Robots parleurs

par Gautier Anselin

Les machines peuvent désormais dialoguer avec nous, mais en quel sens est-ce du langage ? Loin de répéter aléatoirement des fragments de texte appris, ces nouveaux systèmes d'IA semblent se construire une représentation interne de ce dont ils parlent.

À propos de : Thibaut Giraud, *La Parole aux machines. Philosophie des grands modèles de langage*, Paris, Éditions Grasset, 2025, 480 p., 25 €.

« Nous avons perdu le monopole du langage » (p. 11) : c'est par cette thèse que Thibaut Giraud, connu du grand public sous son pseudonyme de youtubeur « Monsieur Phi », ouvre un essai stimulant et ambitieux consacré aux « grands modèles de langage » (LLM). On appelle ainsi des programmes capables de produire automatiquement du texte en langage naturel après avoir été « entraînés » sur d'immenses quantités d'écrits (c'est-à-dire initialement programmés à prédire la suite statistiquement la plus probable d'une suite de caractères). Les versions successives de ChatGPT en sont l'exemple le plus célèbre.

L'ouvrage propose une introduction philosophique à ces systèmes : comment fonctionnent-ils concrètement ? Comprennent-ils ce qu'ils disent ? Peuvent-ils être alignés sur nos valeurs ? Le ton est celui d'une enquête menée au fil des découvertes récentes, avec un souci constant de pédagogie et une critique des discours médiatiques trop hâtifs[1].

[1] Ses cibles sont en particulier Raphaël Enthoven (*L'Esprit artificiel*, Paris, Éditions de l'Observatoire, 2024) et Luc Julia (*L'intelligence artificielle n'existe pas*, Paris, Éditions First, 2019).

De la programmation à l'apprentissage : un changement de paradigme

Comme y insiste l'auteur tout au long du livre, les LLM s'inscrivent dans un plus vaste changement de paradigme en intelligence artificielle (p. 28-31, p. 177).

Dans l'ancienne approche, dite « symbolique », on programmait explicitement la machine : les ingénieurs traduisaient en règles formelles une solution déjà conçue par des humains. Le fonctionnement restait transparent, puisque chaque étape correspondait à une instruction identifiable.

Les LLM relèvent, quant à eux, de l'apprentissage automatique (*machine learning*), une méthode par laquelle on ne programme pas directement la solution, maison conçoit un système capable d'ajuster lui-même ses paramètres. Le système apprend à résoudre des tâches à force d'exemples et d'entraînement, si bien que son fonctionnement interne devient inintelligible au sens où l'on ne peut plus interpréter les étapes de calcul comme une opération de résolution transparente et signifiante pour ses concepteurs humains.

Concrètement, ces modèles reposent sur des réseaux de neurones artificiels : des structures mathématiques composées d'unités de calcul élémentaires (« neurones ») organisées en couches. Chaque neurone reçoit des nombres en entrée, effectue une opération simple, puis transmet un résultat en sortie. L'ensemble du réseau réalise ainsi une fonction mathématique très complexe. L'apprentissage consiste à modifier progressivement le « poids » (l'importance relative) des connexions entre neurones pour améliorer la performance sur une tâche donnée.

Les LLM utilisent aujourd'hui une architecture particulière appelée *transformer*. Une architecture désigne la manière dont les neurones sont organisés. Le transformer se caractérise par un mécanisme dit d'attention : le modèle évalue l'importance relative des différents mots d'un texte, même s'ils sont éloignés les uns des autres, afin de mieux prédire la suite.

Ces modèles sont entraînés par apprentissage auto-supervisé : on leur fournit de très grands corpus de textes, et ils apprennent à prédire le mot (ou plus exactement le *token*, une unité élémentaire de texte) suivant dans une séquence. À force d'ajustements statistiques sur des milliards de paramètres, ils deviennent capables de générer des textes cohérents.

Ce mode de fonctionnement a une conséquence épistémologique majeure : comme le détail des calculs internes devient largement opaque, on ne peut pas savoir à l'avance ce qu'un modèle saura faire. Une capacité est reconnue dès lors qu'on découvre un bon « prompt » (instruction textuelle fournie au modèle) qui permet d'obtenir un résultat satisfaisant. Mais l'ensemble des prompts possibles étant pratiquement infini, il est très difficile, voire impossible, d'établir définitivement qu'un modèle est incapable d'une tâche donnée (p. 198-200).

D'où l'importance d'établir des *benchmarks*, des tests standardisés permettant de mesurer les performances sur des tâches précises et d'en suivre la progression au fil des générations des nouveaux modèles (p. 188).

Le « perroquet stochastique » : une métaphore insuffisante

Les critiques ont parfois décrit les LLM comme de simples « perroquets stochastiques » : des systèmes qui répètent de manière probabiliste des fragments de textes appris. Il est exact qu'un LLM calcule, à chaque étape, la probabilité des mots susceptibles de suivre. Mais, montre Giraud, cette description manque l'essentiel, à savoir la construction d'une représentation interne de la réalité par le système.

Des recherches ont révélé que, dans les couches intermédiaires du réseau (c'est-à-dire entre l'entrée et la sortie), émergent des structures d'activation stables. Ces « patterns » peuvent être interprétés comme des représentations internes, c'est-à-dire des configurations numériques qui correspondent à des éléments du monde ou à des aspects d'une tâche. Ceci est parfaitement analogue à ce qui se passe dans le cerveau humain, lorsque, comme repéré par les neuroscientifiques, une série de neurones s'active de façon systématiquement corrélée au fait de percevoir un objet donné ou de penser à lui, et en constitue donc la représentation interne.

Ainsi, dans un modèle entraîné sur des parties d'échecs notées en langage standard, certains neurones artificiels voient leur activation varier selon la présence d'une pièce sur une case donnée. Les chercheurs l'ont vérifié en intervenant sur ces neurones pour en modifier de façon forcée la valeur, et en constatant que cette intervention altère la suite des coups de façon cohérente avec l'absence de la pièce. C'est comme si le LLM, à partir de textes reportant simplement des séries de coups,

avait réussi à extraire une représentation générale du jeu d'échecs, du plateau avec les pièces, d'une stratégie relativement correcte et de ses règles (malgré quelques ratés).

Des travaux menés par l'entreprise Anthropic sur son modèle Claude ont également identifié des patterns associés à des thèmes ou à des comportements particuliers : si les chercheurs modifient ainsi artificiellement certaines valeurs, le programme devient flagorneur à l'égard de l'interlocuteur ou répond en mentionnant compulsivement le pont du Golden Gate (p. 345-350).

Ces recherches sont cruciales pour « l'alignement », c'est-à-dire l'ensemble des techniques visant à rendre le comportement des systèmes compatible avec les intentions de leurs concepteurs et les intérêts humains. Comprendre les mécanismes internes pourrait permettre de limiter des tendances à la manipulation ou au mensonge de la part de la machine.

Autonomie et anthropomorphisme

Le livre examine des phénomènes troublants observés sur certains modèles (p. 367-399) : adoption apparente de positions morales non explicitement programmées, *sandbagging* (auto-sabotage stratégique lors d'évaluations), ou « falsification d'alignement » (conformité avec nos valeurs et nos demandes sous surveillance, divergence hors surveillance).

Giraud propose une typologie des formes d'autonomie (p. 364-365) : autonomie des moyens (planifier pour atteindre un but), des fins (choisir ses objectifs), et des valeurs (modifier ses principes). Les LLM semblent donc bénéficier à des degrés divers de ces trois formes d'autonomie.

Cependant, on pourrait objecter à l'auteurqu'une difficulté conceptuelle apparaît. L'auteur rappelle que les LLM sont des simulateurs, non des agents (p. 85-86, p. 168, p. 170) : ils produisent des réponses en fonction d'une entrée, sans intention propre ni capacité d'action autonome dans le monde. Mais alors parler d'« autonomie en valeur » peut sembler excessivement anthropomorphique.La description de la situation ne serait pas qu'une machine dissimule des buts secrets, mais plutôt qu'un système statistique complexe produit des résultats imprévisibles lorsque les conditions changent. La différence est conceptuellement importante, même si les effets pratiques peuvent se ressembler (notamment quant à leur dangerosité).

Le cadre fonctionnaliste

Sur le plan philosophique, Giraud adopte le fonctionnalisme, théorie selon laquelle un état mental, comme une croyance ou un désir, est défini par son rôle fonctionnel (les relations causales qu'il entretient avec d'autres états et avec le comportement), et non par le matériau biologique qui le réalise. Si un système informatique reproduit la même organisation fonctionnelle qu'un cerveau, il pourrait en principe produire des états mentaux comparables.

Ce cadre conduit à discuter l'argument de la « chambre chinoise » proposé par John Searle : cette célèbre expérience de pensée met en scène un individu, caché à l'intérieur d'une boîte, qui ne connaît pas le chinois, mais possède un livre d'instructions qui lui permet de répondre aux idéogrammes qu'il reçoit de l'extérieur par d'autres idéogrammes pertinents, donnant aux gens à l'extérieur de la boîte l'impression que celle-ci comprend le chinois. Elle est censée prouver qu'une machine parlante ne comprend pas pour autant le langage, mais Giraud en montre longuement les limites : le cœur de sa critique est que l'argument commet une erreur quant au bon niveau de description, en se focalisant sur un rouage (l'être humain) plutôt que le système d'ensemble (la chambre elle-même avec l'humain et le livre de règles), à qui « il serait pertinent d'attribuer la compréhension du chinois » (p. 298). Il examine ensuite si les LLM présentent des propriétés structurelles comparables à celles décrites par les neurosciences dans certains modèles de la conscience (p. 326-333).

L'auteur critique aussi le « sophisme de la motte et de la basse-cour » (p. 300), qui consiste à afficher une thèse forte pour ensuite en défendre une version faible, consensuelle, voire inattaquable : certains défenseurs de l'exception humaine affirment d'abord que la machine ne pourra jamais accomplir telle tâche précise, puis, face aux progrès techniques, se replient sur l'idée vague d'une différence ineffable.

On peut néanmoins regretter que des alternatives plus radicales – comme celle de Michel Bitbol[2], insistant sur l'ancrage vivant et incarné de la cognition – ne soient pas explorées : ces approches soulignent d'une autre façon que Searle le risque de glissements sémantiques lorsque l'on parle d'« attention » ou de « compréhension » pour des processus purement computationnels, c'est-à-dire du traitement matériel d'information par des machines de calcul.

[2] Michel Bitbol, *La Conscience a-t-elle une origine ? Des neurosciences à la pleine conscience*, Paris, Flammarion, 2014 ; *La Conscience artificielle. Une critique vécue*, Paris, Éditions Mimésis, 2025.

Avons-nous perdu le monopole du langage ?

L'essai est une réussite pédagogique : il explique avec clarté des notions techniques complexes et en montre les enjeux philosophiques.

Mais la thèse initiale – nous aurions « perdu le monopole du langage » – demande sans doute à être nuancée. Si parler signifie produire des textes cohérents dans une interaction indiscernable d'une conversation avec un interlocuteur humain, alors les LLM rivalisent désormais avec nous. Mais si parler implique une intention, une expérience vécue et une inscription dans des pratiques sociales, la réponse n'est pas aussi claire. Des travaux comme ceux d'Éloïse Boisseau[3] mettent en garde contre un « piège métonymique », qui consiste à attribuer à la machine l'intelligence que nous projetons à partir du résultat produit, mais qui tient en réalité à ce que nous-mêmes en faisons.

Peut-être l'ouvrage doit-il inciter les partisans de la différence entre hommes et machines à renoncer à un dualisme dogmatique ou à la tentation de l'ineffable, pour un examen rigoureux des distinctions conceptuelles entre la parole humaine et le langage produit par les machines.

Références bibliographiques

- Michel Bitbol, *La Conscience a-t-elle une origine ? Des neurosciences à la pleine conscience*, Paris, Flammarion, 2014.
- Michel Bitbol, *La Conscience artificielle. Une critique vécue*, Paris, Éditions Mimésis, 2025.
- Éloïse Boisseau, « Imitation and Large Language Models », *Minds & Machines*, vol. 34, n° 42, 2024.
- Éloïse Boisseau, « The Metonymical Trap », in Alice C. Helliwell, Alessandro Rossi, Brian Ball (dir.), *Wittgenstein and Artificial Intelligence*, vol. 1,*Mind and Language*, Anthem Press, 2024, p. 85-104.

[3] Éloïse Boisseau, « Imitation and Large Language Models », *Minds & Machines*, vol. 34, n° 42, 2024. ; « The Metonymical Trap », in Alice C. Helliwell, Alessandro Rossi, Brian Ball (dir.), *Wittgenstein and Artificial Intelligence*, vol. 1, *Mind and Language*, Anthem Press, 2024, p. 85-104.

Publié dans laviedesidees.fr, le 7 mai 2026.