re)traduire ou les reverser dans la réalité sociale, politique, économique et culturelle dont ils sont issus, notamment l'arrimage entre capitalisme de plateforme et *desiderata* éthiques aujourd'hui.

Notre analyse s'articule en trois temps, qui correspondent à ces trois moments évoqués à l'instant : i) prendre au sérieux ; ii) comprendre ; et iii) circonscrire les apories des modèles de traitement automatique du langage. Dans une première section, il

1. Gebru, T. [@timnitGebru], (2 décembre 2020), texte de gazouilli [tweet], *Twitter*, <https://twitter.com/timnitGebru/status/1334345550095912961>, consulté le 19 juillet 2021, italiques ajoutés.

s'agira de saisir ces plus récents modèles langagiers comme des constructions sociales et des assemblages sociotechniques (Schartz, 1989 ; Woolgar, 1985). BERT — ou *Bidirectional Encoder Representations from Transformers* — a été introduit par Google en 2018, puis intégré à son moteur de recherche principal. Il collecte l'information sur Wikipédia par exemple et lit de droite à gauche et inversement afin d'identifier plusieurs connexions parallèles et prédire certains termes manquants. Le GPT-3 de OpenAI est plus récent encore ; avec ses 175 milliards de paramètres, il est réputé dépasser de 400 fois en puissance le modèle de Google en « encodant » la textualité et ouvrant de ce fait énormément de possibilités d'écriture — journalistique, informatique, administrative, etc. Ce que ces deux modèles ont ainsi en commun, c'est de ne pas exactement être des boîtes noires, mais plutôt les objets d'un développement historique particulier ; dit développement qui est pour beaucoup celui de ses difficultés et limitations. Dans une deuxième section de l'article, c'est bien la signification d'ensemble de cette automatisation poussée qui aura à être interrogée. Qu'en est-il de la conception épistémologique promue à travers ces architectures de données et ces régressions statistiques ? Qu'en est-il à la fois de la médiation et du destinataire du langage dans ce type de machine connexionniste et cybernétique ? Ces questions incitent à opérer un certain détour par l'herméneutique — celle de Paul Ricœur sera surtout privilégiée ici, entre autres parce que sa notion de « monde » permet de penser une sémantique, une référence et un « être-à-dire » de la textualité qui donne la mesure de comment parfois, sinon souvent, les modèles d'intelligence artificielle apparaissent « shockingly good, and completely mindless » (Heaven, 2021). Ce monde dont parle Ricœur, autrement dit, est ce qui peut permettre de repenser le lien entre sens et réflexivité, cette dernière étant elle-même entendue comme celle du lecteur, mais aussi plus largement comme la réflexivité retrouvée du monde réel, contextualisé en tant que société, culture, économie et politique. Dans sa troisième et dernière section, l'article cherchera de ce fait à développer une compréhension sociologique et critique du déploiement de ces machines interprétantes problématiques, mais parfaitement pratiques et s'immisçant dans la vie quotidienne que sont BERT et GPT-3. La valeur de ces modèles est inséparable d'un marché de l'extraction de la donnée et de la signification dans lequel certains prospèrent davantage que d'autres et pour lequel, comme l'affaire Gebru le montre plutôt bien, l'éthique devient une sorte de justification et même de commodité.

1. UNE TROP BRÈVE HISTOIRE DES VOLONTÉS D'AUTOMATISATION DU LANGAGE

Ces dernières années marquent une évolution majeure dans le champ du traitement automatique du langage naturel. Pour la première fois, des modèles langagiers relevant d'une architecture dite « *transformer* » permettent de générer des textes suffisamment cohérents pour bluffer leurs lecteurs, et ce, sans relever d'une logique déductive, symbolique et préalablement décidée par un programmeur comme ce fut pendant longtemps le cas (Buchanan, 2005 ; Balpe, 1991). Fondé sur le mécanisme de l'apprentissage machine et particulièrement le *deep learning*, ce type de modèle propose d'imiter

certains des mécanismes cognitifs du cerveau, notamment par le biais de « neurones » artificiels — en réalité des mini-programmes qui s'activent ou se désactivent en fonction du résultat de leur calcul. Comme pour le cerveau humain, la force du mécanisme tient à la mise en réseau d'un grand nombre de ces mini-programmes. Longtemps demeurée à la marge du champ de l'intelligence artificielle, cette méthode est soudainement revenue sous le feu des projecteurs lors du concours ImageNet de 2012 remporté par l'équipe de Geoffrey Hinton, celle-ci ayant conjugué grande puissance de calcul, vaste ensemble de données et cette méthode justement qualifiée de connexionniste (Cardon, Cointet et Mazières, 2018 ; Domingos, 2015).

Dans l'évolution récente du TALN, quatre grandes dates peuvent être distinguées qui correspondent à quatre publications majeures. La première a lieu un an après la grande démonstration de Hinton lors de la publication de l'article « Efficient Estimation of Word Representations in Vector Space » (Mikolov, Chen, Corrado *et al.,* 2013). Rédigé par une équipe de Google — Jeff Dean en est, entre autres, cosignataire —, l'article propose un groupe de modèles langagiers intitulé Word2Vec dont l'objectif est de reconstruire le contexte linguistique dans lequel les mots sont utilisés. Word2Vec, comme l'essentiel des technologies reposant sur l'apprentissage machine, s'appuie très largement sur le principe de régression, une méthode d'analyse statistique permettant de situer une variable en fonction de ses corrélations avec d'autres. Grossièrement résumée, il s'agit de situer la variable — « le sens » — d'un mot en fonction des variables — « les sens » — d'autres mots qui l'entourent. Comme son nom l'indique, Word2Vec vise à transformer les mots en vecteurs, autrement dit à modéliser à l'aide d'algorithmes l'information qu'ils contiennent. Word2Vec « vectorise » en pratique les mots par le biais de deux architectures distinctes et complémentaires : l'une, nommée CBOW, va chercher à prédire un mot en fonction de ses cinq mots à droite et cinq mots à gauche. L'autre, intitulée Skip-gram, fait exactement l'inverse, et va chercher à prédire les mots du contexte en fonction d'un mot donné. La logique mise en place est toujours ainsi prédictive : le modèle doit être capable d'attribuer le « bon » vecteur à chaque mot. En dépit de ses succès au début des années 2010, Word2Vec est fortement limité. Le modèle de langage n'attribue en particulier qu'un seul sens par mot et ne vectorise que les mots pris individuellement en sorte que le sens d'une phrase même relativement simple persiste à lui échapper (Horn, 2017 ; Cusin-Berche, 2003) — il s'agira très certainement d'y revenir.

Pour régler ces nombreuses limites, Vinyals et Le (2015) — eux aussi de Google — publient peu après un article intitulé « A Neural Conversational Model ». Celui-ci propose assez simplement d'appliquer une approche séquentielle à Word2Vec visant à modéliser le sens d'un texte en reliant certaines séquences avec d'autres — formant ainsi une forme plus longue ou « réseautique » de cartographie textuelle (Sutskever *et al.*, 2014). Grâce à cette approche, la modélisation peut désormais s'appliquer à de plus larges séquences, notamment des phrases : les phrases précédant et celles suivant la phrase cible sont ainsi également prises en compte et permettent d'assurer un minimum d'appréhension contextuelle au modèle. Malgré ces progrès, les systèmes de type

Word2Vec demeurent toujours fondés sur cette approche où un mot ne peut avoir qu'une signification. C'est cette limite que l'article « Deep Contextualized Word Representations » (Peters, Neumann, Iyyer *et al.*, 2018) veut dépasser. Les auteurs proposent une nouvelle architecture, dite « *Embeddings from Language Models* » ou ELMo, à l'intérieur de laquelle le modèle peut désormais reconnaître la nature dynamique — mouvante, situationnelle — de la signification des mots. En pratique, chaque mot se voit assigner un coefficient ou un « poids » en fonction de son influence dans la phrase. Un mot comme « glace » peut désormais recouvrir différents sens en fonction d'un certain contexte — « je mange une glace » et « je me regarde dans une glace ». Surtout, ELMo permet pour la première fois d'envisager des modélisations qui n'apprennent pas du texte de façon purement ordonnée en proposant désormais une « lecture » de trois façons différentes : du début à la fin, tout d'abord ; puis de façon inversée — de la fin au début — ; puis en combinant les sens vectorisés des deux types d'analyse. Aussi, ELMo marque le réel début des modèles préentraînés permettant aux utilisateurs de ne pas avoir à entraîner « *from scratch* » (à partir de rien) leurs modèles sur d'énormes corpus de données — pratique extrêmement coûteuse, ne serait-ce qu'en temps et en puissance de calcul[2].

Enfin, l'article intitulé « Attention is All You Need » (Vaswani, Shazeer, Parmar, *et al.,* 2017) marque le moment où l'architecture transformeur vient pour ainsi dire sceller le sort du champ. Les modèles séquentiels préalables avaient somme toute du mal à conserver l'information sur la priorisation des termes entre eux : pour reprendre l'exemple précédent, des informations d'une phrase simple — « je mange une glace », « je me regarde dans une glace » — étaient difficiles à conserver sur de plus longues séquences — « puis la glace est tombée ». L'architecture transformeur délaisse cette approche et ses nombreux problèmes en termes de mémoire, de vitesse de calcul, de position des mots, etc., en se proposant d'identifier le contexte qui confère du sens à chaque mot ; ce dernier étant alors l'objet d'un traitement en parallèle. Cela implique l'utilisation à la fois d'un encodeur et d'un décodeur — et, de fait, d'un grand nombre d'entre eux sur de multiples niveaux agissant et rétroagissant de manière cybernétique. L'encodeur transforme l'information en code en donnant une valeur calculée à un mot ; un décodeur fait exactement l'inverse et transforme le code en information en allant « calculer » un mot à partir d'une valeur. Surtout, le caractère réellement novateur de l'architecture transformeur tient au mécanisme d'attention mis en place. L'idée est de calculer un « produit matriciel pondéré », autrement dit un score matriciel qui détermine le niveau d'attention qu'un mot devrait avoir envers d'autres mots — d'aucuns pourraient aussi parler, plus simplement, de dépendance situationnelle. Un

2. Contrairement à l'usage dans le champ de la reconnaissance d'image — où tout un chacun pouvait télécharger sur ImageNet des modèles préentraînés de reconnaissance de visage, par exemple — le champ du traitement automatique du langage naturel apparaissait avant ELMo comme un milieu peu unifié, où chaque groupe de recherche ou entreprise devait nécessairement partir de zéro, avec ses propres données et puissance de calcul disponible. Tirant exemple de la reconnaissance d'image, différents types de modèles préentraînés voient le jour en même temps qu'ELMo, notamment ULMFit ou encore le premier système transformeur d'OpenAI (Ruder, 2021).

encodeur peut ainsi calculer plusieurs « têtes d'attention », lesquelles fonctionnent de manière bidirectionnelle : un poids d'attention est calculé en entrée et produit un vecteur de sortie. Cette bidirectionnalité en profondeur a pour avantage majeur de permettre un traitement de l'information en parallèle des différentes têtes d'attention, et donc des différentes couches d'encodeurs. Il en découle des temps d'entraînement des modèles de langage considérablement réduits par rapport aux approches séquentielles de type Word2Vec.

Ce sont ainsi ces nouveaux mécanismes de l'attention propres aux architectures transformeur qui sont à la source des immenses succès actuels des modèles langagiers, en particulier le BERT de Google et le GPT-3 d'OpenAI comme fers de lance d'une bataille de tous les instants que se livrent les GAFAM dans leur quête pour la maîtrise de l'intelligence artificielle (Thibout, 2019 ; Horowitz, 2018). BERT ou *Bidirectional Encoder Representations from Transformers* constitue un modèle relativement encore « petit » comparé à GPT-3 puisqu'il a été préentraîné sur environ 3,3 milliards de mots et représente 345 millions de paramètres (Devlin, Chang, Lee, *et al.*, 2018). Son objectif principal est de mettre fin aux recherches formalisées à partir de mots clés ; un but qui peut sembler trivial de prime abord, mais qui est au cœur de la mission que s'est attribuée la compagnie « d'organiser l'information à l'échelle mondiale pour la rendre universellement accessible et utile »[3]. Pour y parvenir, Google doit permettre à ses utilisateurs de s'exprimer de la façon la plus naturelle, conviviale et dialogique possible[4]. BERT atteint en partie cet objectif en se focalisant sur la partie encodeur de l'architecture, celle qui transforme l'information, la requête écrite ou parlée — mais aussi des textes à traduire par exemple — en code et en vecteur comme pour en saisir les contours : qui fait quoi et où, etc. dans cette phrase X ou cet extrait Y. BERT, autrement dit, et surtout, « comprend » au sens d'une extraction des éléments pertinents tels que restitués dans des ensembles plus englobants. Le système architectural transformeur fonctionne à ce titre comme une interface entre le langage naturel d'entrée (la requête) et le résultat de sortie : cet autre langage, purement informatique et calculatoire, est ainsi d'une très grande flexibilité. À noter également que BERT est *open source*, ce qui participe d'une stratégie d'entreprise de création de valeur assez spécifique à Google[5] — il s'agira d'y revenir.

3. Google, « Notre approche de la recherche », [en ligne], <https://www.google.com/intl/fr/search/howsearchworks/our-approach/>, consulté le 24 janvier 2023.

4. Prabhakar Raghavan, vice-président de Google, explique ainsi que l'objectif ultime est de répondre *directement* et *intelligemment* aux besoins des utilisateurs : « 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 ?" » (cité dans Levy, 2021).

5. « 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). Voir aussi Devlin et Chang (2018). Sur la stratégie *open source* de Google (opposée à celle de Microsoft, notamment), voir Janakiram (2017).

Quant à lui, GPT-3 ou *Generative Pre-trained Transformer* représente à l'heure d'écrire ces lignes le modèle de langage le plus puissant avec quelque 570 gigaoctets de données et 175 milliards de paramètres à l'entraînement (Brown, Mann, Ryder *et al.*, 2020). GPT-3 vise explicitement à générer du texte. Contrairement à BERT donc, il privilégie la partie décodeur de son architecture, celle qui permet plus précisément de transformer un code en information, c'est-à-dire inférer des mots manquants, compléter des phrases, etc. Bien loin d'être « *open source* », GPT-3 est actuellement commercialisé via son API (*application programming interface*) ; choix s'inscrivant dans une stratégie d'entreprise qui vise à maîtriser l'écosystème économique sur lequel se fonderont bon nombre d'entreprises futures — sur son blogue, OpenAI rapportait en mars 2021 que plus de 300 compagnies tiraient usage de cette API, un nombre toujours croissant. Parmi les applications déjà disponibles, nous pouvons toutefois déjà citer CopyAI, qui permet par exemple de générer des slogans et des descriptions de produits pour les entreprises, ou encore Fable, qui propose de modéliser des personnages tirés de romans pour discuter avec eux (Scott, 2020)[6].

Parce qu'ils sont des assemblages sociotechniques, il va sans dire que ces modèles peinent à être parfaits ou même à la hauteur des discours légitimant leur usage et plus généralement tout ce qui relève de la magie de l'IA (Roberge, Senneville et Morin, 2020 ; Elish et Boyd, 2018). Le fait est que tout ne va pas pour le mieux dans le meilleur des mondes du TALN et qu'à bien y regarder son déploiement tient davantage de cet autre principe ou paradigme du *garbage in, garbage out* — dit principe « gigo » par ailleurs connu des scientifiques œuvrant dans le champ (Kilkeny et Robinson, 2018). En entrée, il faut voir que si la langue est apparemment modélisable par le calcul, l'architecture transformeur ne peut y parvenir qu'à partir d'une ressource qui constitue elle-même une construction sociale : la base de données. Cette dépendance des modèles de langage envers leurs sources d'entraînement est assez largement traitée dans la littérature (Hutchinson, Prabhakaran, Denton *et al.,* 2020 ; Lebrun, 2018). C'est que l'architecture transformeur demeure fondée sur le principe de régression exposé plus tôt, lequel vise à situer une variable (un mot) en fonction de ses corrélations avec d'autres variables (les autres mots de la base de données). Ce simple procédé mathématique construit ainsi une approche du langage fondée sur le principe du *winner takes all* ; autrement dit, le modèle de langage promeut les constructions langagières les plus statistiquement probables en fonction du jeu de données sur lequel il est entraîné. Aussi, le choix des textes sur lesquels ces modèles langagiers sont développés participe d'une certaine représentation du monde, dont la dimension symbolique, sinon idéologique, n'est souvent révélée qu'une fois les modèles mis en application — par les biais racistes, misogynes ou simplement les incohérences qui en découlent. En l'occurrence, il s'agit là d'un des constats les plus assurés dans le texte ayant mené au départ-renvoi de Gebru de chez Google :

6. Pour voir directement les compagnies citées, voir CopyAI, [en ligne], <www.copy.ai/ >, consulté le 18 mars 2021, ainsi que Fable, [en ligne], <https://fable-studio.com/>, consulté le 4 juin 2021.

> 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, Gebru, Mac-Millan *et al.*, 2020 : 4).

Ce qui est ainsi un problème en entrée le devient en sortie avec un nombre de dérives potentielles et avérées des plus élevés. L'une des applications les plus craintes, à ce sujet, est celle usuellement qualifiée d'*astroturfing* par laquelle est générée automatiquement une pléthore de microdiscours comme pour simuler un mouvement de masse accréditant telles organisations, idées, etc. (Kovic, Rauchfleisch, Sele *et al.*, 2018 ; Zhang, Carpenter et Ko, 2013). Le dernier rapport du National Intelligence Council (2021) intitulé « Global Trends 2040 » fait en effet de la propagande propulsée par IA l'une de ses dix préoccupations majeures en termes de sécurité économique et politique[7]. En l'occurrence, de faux profils dont les contenus sont automatiquement générés parsèment déjà les réseaux sociaux qui sont utilisés quotidiennement par des millions, sinon des milliards d'individus et qui de ce fait sont sujets à la désinformation, à la manipulation et à la promotion de discours haineux (Keller *et al.*, 2020). D'autres exemples de biais intégrés à même BERT et GPT-3 existent aussi, qui sont partie liée à l'obédience probabiliste de ces modèles. AI Dungeon, version informatisée de Donjon & Dragon propulsé par GPT-3, a ainsi fait parler de lui en avril 2021 pour avoir notamment permis la génération de récits mettant en scène des relations sexuelles impliquant des enfants — un phénomène qui n'était évidemment pas prévu par OpenAI (Simonite, 2021). Dans les suites de l'ouvrage *Algorithms of Oppression : How Search Engines Reinforces Racism,* Noble et d'autres ont pour leur part très largement exposé les biais qui ont toujours été intégrés aux environnements Google, et ce, que ce soit les différents modèles de langage ayant précédé BERT ou la manière dont ce dernier est aujourd'hui loin de régler ces difficultés (2018 ; Bhardwaj, Majumder et Poria, 2021 ; Hutchinson, Prabhakaran, Denton *et al.*, 2020 pour n'en nommer qu'une infime partie). L'article « Stochastic Parrots » participe de cette même critique :

> The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP […]. However, the 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.*, 2020 : 4).

Malgré ses conséquences délétères pour ses autrices — Gebru en particulier —, le texte « Stochastic Parrots » n'est pas particulièrement novateur. Comme le rapporte *Wired*, « the paper was not intended to be a bombshell »[8]. L'article se contente en effet d'explorer trois grandes problématiques liées à la taille sans cesse grandissante des modèles

7. Le rapport explique notamment « [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 » (NIC, 2021 : 97).

8. « 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

langagiers : leur coût environnemental d'abord ; leur caractère formel et rigidifiant permettant aux biais à la fois de se structurer et de passer souvent inaperçus ensuite ; et les solutions qui pourraient permettre d'atténuer les risques liés à leur utilisation enfin. Par manque d'espace, c'est essentiellement à la deuxième de ces thématiques à laquelle il s'agit de s'attacher ici. Dans les sections qui composent le cœur de sa démonstration, l'article rappelle que les modèles ne sont entraînés que sur la *forme* du langage et non sur son fond (les mots ou les phrases et leurs sens composés par ces ensembles de caractères). Pour reprendre l'argument en termes saussuriens, un modèle ne pourra jamais maîtriser que le signifiant du langage, jamais le signifié — un argument d'ailleurs développé dans un autre article par Bender et Koller (2020). « Stochastic Parrots » s'appuie sur cet argument pour dénoncer le caractère trompeur ou illusoire des succès actuels de modèles comme BERT et GPT-3, qui semblent maîtriser le langage alors qu'ils n'en auront jamais qu'une appréhension statistique :

> Text generated by an LM 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.*, 2020 ; italiques ajoutés).

2. PROBLÉMATISER ET COMPRENDRE LE MONDE DES MACHINES HERMÉNEUTIQUES

Cette histoire même rapide du traitement automatique du langage naturel doit servir à montrer en quel sens elle est très justement à propos du sens et de la signification. L'enjeu est essentiellement là. D'abord, il apparait bien qu'une certaine revendication herméneutique de l'IA ne puisse être ignorée, déniée ou simplement rejetée du revers de la main. Cette revendication est déjà disséminée à travers un vaste environnement de chercheurs, comme Hinton déclarant que les modèles en développement « are going to do things like common reasoning » à des chefs d'entreprise numérique parlant de leur plateforme comme d'un « content understanding engine » (Candala) « focus[ed] on understanding the meaning of what people share » (Zuckerberg). Pour prendre cet autre exemple, la compagnie torontoise Cohere qui se spécialise dans la conception de modèles langagiers s'est donné comme devise et mission de « build machines that understand the world »[9]. Ainsi, si toutes ces revendications doivent être prises au sérieux, cela ne veut pas dire qu'elles doivent être acceptées telles quelles. Ensuite donc, il apparaît bien qu'il faille mieux cerner ce dont il est question, c'est-à-dire mieux saisir ce qui est à la fois la portée et les limites de ces machines herméneutiques. L'effort intellectuel, autrement dit, en demeure encore et toujours un de problématisation (Romele *et al.*, 2020 ; Hongladorom, 2020 ; Introna, 2000 ; Dreyfus, 2007).

---

who saw the paper after it was submitted for publication found the document "middle of the road" » (Simonite, 2021).

9. Cohere, [en ligne], <https://cohere.ai/about>, consulté le 19 juillet 2021.

À regarder les critiques les plus souvent adressées à ce type particulier de « gestion » automatisée du langage et de la signification, il est possible de voir qu'elles représentent des variations sur le thème de Hans le Malin, ce cheval soi-disant « intelligent » du tournant du 20e siècle, qui semblait trouver les réponses à des problèmes arithmétiques sur un tableau, mais qui, de fait, ne faisait que répondre aux stimuli et indications de son maître. Pour Crawford (2021), par exemple, il y a là l'incarnation d'une volonté d'anthropologiser le non-humain de même qu'une certaine mise en spectacle de ce qu'est l'intelligence cachant assez mal tout un jeu de rapports institutionnels et de tensions politiques — il s'agira très certainement d'y revenir. Pour d'autres, l'image de Hans le Malin sert à illustrer la légèreté, sinon la superficialité herméneutique de l'IA et de ses modèles langagiers ; comme le souligne le commentaire de Pavlus(2019) : « even a simulacrum of understanding has been good enough for natural langage processing[10]. Ceci étant, ces dernières années, c'est sans doute Gary Marcus qui a le plus fait pour cerner les différentes manières par lesquelles ce qui est réputé « profond » dans tout ce qui est *deep learning* ne reste qu'une propriété architecturale et technique — et donc non symbolique et herméneutique (Marcus, 2019a ; Marcus et Davis, 2019b ; Marcus et Davis, 2020). Son argumentaire est tripartite. Primo, ce type de modèle est dépourvu de ce qu'il nomme la compositionnalité (*compositionnality*), à savoir cette capacité à jouer avec des significations complexes et le plus souvent intriguées. Sur ce premier point, Marcus est assez près par exemple de cette idée de cercle herméneutique — chez Gadamer notamment — à travers laquelle le tout et la partie dialoguent tant et si bien qu'ils peuvent espérer en arriver à une forme de vérité qui est davantage qu'un simple assemblage méthodique (Marcus, 2019a ; Gadamer 1996 [1960] ; voir aussi Andersen, 2020). Secundo, Marcus insiste pour dire que des modèles comme BERT ou GPT-3 ont « no good way to incorporate background knowledge » (2019a). Des catégories et outils sont mis de l'avant, tels ceux de probabilité, de distance, de variation ou de seuil, qui ont leur logique propre, horizontale pour ainsi dire. Certes, ils calculent des significations, mais sans vouloir ni pouvoir puiser dans leur richesse historique, culturelle, etc. Et c'est ce qui se traduit, *tertio*, en un substantiel enjeu sémantique :

> 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 (2020 : 5).

C'est cette dernière notion de « monde » qui semble donner la mesure ici, même si, à l'évidence, elle n'est pas sans ambiguïté. Marcus en fait usage, mais la définit assez peu — ce qui est par ailleurs le cas pour tout ce qui concerne la forme du langage chez

10. Cette idée de « simulacre » n'est pas tant entendue ici en son sens postmoderne et pour tout dire baudrillarien, mais sans doute plus simplement comme l'émergence de solutions sous la main et acceptées surtout pour leur efficacité. C'est entre autres à cela que Floridi et Chiriatti font référence lorsqu'ils notent comment GPT-3 « represents the arrival of a new age in which we can now mass produce good and cheap semantic artifact » (2020 : 690).

Bender et Gebru, tel que vu ci-haut. Comment et pourquoi ainsi les mots, la signification et le monde apparaissent-ils tellement indissociables ? En l'occurrence, c'est ce type d'interrogation qui est central à la réflexion herméneutique de Paul Ricœur autour de la textualité ; dite réflexion qui peut de ce fait être revisitée à l'âge du traitement automatique du langage naturel (Ricœur, 1986 ; Moore, 1990 ; Roberge, 2011). « La chose du texte, voilà l'objet de l'herméneutique, écrit le philosophe. Or la chose du texte, c'est le monde qu'il déploie *devant* lui » (Ricoeur, 1986 : 126, italiques ajoutés). Quelque chose est fixé par l'écriture qui n'est pas réductible à l'intention de son auteur ou aux conditions sociales de sa production — derrière ou en deçà, Ricœur cherchant par-là à se prémunir contre un certain romantisme et un certain déterminisme. Pour tautologique que cela puisse sembler, le monde du texte est *son* monde, comme pour en signaler l'autonomie et l'objectivité, non pas une, mais deux fois. D'une part, en effet, la textualité au sens de Ricœur relève d'une dynamique et d'une structuration interne qui ne sont pas sans rappeler la compositionnalité discutée par Marcus. Mais d'autre part, et sans contraction aucune, faut-il voir que tous les textes sont toujours à propos de quelque chose, à savoir qu'ils ont tous leur propre référence dans un monde qu'ils ouvrent et découvrent. Ce monde n'est pas la réalité en tant que telle puisque du coup cela exclurait toute œuvre de fiction. Non, le monde dont il est question est bien celui de la signification qui s'y déploie, d'une certaine universalité dans le discours qui en représenterait la « valeur de vérité » ou l'« être-à-dire » (Ricoeur, 1986 : 34). Ce que tente Ricœur, pour le dire encore autrement, est de penser le monde de la textualité comme médiation et suggestion, comme ce qui est pour ainsi dire donner à l'interprétation. La réflexion est alors résolument d'ordre ontologique et phénoménologique — l'auteur parlant ailleurs de « transcendance immanente » de la textualité par exemple (1984). Dans un texte, fondamentalement, se trouvent mises en jeu des « valeurs sensorielles […] et axiologiques qui font du monde un monde *habitable* » (Ricoeur, 1986 : 24 ; italiques dans l'original). Ontologiquement et phénoménologiquement, cela revient à dire qu'il y est toujours aussi question de l'expérience humaine en sorte que, très justement, le propos de Ricœur cherche à allier ou à arc-bouter différentes possibilités qui ne sont que difficilement compatibles *a priori* : expérience et réflexivité, texte et action, explication et interprétation-compréhension comme plus largement philosophie et sciences humaines et sociales (Ricœur, 1977 ; Ricoeur, 1991).

Ce rapide détour par le monde du texte ne peut que soulever la question de sa destination : pourquoi devient-il parlant et pour qui ? Tout le problème des modèles comme BERT ou GPT-3 est qu'ils apportent surtout des solutions éthérées à cet enjeu, à savoir qu'ils ont une infinie difficulté à construire un monde sensé qui, de ce fait, veuille véritablement dire quelque chose pour quelqu'un. Ricœur, pour continuer avec lui, voyait cet horizon de la textualité et comment donc elle forçait une réflexion sur les rapports multiples — si complexes et ambigus — entre monde et appropriation, interprétation des textes et compréhension de soi (Roberge, 2008). « La lecture est comme l'exécution d'une partition musicale, écrit-il, elle marque l'effectuation, la venue à l'acte des possibilités sémantiques du texte » (Ricœur, 1986 : 153). Ainsi, ce à

quoi une théorie herméneutique comme celle de Ricœur convie n'est rien de moins qu'à l'élaboration d'une anthropologie philosophique (voir 1960a et b ; 1989). La compréhension est autant effort que reconnaissance : « se comprendre, c'est se comprendre devant le texte et recevoir de lui les conditions d'un soi autre que le moi qui vient à la lecture » (Ricoeur, 1986 : 31). Il est question d'un détour par lequel « je ne me trouve qu'en me perdant » (Ricoeur, 1986 : 115) en sorte que, certes, il puisse y avoir là un acte ou une expérience en forme de gageure plus ou moins difficile, mais aussi en une forme assumée de conviction. Pour Ricœur, l'appropriation dont il s'agit est davantage nécessaire que facile comme si, justement, la réflexion herméneutique représentait un appel ou un défi.

Or voilà, c'est ce type de défi herméneutique que refusent aujourd'hui de relever l'IA et le traitement automatique du langage naturel, BERT et GPT-3 en tête. Soit l'exemple des discussions autour de l'« interprétabilité » et de l'« explicabilité » des machines ayant fait couler passablement d'encre ces dernières années (Biran et Cotton, 2017 ; Gilpin *et al.,* 2018). Pour les sciences computationnelles, l'enjeu consiste entre autres à se départir de cette image (polluée) de black box, en montrant les modèles dans leur simplicité et leur transparence avec le but avoué d'augmenter la confiance envers ceux-ci. Dietterich illustre cette position plutôt bien lorsqu'il note par exemple que l'objectif est « to translate our fuzzy notion of interpretation and understanding into concrete, *measurable* capabilities » (2019 ; italiques ajoutés). Interprétabilité et explicabilité, autrement dit, sont des néologismes d'ordre pratique, sinon technique et instrumental, qui partagent la logique de l'automatisation avec les termes connexes de prédiction, optimalisation, généralisation et ainsi de suite. Quelques-uns, de fait, ont fait valoir que tout cela était conceptuellement pour le moins confus, qu'il y avait notamment « *conflation* » (Miller, 2019) entre explicabilité et interprétabilité ou que cette dernière était « ill-defined » (Lipton, 2016). D'autres ont poursuivi sur cette lancée en notant qu'il y avait là une sorte de réassignation des paramètres du débat (Mittelstadt *et al*., 2019) et ce, au double sens d'une traduction et d'un appauvrissement. Bref, l'insoutenable légèreté de la discussion en vogue dans le champ de l'IA tient à ce qu'elle ne remet rien ou si peu en cause alors que, très justement, ce défi de l'herméneutique est celui du désenclavement, de la mise en abyme et de la problématisation. Parce que fondamentalement, c'est bien de cela dont il s'agit ; comme le font remarquer Mittelstadt *et al*., la réflexion au sein même du champ « might benefit from viewing the problem […] more broadly » (2019 : 7 ; voir aussi Campolo et Crawford, 2020). Qu'est-ce que comprendre et interpréter à l'ère du traitement automatique du langage naturel ? Quelle sorte de monde, de sujet, d'expérience et de doute cela met-il en jeu ? Poser ces questions incite à (re)penser l'herméneutique comme partie-liée à une recherche de réflexivité — à la fois individuelle et collective, celle d'un sujet, mais aussi d'une société, d'une culture, etc.

Si donc les mots, le monde et l'expérience sont tellement indissociables, c'est bien parce que ce monde peut se dire de différentes façons. Ce n'est pas par hasard que cette polysémie est présente chez Bender *et al.* ou chez Marcus, et ce n'est pas par hasard non plus qu'elle l'est déjà chez Ricœur. L'herméneutique est *contextualisante*, à savoir

que le monde est autant dans le texte que le texte est dans le monde. La question herméneutique de l'automatisation du langage, autrement dit, est celle d'une certaine prégnance ou d'un certain ancrage de la réflexion dans ce qui — à défaut de meilleure expression — peut être dit de la réalité. Cela se voit d'abord sur le plan de la signification ; comme le note le commentaire de Romele, « meaningfulness and truthfulness are directly "encapsulated" into the notion of information rather than being problematized in their context-dependency » (Romele *et al.*, 2020). Cela se voit ensuite dans une historicité de la compréhension voulant qu'un sujet soit toujours néanmoins situé dans le temps et l'espace et que cette situation teinte nécessairement sa lecture de ce qui advient[11]. Cela se voit enfin, et pour aller à l'essentiel, dans l'objet même de ce qui occupe le présent article : l'IA, le traitement automatique du langage naturel, BERT et GPT-3. De fait, dès les premières ébauches sociologiques de ce vaste champ de technologies, on a insisté pour montrer qu'il était « socially constituted » (Schartz, 1989 ; voir aussi Woolgar 1985, notamment). Non pas que le déterminisme soit triomphant — ce qui, comme il a été vu *supra*, ne pourrait satisfaire une perspective herméneutique comme celle de Ricœur —, mais bien plutôt qu'il a quelque chose d'une co-construction, d'une référentialité croisée ou d'une résonance entre contextualité et avancées technologiques[12].

Qu'en est-il ainsi de ce monde aujourd'hui ? De *notre* monde ? Qu'est-ce qui le caractérise tant et si bien que cela puisse rendre possible, par exemple, le genre de scandale autour de la personne de Timnit Gebru et de la publication de « Stochastic Parrots » ? Entre autres choses fondamentales, force est de constater que nous vivons de plus en plus au sein non seulement d'une « plateformisation » accrue du Web et de la culture numérique (Helmond, 2015), mais encore d'une « datafication » de plus en plus poussée de la vie quotidienne (Van Dijck, 2014). Lorsque, comme ci-haut, le PDG de de Facebook dit que sa plateforme « focuses on understanding the meaning of what people share », c'est de cela dont il est question. Individuellement et collectivement, il s'agit de *nos* données, de *nos* informations et d'un travail sans cesse opérant allant jusqu'à la manière dont on (re)construit le langage, l'écriture, la lecture, etc. Et c'est cela qui ne manque jamais d'être problématique, à savoir justement que ce sont ces mondes de signification qui se trouvent de plus en plus sous l'emprise d'une appropriation qu'il serait ici possible de qualifier d'*autre* ou d'*hétéronome* — en l'occurrence, ce sera la tâche de la section à venir que de la définir plus avant. Aussi, cela incite très largement l'herméneutique à repenser l'enjeu de l'interprétation-compréhension comme enjeu de sociologie critique en tentant par exemple de réfléchir à l'économie politique qui ne manque pas d'aller avec le déploiement de modèles de traitement automatisé du langage tels BERT ou GPT-3.

---

11. En l'occurrence, il y a là une grande partie du débat entre Gadamer et Habermas à propos de la *Vorstruktur des Verstehens* ou préstructure de la compréhension (Roberge, 2011).

12. Ce qui est par ailleurs un des préceptes de base d'une vaste littérature en *Science and Technology Studies* (STS), allant par exemple de Bijker (1995) à Holton et Boyd (2019).

3. CIRCONSCRIRE LES APORIES : ENTRE HERMÉNEUTIQUE CRITIQUE ET *CRITICAL AI STUDIES*

La plateformisation-datafication du monde *hic* et *nunc* est ce contexte particulier ayant des implications et des origines tout aussi particulières, pratiques, terre à terre presque. C'est ce *mode opératoire* qu'il reste à comprendre et interpréter — non pas qu'il soit franchement caché, mais qu'il ne soit pas tout à fait thématisé ni mis de l'avant non plus. Concrètement, l'histoire récente de l'IA est surtout celle d'un déploiement pragmatique qui est de ce fait plus utilitariste que réflexif. Il est question d'optimiser des solutions comme formes automatisées d'action et de prise de décision. Cela vaut par exemple dans les domaines des véhicules autonomes, du diagnostic du cancer par imagerie algorithmique et encore bien d'autres, y compris le traitement automatisé du langage (Stilgoe, 2018). Ce qui est alors commun à l'ensemble de ces applications et modèles est de relever d'un *modus operandi* ingénieurial qui, lui-même relié à ce qu'une sommité du champ telle Pedro Domingos nomme son « black art » (cité dans Campolo et Crawford, 2020 : 7-8). Entraîner et calibrer un modèle, c'est le « bidouiller » ; c'est opérer un « tweaking to the level of detection that is *useful to you* » (Amoore, 2019 : 6 ; italiques ajoutés). Et c'est entre autres ce qui explique le caractère pour souvent bêta et encore imparfait de la mise en service de ces solutions. Des choix sont opérés qui répondent néanmoins à une certaine logique, pression et urgence. Ce qui, une fois de plus, soulève des questions parfaitement concrètes et pratiques : « *what* is being optimized, and for *whom*, and *who* gets to decide » ? (Crawford, 2021 : 9 ; italiques dans l'original). En outre, il s'agit ici de suivre d'un pas supplémentaire l'autrice de l'*Atlas of AI* lorsqu'elle note le caractère éminemment politique de tous ces enjeux. De nouveaux rapports de force s'instaurent prosaïquement, mais assurément. Pour Crawford, ce à quoi l'on assiste aujourd'hui est un phénomène de « shifting tectonics of power in AI » (*ibid.*, 11). De près en près, le contrôle de la technologie donne accès à des ressources de contrôle. La distribution du pouvoir se réorganise ainsi davantage au sens d'une agrégation que d'une meilleure égalité ou symétrie.

Politique et économie sont intimement liées bien sûr, et dans le cas de l'IA et du traitement automatique du langage naturel, cela passe par une adaptation particulière du capitalisme contemporain (Srnicek, 2017). Une des raisons fondamentales pour laquelle les GAFAM de ce monde investissent dans le développement de machines interprétatives comme BERT et GPT-3 tient à l'avantage compétitif, sinon la position dominante qui peut en être tirée. Comme le note ce commentaire connu de *Wired* par exemple, il y a là une forme de desiderata hautement performatif qui « make[s] tech giants harder to top » (Simonite, 2017). Non pas que ces compagnies soient sororales — entre elles — ou conspirationnistes — contre le reste du monde —, mais plutôt que l'ensemble de leurs efforts en termes d'innovation relève d'une seule et même « cooperative struggle » (Crandall, 2010). À risquer l'analogie : si chacune occupe une position particulière sur l'échiquier, toutes jouent même partie d'échecs qu'est le traitement du langage naturel dans le cadre de cet article. Tel qu'entrevu dans la première section, le BERT de Google est indissociable d'une certaine histoire de liens avec les milieux aca-

démiques et faisant en sorte que soit privilégié un modèle d'*open access* et d'*open science.* BERT, autrement dit, fait montre d'une ouverture, même si ce n'est que pour des raisons stratégiques. Parce que, de fait, l'avantage de Google est de pouvoir amener tout un chacun dans *son* environnement — familier, sauvegardé dans le Cloud, permettant la transition simple vers différents appareils, etc. De son côté, GPT-3 procède d'une d'une stratégie propriétaire plus directe, sinon agressive, à l'image de l'écosystème Microsoft dont il est maintenant partie intégrante. À ce titre, la titularité du droit d'auteur sur le contenu généré par ces modèles de langage appartient par principe à l'entreprise exploitant le modèle. Le droit d'auteur, s'appliquant tant au texte qu'au code informatique généré, appartient donc à l'entreprise qui le produit, ici OpenAI et potentiellement Microsoft (Lebrun, 2018). Il est dès lors possible de voir poindre l'énorme enjeu d'une économie où les entreprises, utilisant des services comme GPT-3, BERT et consorts, ne posséderaient pas ou plus de droit sur ce qu'ils génèrent, ni même sur le code informatique à partir duquel fonctionne leur produit. Cette problématique est nouvelle, mais il y a fort à parier qu'il s'agit là du principal enjeu du XXI^e^ siècle en matière de droit d'auteur.

Mais, encore une fois, il ne saurait être question pour la réflexion critique de tout réduire à des rapports économiques. Lorsque par exemple un chercheur réputé dans le champ comme Yoshua Bengio souligne que les modèles d'IA sont devenus « very valuable for GAFAM », il indique sans doute malgré lui des possibilités plus vastes, même si plus ambiguës. Ce sont ces possibilités qui, entre autres choses d'importance, vont venir se cristalliser dans l'affaire Gebru. Il faut se remémorer son gazouilli à la sortie : « @jeffdean I realize how much large langage models are *worth* to you now » (Gebru, *op. cit.*). Le problème avec la valeur du traitement automatique du langage naturel est d'être sonnante et trébuchante, d'une part, tout en étant axiale, normative et symbolique, de l'autre. D'où la nécessité d'une herméneutique critique autour d'une économie politique du sens et de la signification et d'où le pourquoi d'un apport aujourd'hui des *Critical AI studies* en devenir. Gebru, elle-même, semble peiner à voir la portée du double sens qu'elle met pourtant en mot comme si, justement, elle hésitait à dire que c'est cette ambiguïté qui est la plus parlante.

Une réflexion plus large et distante peut quant à elle faire état de tout le problème qu'il y a aujourd'hui à « assetizer » (Birch et Muniesa, 2020) non seulement les données, mais encore les modèles langagiers et le langage en tant que tel. Optimiser-réduire, enrichir-appauvrir, commun-particulier, ce devenir-ressource du langage fait partie de ces couples incertains dont le sens émerge, de par l'écart le séparant de l'appropriation, telle que vue *supra*. Ce sens n'est plus tant réflexif qu'extractiviste. Il s'agit ainsi à la fois de suivre la ligne argumentative de Birch *et al.*, comme celle de Crawford et d'autres, pour qui se mettent en œuvre à l'heure actuelle une justification et une croyance « that everything is data and is there to be taken » (Crawford, 2021 : 93) —, ce que résument à leur tour parfaitement Shaev *et al.* en parlant de « plateform' *meaning extraction* » (2019 ; italiques ajoutés). Le modèle d'affaires général travaille à devenir un modèle du monde, à savoir qu'il met en place une nouvelle normalité dont il est

autant le garant que le principal bénéficiaire. Le mythe de l'IA continue sans remise en question, à ceci près, une fois de plus, que c'est la tâche d'une herméneutique critique que de poser des questions et de montrer comment tout de l'IA, jusqu'à BERT ou GPT-3, relève d'une construction et d'une contingence pour lesquelles d'autres possibles sont imaginables.

Soit l'enjeu de la tournure éthique qu'a parfois, sinon souvent, pris le débat autour de l'affaire Gebru. De fait, il est de bon aloi que de vouloir associer IA et éthique comme si la *hype* de l'une ne pouvait aller sans celle de l'autre et comme si, dans cette discussion croisée, il n'y avait pas toute une industrie à la fois publique et privée de production de discours (Jobin *et al*., 2019 ; Roberge, Senneville et Morin, 2020). Or voilà, cette association ne va jamais de soi et est toujours plutôt problématique. Des auteurs comme Mittelstadt ont par exemple montré comment les grands principes mis de l'avant sur la scène internationale étaient très justement vagues et formels en plus de représenter « a reason not to pursue new regulation » (Mittelstadt, 2019 : 501 ; voir aussi Wagner, 2018). De même, Elish et Boyd (2018) ont insisté sur les aspects normatifs et politiques allant de pair avec cette « ability to manufactrure legitimacy » des discours éthiques en vogue. Et c'est cela qu'expose l'*épreuve* ou la *crise* Gebru. Lorsque vient le temps de s'évaluer ou de s'amender, Google demeure juge et partie. Ce que la compagnie veut dire — ou faire comprendre — diverge de ce qu'elle doit faire pour assez simplement continuer à exister. « As Google underscore in its treatment of Gebru […], souligne le commentaire de Hao, the few companies rich enough to train and maintain large language model investments have an heavy financial interest in declining to examine them carefuly » (2021 : 2)[13]. Une partie importante de « Stochastic Parrots » est entre autres à discuter de discrimination et de biais — genrés, raciaux, etc. — sans viser uniquement Google. Ce qui est presque pire, dans la mesure où cela signale que le problème est plus fondamental, structurel, pour ainsi dire. L'article parle de « real harm » et d'un besoin simultanément immédiat et insatiable d'imputabilité, comme si c'était alors le sens même de la critique qui devenait éminemment pratique et qui donnait à penser que toute vérité n'est pas bonne à dire.

Ce sort de la critique au travers de l'affaire Gebru intéresse au plus haut point une perspective comme la nôtre. L'herméneutique critique et les *Critical AI Studies* sont de fait intimement liées à ces exercices de réflexivité *in situ*, au discours sur le discours et au développement d'une économie politique de la signification. Son enjeu est ainsi parfaitement résumé par Hanna et Whittaker :

> 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

13. Un témoignage de « l'intérieur » quant à cette même idée se trouve chez Lemoine : « Google has moved from being the compagny whose motto is 'Don't be evil' to being the compagny 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).

> understanding of these systems, while those most likely to be harmed are shut out of knowledge creation and influence (2020).

Ne s'agit peut-être que d'ajouter que ce qui apparaît se jouer dans cette affaire n'est rien de moins que la possibilité d'une culture critique. Qu'est-ce qui peut encore être discuté *dans* l'automation du langage ? Qu'est-ce qui peut encore être discuté *sur* ou *à propos* de celle-ci ? Ces questions méritent de rester ouvertes. Gebru est pour sa part excédée, mais fondamentalement elle a raison : « "responsible AI" at Google promote those good at ethics washing and ensuring the marginalization of those already marginalized. I'm telling you after all this they have zero shame »[14].

CONCLUSION

En s'interrogeant sur la possibilité d'une sociologie herméneutique, ce numéro thématique de *Sociologies et sociétés* est l'occasion de réfléchir à des enjeux vastes et complexes — le sens, l'interprétation, la textualité, le symbolique, etc. —, mais aussi à comment ceux-ci s'incarnent dans des objets et défis précis. Notre contribution est voulue comme ce « général particulier » ou, pour le dire encore autrement, comme cette « interrogation fondamentale ici et maintenant ». C'est que l'intelligence artificielle est bel et bien aujourd'hui à la conquête du langage ; tenant autant d'une forme de *zeitgeist* que de développement technologique. Des modèles comme BERT et GPT-3 deviennent des machines interprétantes pour le moins performantes ; ce qui, bien sûr, ne va pas sans un nombre important de revendications. En l'occurrence, c'était le sens de la première section de l'article que de les prendre au sérieux. L'histoire récente du traitement automatisé du langage naturel est liée aux avancées en apprentissage profond et comment, entre autres, ce type d'architecture et de réseautique fondé sur un principe de régression statistique permet aujourd'hui un traitement en parallèle d'une grande quantité de données que le modèle n'a pas besoin de « comprendre » pour calculer efficacement. La force des transformeurs tient ainsi pour beaucoup à leur souplesse : le rapport entre les lettres et les chiffres, les mots et les codes, ou encore les phrases et les vecteurs, étant de ce fait joué et rejoué dans un flux continu. Les modèles s'adaptent aux plateformes et à la culture numérique, ce qui au moins en partie permet d'oblitérer certaines de leurs faiblesses. Tel que vu, ces dernières ne sont pas exactement cachées, mais peinent néanmoins à émerger. Lorsque Gebru et compagnie, par exemple, entament cette discussion, cela est surtout fait à travers une remise en question de l'amont et de l'aval, c'est-à-dire des biais dans la constitution des bases de données et de leurs incidences sur les populations. Le cœur — herméneutique — de la problématique, lui, demeure plus ou moins intact ; ce qui est même sans rien dire de la réception du propos de la chercheure de la part de l'industrie, Google en tête.

Il s'agit alors de mieux problématiser pour mieux comprendre. Comme la section deux de l'article a cherché à le montrer, l'enjeu du traitement automatique du langage

14. Gebru, T. [@timnitGebru], (2 décembre 2020), texte de gazouilli [tweet], *Twitter*, <https://twitter.com/timnitGebru/status/1334345550095912961>, consulté le 19 juillet 2021.

naturel est fondamentalement de nature sémantique. GPT-3 a, suivant l'exemple de Marcus évoqué plus avant, « no sense whatesoever about how […] worlds related to the word » (Marcus et Davis, 2019a). Ce qui n'est pas simple, certes, puisque cette notion de « monde » est suffisamment riche et englobante pour être polysémique. Et c'est ici qu'un détour par l'herméneutique — celle de Ricœur, notamment — est porteur dans la mesure où un monde peut être celui d'un texte comme valeur de vérité et rapport à l'appropriation, de même que celui d'un contexte, à savoir notre monde par et pour l'histoire, la culture, etc. Une partie importante du mérite de la position ricœurienne relève de sa capacité à faire tenir ensemble ces deux possibilités comme si justement il ne fallait pas choisir, mais réfléchir à leurs innombrables interactions. Réalité et interprétation se relancent l'une et l'autre, comme le font par ailleurs signification et critique. Dans le cas qui nous occupe, cela permet d'actualiser l'herméneutique pour en faire une réflexion sur l'IA et la manière dont elle s'approprie quelque chose de nous via l'automatisation à la fois des données et du langage. Tout le problème est qu'il faille désintriquer une nouvelle normalité qui soit indissociablement technologique, culturelle, sociale, économique et politique. *Circonscrire les apories* — comme a cherché à le faire la troisième et dernière partie de l'article — est ainsi à montrer des déclinaisons multiples des thèmes du pouvoir, des inégalités et de leurs justifications, éthiques ou autres. Le plus fondamentalement du monde, le sens de l'IA et du traitement automatisé du langage naturel est d'être une extraction du sens et de la signification. Et c'est sans doute là qu'il faut choisir. Gebru, elle, a choisi. Sa rebuffade est peut-être parfois subjective et pas parfaitement calibrée, mais elle a l'insigne avantage d'assumer sa charge politique en indiquant qu'une critique est toujours possible, *a fortiori* lorsqu'elle puise sa source dans l'expérience et qu'elle fait écho à l'idée même d'une société.

RÉSUMÉ

Le déploiement aujourd'hui de modèles sémantiques automatisés tels le BERT de Google ou le GPT-3 d'OpenAI se montre comme un remarquable défi pour l'inscription de l'herméneutique au cœur même du projet des sciences sociales. L'intelligence artificielle est bel et bien à la conquête du langage. Cela implique d'abord qu'il faille prendre au sérieux les possibilités et la puissance de tels modèles, en se penchant sur l'histoire récente des avancées technologiques en apprentissage profond et les *modi operandi* de ces machines interprétantes. Cela implique ensuite de s'attarder au type de compréhension mis en jeu, à savoir principalement comment le calcul de probabilité, de variation et de seuil par exemple vient vectoriser le langage pour le restituer à la manière d'un perroquet. Larticle aborde le renvoi par Google de la chercheure Timnit Gebru suivant la parution de « On the Danger of Stochastic Parrots » pour montrer comment la valeur du traitement automatisé du langage tient tant au monde qu'il met de l'avant qu'à sa référence à un contexte précis. Cela, enfin, doit permettre de circonscrire les apories économiques, politiques et éthiques autour de ces modèles, notamment le fait que les plateformes les développant font l'impasse sur la manière dont ils procèdent par extraction et instrumentalisation du sens. À terme, c'est ce lien étroit entre signification et déplacement des centres de pouvoir qui devient l'enjeu central des *Critical AI Studies*.

Mots clés : *Critical AI Studies*, traitement automatique du langage naturel, GPT-3, BERT, Timnit Gebru, économie politique la signification.

ABSTRACT

BERT, GPT-3, Timnit Gebru and us: How artificial intelligence subsumes language

The deployment of automated semantic models, such as Google's BERT or OpenAI's GPT-3, poses a serious challenge to hermeneutics and its centrality to the social sciences. Artificial intelligence is indeed conquering language. The technological advances associated with deep learning and the *modi operandi* of interpreting machines force us to seriously consider the possibilities and power of such models. More understanding is needed about the type of automated "comprehension" advanced by these models and how their calculations of probability, variation and threshold, for example, can vector language and render it as a form of "parroting." In this paper and to this end, we examine Google's dismissal of the researcher Timnit Gebru following the publication of "On the Danger of Stochastic Parrots," to demonstrate how the value of automated language processing stems as much from the world it puts forward as from its reference to a specific context. Finally, this should allow us to circumscribe the economic, political and ethical aporias around these models, including the fact that the platforms developing them fail to consider the way in which they proceed by extracting and instrumentalizing meaning. In essence, it is the intimate relationship between meaning and the displacement of centres of power that is fundamental to Critical AI Studies and their evolution.

Key words: Critical AI Studies, automatic natural language processing, GPT-3, BERT, Timnit Gebru, political economy of meaning

RESUMEN

BERT, GPT-3, Timnit Gebru y nosotros. La inteligencia artificial a la conquista del lenguaje

El despliegue actual de modelos semánticos automatizados como el BERT de Google o el GPT-3 de OpenAI, está demostrando ser un desafío notable para la inscripción de la hermenéutica en el corazón mismo del proyecto de las ciencias sociales. De hecho, la inteligencia artificial va a la conquista de la lengua. Esto implica, en primer lugar, que sea necesario *tomar en serio* todas las posibilidades y el poder de tales modelos, es decir, lo que es particularmente la historia reciente de los avances tecnológicos en el aprendizaje profundo y el *modus operandi* de estas máquinas interpretadoras –la lectura bidireccional y los "transformadores", en particular. Esto implica *entender* mejor el tipo de comprensión que está en juego, a saber, principalmente, cómo el cálculo de probabilidades, de variación y de umbral, por ejemplo, vienen a vectorizar el lenguaje para restituirlo a la manera de un loro. El artículo toma nota del caso de la referencia de Google, de la investigadora Timnit Gebru, autora del texto "On the Danger of Stochastic Parrots" (El peligro de los loros estocásticos), para mostrar cómo se encuentra el valor del procesamiento automatizado del lenguaje en el mundo que presenta, así como su referencia a un contexto específico. Esto, por último, debe permitir *circunscribir las aporías* económicas, políticas e incluso éticas en torno a estos modelos y, entre otros, el hecho de que las plataformas que las desarrollan ignoran la forma en que proceden mediante la extracción, mercantilización e instrumentalización del significado. En última instancia, es este estrecho vínculo entre el significado y el desplazamiento de los centros de poder lo que deviene el tema central de los *Critical AI Studies* (Estudios críticos de IA) y su futuro desarrollo.

Palabras clave: *Critical AI Studies*, procesamiento automático del lenguaje natural, GPT-3, BERT, Timnit Gebru, economía política del significado.

BIBLIOGRAPHIE

Amoore, L. (2019), « Doubt and the Algorithm : On the Partial Accounts of Machine Learning », *Theory, Culture & Society*, vol. 36, n° 6, p. 147-169.

Andersen, J. (2020), « Understanding and Interpreting Algorithms : toward a Hermeneutics of Algortihms », *Media, Culture & Society*, vol. 42, n° 7-8, p. 1479-1494.

Balpe, J.-P. (1991), « Macro-structures et micro-univers dans la génération automatique de textes à orientation littéraire », *in* Magné, B. et Balpe, J.-P. (dir.), *L'imagination informatique de la littérature*, Paris, Presses universitaires de Vincennes, p. 128-149.

Bender, E. M., T. Gebru, A. 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*, p. 610-623.

Bender, E. M. et A. 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*, p. 5185-5198.

Bhardwaj, R., N. Majumder et S. Poria (2021), « Investigating gender bias in BERT », *Cognitive Computation*, p. 1-11.

Bijker, W. (1995), *Of Bicycles, Bakelites and Bulbs*, Cambridge (MA), MIT Press.

Biran, O. et C. V. Cotton (2017), « Explanation and justification in machine learning : A survey », *IJCAI-17 Workshop on Explainable AI (XAI).*

Birch, K. et F. Muniesa (dir.) (2020), *Assetization: Turning Things into Assets in Technoscientific Capitalism*, Cambridge, MIT Press.

Brown, T. B., B. Mann, N. Ryder *et al.* (2020), « Language models are few-shot learners », *arXiv preprint*, p. 1-75.

Buchanan, B. G. (2005), « A (very) brief history of artificial intelligence », *AI Magazine*, vol. 26, n° 4, p. 53-60.

Campolo, A. et K. Crawford (2020), « Enchanted determinism : Power without responsibility in artificial intelligence », *Engaging Science, Technology, and Society*, vol. 6, p. 1-19.

Cardon, D., J. P.Cointet et A. Mazières (2018), « La revanche des neurones », *Reśeaux*, n° 5, p. 173-220.

Crandall, J. (2010), « The Geospatialization of Calculative Operations : Tracking, Sensing and Megacities », *Theory, Culture & Society*, vol. 27, n° 6, p. 68-90.

Crawford, K. (2021), *The Atlas of AI*, New Haven, Yale University Press.

Cusin-Berche, F. (2003), *Les mots et leurs contextes*, Paris, Presses Sorbonne Nouvelle.

Devlin, J., M.-W. Chang, K. Lee *et al.* (2018), « Bert : Pre-training of deep bidirectional transformers for language understanding », *arXiv preprint,* p. 1-16.

Devlin, J. et M.-W. Chang (2018), « Open Sourcing BERT : State-of-the-Art Pre-training for Natural Language Processing » [en ligne], <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>, consulté le 15 juin 2021.

Dietterich, T.-G. (2019), « What does it mean for a machine to "understand" ? » [en ligne], <https://medium.com/@tdietterich/what-does-it-mean-for-a-machine-to-understand-555485f3ad40>, consulté le 21 juillet 2021.

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

Elish, M. C. et D. Boyd (2018), « Situating methods in the magic of Big Data and AI », *Communication Monographs*, vol. 85, n° 1, p. 57-80.

Floridi, L. et M. Chiriatti (2020), « GPT-3 : Its nature, scope, limits, and consequences », *Minds and Machines*, vol. 30, n° 4, p. 681-694.

Gadamer, H.-G. (1996 [1960]), *Vérité et méthode. Les grandes lignes d'une herméneutique philosophique*, Paris, Seuil.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Baiwa, M. Specter et L. Kagal (2018), « Explaining explanations : An overview of interpretability of machine learning », *2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (DSAA), p. 80-89.

Hanna, A. et M. Whittaker (2020), « Timnit Gebru's Exit From Google Exposes a Crisis in AI », *Wired* [en ligne], <https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/#:~:text=Google's%20appalling%20treatment%20of%20Gebru,IBM%2C%20and%20yes%2C%20Google>, consulté le 27 octobre 2022.

Hao, K. (2020), « We read the paper that forced Timnit Gebru out of Google. Here's what it says », *MIT Technology Review* [en ligne], <www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>, consulté le 19 juillet 2021.

Heaven, W. D. (2020), « OpenAI's new Language Generator GPT-3 is shockingly good — and completely mindless », *MIT Technological Review* [en ligne], <https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/>, consulté le 27 octobre 2022.

Helmond, A. (2015), « The Platformization of the Web : Making Web Data Platform Ready », *Social Media + Society*, vol. 1, n° 2, p. 1-11.

Holton, R. et R. 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*, vol. 7, n° 2, p. 179-195.

Hongladorom, S. (2020), « Machine hermeneutics, postphenomenology, and facial recognition technology », *AI & Society*, p. 1-8.

Horn, F. (2017), « Context encoders as a simple but powerful extension of word2vec », *arXiv preprint*, p. 1-5.

Horowitz, M. C. (2018), « Artificial intelligence, international competition, and the balance of power », *Texas National Security Review*, p. 1-22.

Hutchinson, B., V. Prabhakaran, E. Denton *et al.* (2020), « Social biases in NLP models as barriers for persons with disabilities », *arXiv preprint*, p. 1-5.

Janakiram, M. (2017), « How Google Turned Open Source Into A Key Differentiator For Its Cloud Platform », *Forbes* [en ligne], <www.forbes.com/sites/janakirammsv/2017/07/09/how-google-turned-open-source-into-a-key-differentiator-for-its-cloud-platform/?sh=7a52302e646f>, consulté le 15 juin 2021.

Jobin, A., M. Ienca et E. Vayena (2019), « The global landscape of AI ethics guidelines », *Nature Machine Intelligence*, vol. 1, n° 9, p. 389-399.

Keller, F. B., D. Schoch, S. Stier et J. Yang (2020), « Political Astroturfing on Twitter : How to Coordinate a Disinformation Campaign », *Political Communication*, vol. 37, n° 2, p. 256-280.

Kilkenny, M. F. Et Robinson, K. M. (2018), « Data quality : "Garbage in—garbage out" », *Health Information Management Journal*, vol. 47, n° 3, p. 103-115.

Kovic, M., A. Rauchfleisch, M. Sele *et al.* (2018), « Digital astroturfing in politics : Definition, typology, and countermeasures », *Studies in Communication Sciences*, vol. 18, n° 1, p. 69-85.

Lebrun, T. (2018), « L'apprentissage machine est une appropriation », *Les Cahiers de propriété intellectuelle*, vol. 30, n° 3, 2018, p. 895-924.

Lemoine, B. (2021, 17 mai), « The History of Ethical AI at Google », [en ligne], <https://cajundiscordian.medium.com/the-history-of-ethical-ai-at-google-d2f997985233>, consulté le 21 juillet 2021.

Levy, S. (2021), « Prabhakar Raghavan Isn't CEO of Google—He Just Runs the Place », *Wired* [en ligne], <www.wired.com/story/prabhakar-raghavan-isnt-ceo-of-google-he-just-runs-the-place/>, consulté le 15 juin 2021.

Lipton, Z. C. (2016), « The mythos of model interpretability », *2016 ICML Workshop on Human Interpretability in Machine Learning*, p. 1-9.

Marcus, G. et E. Davis (2020), « GPT-3, Bloviator : OpenAI's Langage Generator has no idea what it's talking about », *MIT Technology Review* [en ligne], <www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>, consulté le 21 juillet 2021.

Marcus, G. et E. Davis (2019a), « If computers are so smart, how come they can't read ? », *Wired* [en ligne], <www.wired.com/story/adaptation-if-computers-are-so-smart-how-come-they-cant-read/>, consulté le 21 juillet 2021.

Marcus, G. et E. Davis (2019b), *Rebooting AI : Building artificial intelligence we can trust*, New York, Vintage.

Mikolov, T., K. Chen, G. Corrado *et al.* (2013), « Efficient estimation of word representations in vector space », *Proceedings of Workshop at ICLR*, p. 1-12.

Miller, T. (2019), « Explanation in artificial intelligence : Insights from the social sciences », *Artificial Intelligence*, vol. 267, p. 1-38.

Mittelstadt, B. (2019), « Principles alone cannot guarantee ethical AI », *Nature Machine Intelligence,* vol. 1, p. 501-507.

Mittelstadt, B., C. Russell et S. Wachter (2019), « Explaining explanations in AI », *Proceedings of the conference on fairness, accountability, and transparency*, p. 279-288.

Moore, H. (1990), « Paul Ricœur : Action, Meaning and Text », *in* C. Tilley (dir.), *Reading Material Culture. Structuralism, Hermeneutics and Post-Structuralism,* Oxford, Basil Blackwell.

National Intelligence Council (2021), *Global Trends 2040 : A More Contested World* [en ligne], < https://www.dni.gov/files/ODNI/documents/assessments/GlobalTrends_2040.pdf>, consulté le 27 octobre 2022.

Nayak, P. (2019, 25 octobre), « Understanding searches better than ever before » [en ligne], <https://blog.google/products/search/search-language-understanding-bert/>, consulté le 4 juin 2021.

Noble, S. U. (2018), *Algorithms of oppression : How search engines reinforce racism*, New York, NYU Press.

Pasquinelli, M. et V. Joler (2020), « The Nooscope manifested : AI as instrument of knowledge extractivism », *AI and Society*, nº 36, p. 1263-1280.

Pavlus, J. (2019), « Machines Beat Humans on a Reading Test. But Do They Understand ? » [en ligne], <www.quantamagazine.org/machines-beat-humans-on-a-reading-test-but-do-they-understand-20191017/>, consulté le 21 juillet 2021.

Peters, M. E., M. Neumann, M. Iyyer *et al.* (2018), « Deep contextualized word representations », *arXiv preprint*, p. 1-15.

Ricœur, P. (1984), *Temps et récit. La configuration dans le récit de fiction*, Tome II, Paris, Seuil.

Ricœur, P. (1986), *Du texte à l'action. Essais d'herméneutique II*, Paris, Seuil.

Ricœur, P. (1989), « L'homme comme sujet de philosophie », *Anzeiger der philosophische-historische Klasse der Österreichischen Akademie der Wissenschaften*, n° 126, p. 73-86.

Ricœur, P. (1960a), « L'antinomie humaine et le problème de l'anthropologie philosophique », *Il Pensiero*, vol. 5, n° 3, p. 283-290.

Ricœur, P. (1960b), *L'homme faillible*, Paris, Aubier.

Ricœur, P. (1977), « Phenomenology and the Social Sciences », *The Annals of Phenomenological Sociology*, nº 2, p. 145-159.

Ricœur, P. (1991), « L'herméneutique et les sciences sociales », *in* P. Amselek (dir.), *Théorie du droit et science*, Paris, Presses universitaires de France, p. 15-25.

Roberge, J. (2008), *Paul Ricœur, la culture et les sciences humaines*, Québec, Presses de l'Université Laval, coll. « Sociologie contemporaine ».

Roberge, J. (2011), « What is Critical Hermeneutics ? », *Thesis Eleven*, vol. 106, nº 1, p. 5-22.

Roberge, J. et M. Castelle (2020), « Toward an End-to-End Sociology of 21st-Century Machine Learning », *in Cultural Life of Machine Learning*, New York, Palgrave McMillan, p. 1-29.

Roberge, J., M. Senneville et M. Morin (2020), « How to Translate Artificial Intelligence ? Myths and Justifications in Public Discourse », *Big Data and Society*, vol. 7, nº 1, [en ligne], <https://journals.sagepub.com/doi/full/10.1177/2053951720919968>, consulté le 27 octobre 2022.

Romele, A., M. Severo, P. Furia (2020), « Digital hermeneutics : from interpreting with machines to interpretational machines », *AI & Society*, vol. 35, p. 73-86.

Ruder, S. (2018), « NLP's ImageNet moment has arrived » [en ligne], <https://ruder.io/nlp-imagenet/>, consulté le 19 juillet 2021.

Schwartz, R. D. (1989), « Artificial intelligence as a sociological phenomenon », *The Canadian Journal of Sociology/Cahiers canadiens de sociologie*, vol. 14, n° 2, p. 179-202.

Schwartz, H. A. et D. Hovy (2019), « Predictive biases in natural language processing models : A conceptual framework and overview », *arXiv preprint arXiv : 1912.11078*.

Scott, K. (2020, 22 septembre), « Microsoft teams up with OpenAI to exclusively license GPT-3 language model », *Official Microsoft Blog* [en ligne], <https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/>, consulté le 4 juin 2021.

Shaev, Y. (2019), « Machine Learning and the Problem of the Hermeneutic Gap », *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, Kahului, HI, USA, p. 41-45, [en ligne], https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8711093, consulté le 30 janvier 2023

Simonite, T. (2021), « What Really Happened When Google Ousted Timnit Gebru », *Wired* [en ligne], <www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>, consulté le 21 juillet 2021.

Simonite, T. (2021), « It Began as an AI-Fueled Dungeon Game. It Got Much Darker », *Wired* [en ligne], <www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/>, consulté le 4 juin 2021.

Simonite, T. (2017), « AI and "Enormous Data" Could Make Tech Giants Harder to Topple », *Wired* [en ligne], <www.wired.com/story/ai-and-enormous-data-could-make-tech-giants-harder-to-topple/>, consulté le 21 juillet 2021.

Srnicek, N. (2017), *Platform Capitalism*, Londres, Polity.

Stilgoe, J. (2018), « Machine learning, social learning and the governance of self-driving cars », *Social Studies of Science*, vol. 48, n° 1, p. 25-56.

Sutskever, I., O. Vinyals et Q. Le (2014), « Sequence to sequence learning with neural networks », *arXiv preprint*, p. 1-9.

Thibout, C. (2019), « La compétition mondiale de l'intelligence artificielle », *Pouvoirs*, n° 3, p. 131-142.

Van Dijck, J. (2014), « Datafication, dataism and dataveillance : Big Data between scientific paradigm and ideology », *Surveillance & Society*, vol. 12, n° 2, p. 197-208.

Vaswani, A., N. Shazeer, N. Parmar *et al.* (2017), « Attention is all you need », *arXiv preprint*, p. 1-5.

Vincent, J. (2021), « Google is poisoning its reputation with AI researchers », *The Verge* [en ligne], <www.theverge.com/2021/4/13/22370158/google-ai-ethics-timnit-gebru-margaret-mitchell-firing-reputation>, consulté le 20 juillet 2021.

Vinyals, O. et Q. Le (2015), « A neural conversational model », *arXiv preprint*, p. 1-8.

Wagner, B. (2018), « Ethics as an Escape from Regulation : From Ethics-Washing to Ethics- Shopping ? », *in* M. Hildebrandt (dir.), *Being Profiling. Cogitas Ergo Sum*, Amsterdam, Amsterdam University Press, p. 1-7.

Wakabayashi, D. (2020), « Google Chief Apologizes for A.I. Researcher's Dismissal », *The New York Times* [en ligne], <www.nytimes.com/2020/12/09/technology/timnit-gebru-google-pichai.html>, consulté le 19 juillet 2021.

Woolgar, S. (1985), « Why not a sociology of machines ? The case of sociology and artificial intelligence », *Sociology*, vol. 19, n° 4, p. 557-572.

Zhang, J., D. Carpenter et M. Ko (2013), « Online astroturfing : A theoretical perspective », *Proceedings of the Nineteenth Americas Conference on Information Systems*, p. 1-7.