Dealer de mots. La langue comme capital de l'intelligence artificielle

Frédéric Kaplan
Dealer de mots. La langue comme capital de l'intelligence artificielle
Arles, Actes Sud, 2026
Collection « La compagnie des langues »
ISBN 978-2-330-224-12-7

Benjamin Caraco

Responsable de la bibliothèque universitaire Rosalind-Franklin,
Service commun de la documentation (SCD), Université de Caen Normandie ;
chercheur associé au Centre d'histoire sociale des mondes contemporains

En quelques années seulement, l'intelligence artificielle (IA) s'est rendue omniprésente dans notre quotidien numérique. Dans *Dealer de mots. La langue comme capital de l'intelligence artificielle*, Frédéric Kaplan, directeur du Laboratoire d'humanités digitales à l'École polytechnique fédérale de Lausanne (EPFL), revient sur les vingt ans qui ont précédé l'avènement de l'IA générative (IAG) grand public. Pour créer ces IAG, ses concepteurs ont mobilisé ce que Frédéric Kaplan appelle, en s'inspirant de Pierre Bourdieu, un « *capital linguistique* » reposant sur une captation massive et gratuite de nos écrits. Ces derniers sont ensuite passés à la moulinette des algorithmes avant de nous être revendus sous forme de services payants. Ce capital peut ainsi être défini comme « *la somme des données linguistiques accumulées par un acteur* » (p. 32), qu'il s'efforce ensuite de valoriser de diverses manières.

Google : pionnier de l'accumulation du capital linguistique

Cette histoire commence avec le début de l'Internet grand public : comment se repérer dans cette nouvelle masse d'informations ? Par rapport à ses concurrents, l'option adoptée par Google est révolutionnaire. Contrairement aux autres moteurs de recherche, Google ne compte pas les mots présents sur les sites web pour les indexer. Il est trop facile de tricher en gonflant la présence de certains mots-clés populaires. Le célèbre moteur de recherche se fonde lui sur le principe de la citation, comme dans la recherche scientifique, pour déterminer l'importance d'une page web. Son algorithme PageRank s'actualise en permanence pour classer les sites en fonction de leur pertinence, c'est-à-dire leur nombre de citations par URL.

Google fait également un choix déterminant dès le départ : il n'indexe pas simplement les sites qu'il moissonne, mais réalise une copie compressée de l'ensemble du Web. Un choix paradoxal de prime abord, car beaucoup plus coûteux en termes d'infrastructures (serveurs, etc.), mais qui constitue la première pierre du capital linguistique construit par Google. Le moteur de recherche s'impose rapidement comme le meilleur et capte donc un grand nombre d'utilisateurs. Pour se financer, il met aux enchères des mots vendus à des annonceurs publicitaires. Au passage, les termes de recherche employés enrichissent son capital linguistique. En parallèle, via l'autocomplétion, Google développe une « *économie de l'expression* » visant à assister les utilisateurs dans leur communication (recherche en ligne, écriture de courriels, reconnaissance vocale, traduction).

Pour accélérer la constitution de son capital linguistique, l'entreprise se tourne vers les bibliothèques dont elle numérise massivement les fonds. En effet, comme le relève Frédéric Kaplan, dans cette optique d'accumulation linguistique, numériser des livres existants est incomparablement plus rapide que la captation régulière des écrits des internautes. Cette grille de lecture permet de mieux comprendre Google Books, qui a pu apparaître comme une entreprise au service du savoir ou comme la traduction de la volonté de Google de rivaliser avec Amazon et les éditeurs. Toutefois, nombreux sont les livres numérisés qui sont sous droit d'auteur et dont seul un extrait est visible en ligne. De fait, Google a surtout poursuivi l'augmentation de son capital linguistique.

Les équipes de recherche de Google travaillent en parallèle sur les meilleurs moyens d'exploiter ce gisement et aboutissent au modèle « Transformer ». Dépassant les approches précédentes qui visent à comprendre les structures des phrases, « Transformer » travaille sur des grandes masses de données

afin de repérer des régularités. Autrement dit, plutôt que d'apprendre les règles de grammaire, il observe et formule en retour des phrases qui lui apparaissent les plus probables. Le modèle apprend aussi à hiérarchiser.

OpenIA et la course des GAFAM aux IAG

Toutefois, en dépit de son avance et de la masse de données disponibles, ce n'est pas Google qui popularise en premier l'IAG, mais OpenIA avec ChatGPT, dont les fondateurs sont Sam Altman, Elon Musk et Ilya Sutskever, un ancien employé de Google spécialisé dans l'IA. Ce dernier « *développe la conviction profonde, presque religieuse, que seules comptent les performances de calcul des réseaux et la taille des données d'entraînement* » (p. 59). Autrement dit, développer une IA ne serait qu'une question de financement, aux conséquences environnementales et énergétiques non négligeables cependant. Les dirigeants d'OpenIA souhaitent ainsi prendre de court leurs concurrents et s'assurer d'un monopole auprès des futurs utilisateurs de ce type de service. OpenIA ne dispose pas de la base d'entraînement de Google et se tourne donc vers des contenus accessibles du Web, dont Wikipédia, ainsi que vers des manuels de cours. C'est du moins la version officielle ; OpenIA a sûrement utilisé des textes sous droit d'auteur.

Depuis le succès de ChatGPT, chaque GAFAM s'est lancé dans la course aux IAG en s'appuyant sur l'originalité de ses données collectées (issues des réseaux sociaux pour Meta ou X, de techniques de regard pour Apple, etc.). Ces différents acteurs hésitent entre modèles ouverts et fermés, entre discours pleins de promesses ou de menaces relativement à l'IA. Au passage, le développement de l'IA questionne les politiques institutionnelles d'ouverture de la science du fait de risque de captation de ces données par des acteurs privés.

Bouleversements socioculturels

Les conséquences socioculturelles de l'arrivée de ChatGPT sont rapides et sources de bouleversements – entre autres – pour l'éducation, la recherche ou l'édition. Il devient alors très facile d'écrire, alors que la maîtrise de cette capacité est structurante et discriminante dans ces domaines. Une crise de confiance émerge en dépit des tentatives d'adaptation, qui reposent parfois sur l'usage de l'IA pour détecter les textes produits par d'autres IA. Ces dernières sont efficaces pour coder ou pour rédiger comptes rendus de réunions, rapports et autres textes fonctionnels, au risque de saper certains fondements du monde du travail en accroissant le sentiment de vacuité de ces tâches, pourtant nécessaires. Frédéric Kaplan revient aussi sur les « hallucinations » de ces IA, l'aspect boîte noire de leurs données d'entraînement et leurs lacunes, alors que d'aucuns prétendent faire de la science avec elles. En histoire, par exemple, les sources privées et administratives en sont le plus souvent absentes.

L'irruption des IAG questionne le droit d'auteur et la responsabilité juridique des textes produits, ouvrant une série d'interrogations vertigineuses. Est-ce l'utilisateur ou le fournisseur d'IA qui endosse la responsabilité du texte produit ? Les IAG avaient-elles le droit de s'entraîner sur des articles de journaux ou des livres ? Les entreprises privées ont-elles intérêt à laisser leurs employés se servir de l'IA en les alimentant de données potentiellement confidentielles ou sensibles ? Est-on encore auteur d'un livre s'il est fabriqué à partir de prompts ? Les IA peuvent-elles être considérées comme des auteurs ?

Pour financer ces IAG et les colossaux investissements qu'elles impliquent (dont les conséquences matérielles se font sentir, de même que l'augmentation globale des taux d'intérêt), la publicité se renouvelle et de nouveaux services d'expressions sont vendus par abonnement, au risque de faire perdre en compétences leurs utilisateurs. Frédéric Kaplan pointe, à juste titre, les risques d'une « *prolétarisation des savoirs* » et d'« *atrophie expressive* » par l'usage excessif de ces « *prothèses* » : « *Une vision dystopique se dessine. Dans le possible régime post-linguistique qui s'annonce, le locuteur est devenu un locataire* » (p. 124).

Développements récents et perspectives

Frédéric Kaplan identifie un troisième régime du capitalisme linguistique où la simulation permise par l'IA offre de nouvelles perspectives en termes de raisonnement. Dans cette optique, les prochaines IA seraient en mesure de s'auto-améliorer et d'innover dans de nombreux domaines scientifiques. Plutôt qu'une superintelligence à la HAL dans *2001 : l'odyssée de l'espace*, Frédéric Kaplan estime que les IA ressemblent plus à des technologies de simulation à la *Matrix*, ce qui ne revient pas à sous-estimer leurs potentialités pour autant. Et leurs conséquences sur nos emplois.

Le quatrième régime a trait au contrôle du discours, voire de la langue elle-même, à travers la sous-représentation de certains idiomes, peu présents dans les jeux d'entraînement des IA, ou l'émergence d'une prose uniforme. Frédéric Kaplan pointe aussi le risque de contenus incorporés, modifiant subrepticement les discours produits, si les IAG sont potentiellement de plus en plus utilisées par tout un chacun pour s'exprimer. Les enjeux culturels et politiques de telles manipulations ne sont pas minces. Une compétition géopolitique entre États est donc lancée pour

des IA souveraines afin de ne pas rester dépendants de technologies étrangères.

Les entreprises d'IA cherchent en parallèle à renouveler leurs ressources primaires, le Web contenant désormais trop de contenus générés par IA. Anthropic aurait ainsi numérisé des ouvrages achetés d'occasion pour alimenter ses modèles. Une autre question concernant les archives, les bibliothèques et les historiens, émerge également : *quid* de la stabilité des documents produits quand les IA reconfigurent sans cesse, à l'attention de leurs utilisateurs, les sources existantes ? Pour contrer ces différentes tendances, Frédéric Kaplan invite à une alliance entre institutions de l'écrit (archives, bibliothèques et musées) pour constituer un *« socle documentaire »*, vaste, représentatif et ouvert pour alimenter des modèles plus vertueux au service du bien commun.

Une approche originale invitant au débat avec d'autres travaux

Très accessible, se lisant comme un roman, malgré la densité des informations brassées, Dealer de mots propose une approche originale qui ne s'intéresse pas aux seules conséquences sociales et culturelles, déjà bien médiatisées, de l'IA mais à sa construction fondée sur l'accumulation d'un capital linguistique. Frédéric Kaplan rappelle ou ouvre un certain nombre de questionnements essentiels.

Il est toutefois regrettable que l'auteur ne dialogue que très peu avec d'autres travaux sur le sujet. Il est ainsi symptomatique qu'il n'y ait ni note de bas de page – ce qui peut se comprendre en termes de fluidité de lecture – ni bibliographie finale, sauf un rappel des principaux articles et livres de l'auteur sur le sujet. Pourtant, d'autres travaux récents sont complémentaires et auraient pu enrichir la réflexion. *Déshumanités numériques*[1] de Dominique Boullier revient ainsi sur les différentes formes d'IA, qui ne se résument pas à l'IAG. Il formule aussi un grand nombre de préconisations en termes de régulation des usages de l'IAG, quand Frédéric Kaplan s'en tient à quelques suggestions finales en termes d'infrastructures. De son côté, *Pris dans la toile*[2] de Sébastien Broca éclaire les rapports ambivalents entre monde du libre – porteur d'un principe d'ouverture – et monde de la Big Tech – qui sait s'appuyer sur ces ressources libres pour les monnayer. Qu'en est-il par ailleurs de la dévalorisation du capital linguistique quand celui-ci est de plus en plus victime de la « merdification » (pour reprendre le terme de Cory Doctorow, écrivain et militant des libertés numériques) des contenus, alimenté par l'IA ? Enfin, comme se questionnent de plus en plus d'observateurs (par exemple le journaliste économique de *Médiapart*, Romaric Godin, dans son dernier livre *Le problème à trois corps du capitalisme*[3], ou l'économiste américain Paul Krugman[4]), la constitution de ce capital linguistique débouchera-t-elle sur une vraie rentabilité ou assistons-nous à la formation d'une nouvelle bulle financière ? ⊙

1 Dominique Boullier, *Déshumanités numériques,* Paris, Armand Colin, 2025 (coll. Brèches).

2 Sébastien Broca, *Pris dans la toile : de l'utopie d'Internet au capitalisme numérique,* Paris, Seuil, 2025 (coll. Liber).

3 Romaric Godin, *Le problème à trois corps du capitalisme : sur la gestion autoritaire du désastre (et les moyens de lui faire face),* Paris, La Découverte, 2026 (coll. Cahiers libres).

4 Les analyses de Paul Krugman sont issues de son blog : https://paulkrugman.substack.com/