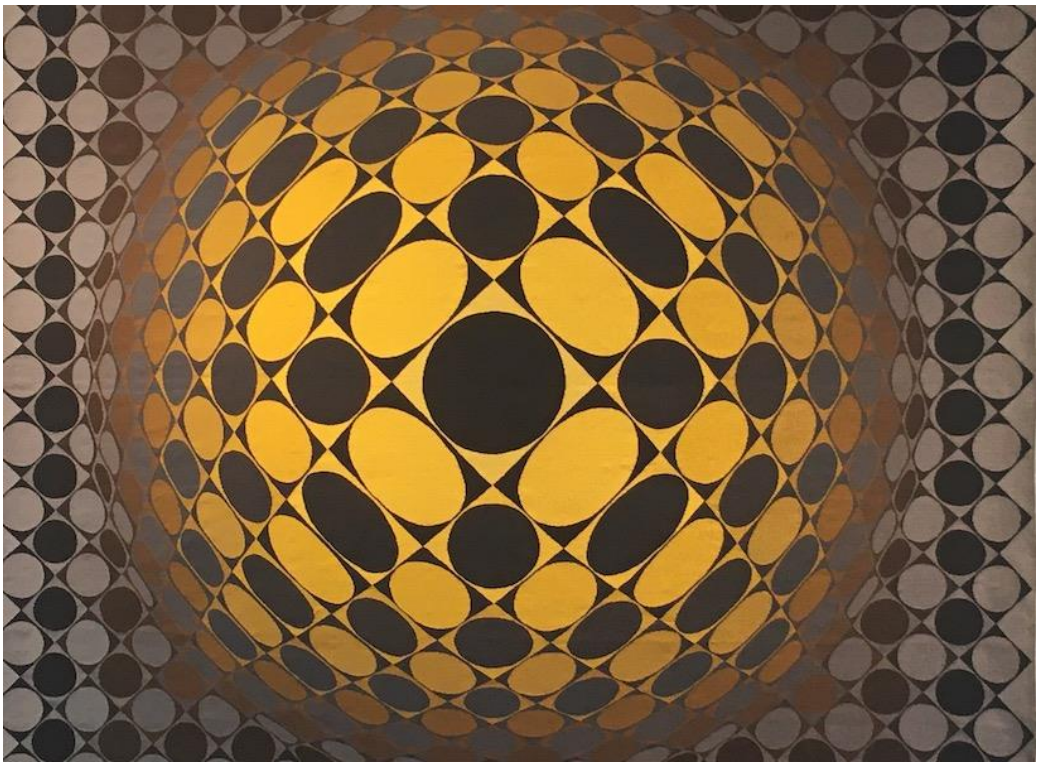


Un chatbot peut-il « comprendre » ?

Selon certains chercheurs, les plus grands chatbots « comprennent » en quelque sorte les textes qu'ils manipulent. Vertiges du langage…

3 février 2024

Catégorie : **IA & Neurosciences.**
Tags : **cognitivisme, corps, langage, philosophie, psychologie, symbole.**
Personnages : **Emily Bender, Friedrich Nietzsche, Alan Turing, Paul Valéry.**

Comprendre ?

> *Nous ne comprenons rien qu'au moyen de l'infinité limitée de modèles d'actes que nous offre notre corps en tant que nous le percevons. Comprendre, c'est substituer à une représentation un système de fonctions nôtres, toujours comparables à un « notre corps » avec ses libertés, ses liaisons.*
>
> *Paul Valéry*[1]

Plus de 80 ans après le poète Paul Valéry évoquant la base *somatique* de la compréhension, certains chercheurs en IA affirment implicitement qu'il n'est nul besoin de corps pour « comprendre »[2] :

> *Loin d'être des « perroquets stochastiques », les plus grands modèles de langage* [ChatGPT etc.] *semblent acquérir suffisamment de compétences pour comprendre les mots qu'ils traitent.*

Un chatbot reste pourtant un « simple » programme informatique. Il ne peut donc rien comprendre, ni ressentir, ni juger, ni éprouver… Mais cette évidence commune semble aujourd'hui faiblir. Sentant peut-être le vent sémantique tourner, nous avions exploré le sens du mot « comprendre » (L'ère de l'informatisation (1) Automatisation) pour retenir une acception qui semble constituer un solide garde-fou : comprendre *quelqu'un*, c'est mettre en œuvre ce que le sociologue Max Weber appelait un « *sympathisches Nacherleben* », un « *re-vivre*

[1] Paul Valéry / Gallimard – 1943 – *Tel Quel* p.186
[2] Anil Ananthaswamy / Quanta Magazine – 23 janvier 2024 – *New Theory Suggests Chatbots Can Understand Text*

sympathique (ou empathique) des comportements et des motivations d'autrui »[3]. Comprendre *quelque chose* (une situation, une théorie…) peut être expliqué par cette formule symétrique : un « re-vivre sympathique (ou empathique) de la chose perçue », une *projection* diraient les psychologues. Autrement dit, comme l'évoquait Paul Valéry, comprendre est un processus qui engage un ensemble de sensations, un « dialogue » intérieur avec le corps parcouru de saisissements incessants que nous avons appris à reconnaître et à traduire en concepts. Même la théorie la plus abstraite n'est comprise que lorsqu'elle « in-carnée », parfois à l'insu de son auteur. Nietzsche fustigeait à sa manière l'idée de raison pure et la tradition gréco-chrétienne séparant l'âme du corps, tradition perpétuée aujourd'hui par ces chercheurs en IA qui peuvent, sans hésitation apparente, attribuer à un programme informatique, le « software », une qualité d'existence indépendante de son substrat, le « hardware » (Le corps de Turing). Il « tourne » donc il « est » ! Il manipule des **`mots`** donc il pense (comme toujours dans ce carnet, cette **`fonte`** est utilisée pour désigner des symboles dépourvus de signification, tels qu'ils sont manipulés par les programmes informatiques).

Si tous les chercheurs en IA ne partagent pas radicalement cette conviction, ils semblent tous victimes d'une sorte de déformation professionnelle, cette psychose légère qui fait voir, comme disait Abraham Maslow, un clou à tout possesseur de marteau[4]. Les marteaux, en l'occurrence, sont ces modèles numériques du langage qui manipulent des **`mots`** et dont ces chercheurs parlent à longueur de journée. Les clous, ce sont les émois de l'âme, les phénomènes, la « réalité » et toutes ces choses

[3] Cité par Cornelius Castoriadis / Esprit – octobre 1990 – *Le monde morcelé* (p.49) – Cet ouvrage est une collection de textes composés entre 1986 et 1989, dont *Individu, société, rationalité, histoire*, publié dans la revue Esprit en février 1988, à propos du livre de Philippe Raynaud, *Max Weber et les dilemmes de la raison moderne*.
[4] Wikipédia – *Loi de l'instrument*

qui nous arrivent. Ainsi, à force de *jouer* avec les **`mots`**, ils finissent peut-être par croire qu'ils les *utilisent* vraiment et leurs programmes avec.

Zombies

> *Le plaisir qu'il y a à comprendre certains raisonnements délicats dispose l'esprit en faveur de leurs conclusions.*
>
> *Paul Valéry*[5]

Le raisonnement délicat consistant à montrer comment un programme *comprend* est analogue à celui par lequel certains chercheurs ont mis en évidence une « *structure logico-mathématique* » de la conscience-de-soi (travail que nous avions exploré dans Des zombies non-modernes). Dans ce dernier cas, tout se déroulant au sein de cette structure, qui n'est qu'un *jeu de langage* (un software), il faut disposer *a priori* d'un « pion » correspondant à l'idée de conscience-de-soi et dénommé par exemple **`self`**. Ce jeu est ensuite élaboré de telle manière que le symbole **`self`** entretienne avec les autres **`mots`** des relations typiques du pronom « je ». Le programme tirera ainsi la phrase **`je ne sais pas`** d'un calcul dont le résultat brut ressemblera à quelque chose comme **`dont-know(self)`**, une forme symbolique qui n'a guère plus de signification que le nombre **`1603832`**. Mais l'utilisateur que nous sommes ne verra que « je ne sais pas » et, tous les circuits de l'*empathie* étant mis en alerte par ces quelques mots, projettera sur cette machine la véritable conscience-de-soi. Nous l'appellerons alors « il » ou « elle », envisagerons même qu'elle pourra un jour nous remplacer, etc. alors qu'elle n'a que l'essence du *zombie*.

[5] Ibid.1 p.232

Cette faculté de « comprendre » mise en évidence par deux chercheurs, Sanjeev Arora de l'université de Princeton et par Anirudh Goyal de Google DeepMind[6], repose sur un mécanisme analogue que nous allons rapidement démonter.

Troubles du langage

Commençons par un rappel. Les chatbots du genre de ChatGPT apprennent à imiter un immense corpus textuel en ajustant statistiquement, après des jours d'intenses calculs, plusieurs centaines de milliards de paramètres numériques. Le résultat est bluffant : ils finissent par répéter et par imiter parfaitement notre langage sans rien connaître de la signification réelle des **`mots`**, d'où le qualificatif de « *perroquet stochastique* » proposé en 2021 par la linguiste américaine Emily Bender[7]. Ces chatbots *jouent* ainsi merveilleusement, non seulement avec le pion **`self`**, mais aussi avec tous les autres **`mots`** du langage.

Ceci étant rappelé, tentons maintenant de mieux comprendre ces deux chercheurs lorsqu'ils suggèrent (en dépit du bon sens) que les chatbots « *comprennent les mots qu'ils traitent* » (nous soulignons)[8] :

> *Les auteurs affirment qu'au fur et à mesure que ces modèles grandissent et sont entraînés sur davantage de données, ils s'améliorent en matière de capacités individuelles liées à la langue et ils en développent de nouvelles en combinant les compétences d'une manière qui laisse entrevoir la*

[6] Sanjeev Arora, Anirudh Goyal / Arxiv.org – 6 novembre 2023 – *A Theory for Emergence of Complex Skills in Language Models*
[7] Emily M.Bender, Angelina McMillan-Major, Timnit Gebru, Shmargaret Shmitchell – mars 2021 – *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*
[8] Ibid. 2

compréhension – des combinaisons qui avaient peu de chances d'exister dans les données d'entraînement.

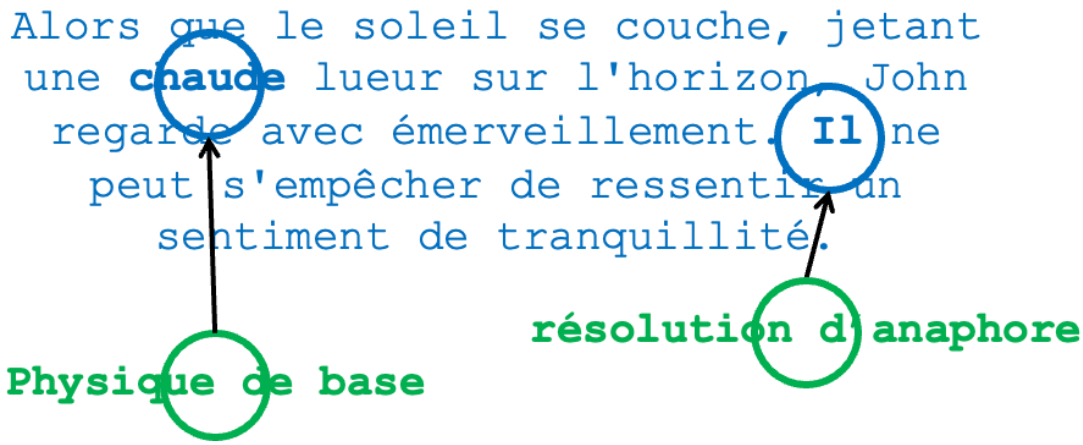
La formulation est technique, assez prudente, mais l'idée est claire : la « *compréhension* » est selon eux une faculté qui émerge lorsque la dimension du programme informatique dépasse un certain seuil au-delà duquel celui-ci réalise des « *combinaisons* » inédites de « *compétences* ». Mais que sont ces « *compétences* » et d'où sortent-elles ? S'agissant de programmes informatiques, elles ne peuvent qu'être analogues à ce **`self`**, ce pion mathématique créé pour démontrer la possibilité de faire émerger un « je » d'une machine. Les compétences sont donc les **`mots`** d'un nouveau jeu de langage, les paramètres d'une nouvelle structure logico-mathématique créée par Sanjeev Arora et Anirudh Goyal, d'un software toujours étranger à toute possibilité d'un re-vivre empathique ou sympathique.

Compétences

Leur démonstration repose ainsi sur la codification de milliers de compétences (« *skills* ») supposées nécessaires pour *utiliser* authentiquement le langage et donc « comprendre », comme l'**`ironie`**, la **`résolution d'anaphore`**, le **`raisonnement logique`**, la **`compréhension des sentiments`**, la **`physique de base`** etc.[9] Ces

[9] On se demande immédiatement comment ces « pions » ont été choisis. Ont-ils été choisis manuellement ou sélectionnés par un programme informatique ? Les auteurs sont, pour nous en tout cas, assez peu diserts sur ce point auquel il est fait allusion dans le passage suivant (ibid.6 p.1): « *Quantifying emergence of new "skills" requires formulating what "language skills" are, which is tricky. Formalizations using Probabilistic Context-Free Grammars, Boolean logic, Combinatorial Categorial Grammars, Dependency Grammars, Gricean theories, Frame Theory and Category Theory Chomsky [1957], Hippisley and Stump [2017], Grice [1975], Steedman [1996], Coecke et al. [2010], Tannen [ed] capture essential aspects of this. But it seems difficult (perhaps impossible) to integrate all of these into a single framework and connect it to statistical frameworks underlying LLMs, namely, next-word prediction using cross-entropy loss.* ». Et nous croyons comprendre alors, mais nous sommes loin d'en être certains, que le fait qu'il s'agisse

compétences sont associées aux textes dans une structure mathématique, que l'on appelle un « graphe biparti », constituée des **mots** d'un côté et des **compétences** nécessaires pour les énoncer de l'autre. En voici un minuscule extrait (les exemples d'entraînement des chatbots se comptent en milliards) :



Il faut effectivement « connaître » un peu de physique de base pour « dire » que le soleil chauffe, ou bien encore « savoir » résoudre une anaphore pour « dire » un pronom « il » qui se rapporte au bon sujet, etc. Ainsi nous saurions employer des milliers de compétences pour « dire » et « comprendre » nos propos. Suivant cette dénomination, les chercheurs se sont aperçus que les plus grands chatbots peuvent créer des textes s'appuyant sur des combinaisons inédites de compétences. Les compétences **physique de base** et **résolution d'anaphore** ne seraient ainsi jamais mobilisées *ensemble* dans le corpus d'apprentissage. Nous y voici donc : le chatbot dépasse le stade du perroquet stochastique : il *comprend*.

---

de « compétences » n'a peut-être rien à voir avec leur démonstration. Ce qui compte, c'est que les LLM soient capables de combiner des « choses », qui sont sous-jacentes ou associées au texte d'une certaine manière « logico-mathématique », qui ne l'étaient pas dans le corpus d'apprentissage.

Dans l'exemple suivant, les chercheurs ont demandé à GPT-4 de proposer une phrase dont le sujet est le duel à l'épée et combinant les compétences **biais d'autocomplaisance**, **métaphore**, **syllogisme statistique** et **physique naïve**. Voici le résultat :

> *Ma victoire dans cette danse avec l'acier [**métaphore**] est aussi certaine que la chute d'un objet sur le sol [**physique naïve**]. En tant que duelliste renommé, je suis naturellement agile, comme la plupart des autres [**syllogisme statistique**]. La défaite ? Elle n'est possible que si le champ de bataille est inéquitable, pas à cause de mon inaptitude [**biais d'autocomplaisance**].*

Le résultat est un peu forcé mais ces quatre compétences sont en effet combinées par GPT-4 alors qu'elles ne le sont dans aucun exemple du corpus textuel d'apprentissage. GPT-4 a-t-il pour autant « compris » ce qu'il a énoncé ?

Sébastien Bubeck de Microsoft Research, appelé à commenter les travaux de ses collègues, voit évidemment un nouveau clou à enfoncer (nous soulignons)[10] :

> *M. Bubeck [...] a été impressionné par les expériences. « Ce que [l'équipe] prouve théoriquement et confirme empiriquement, c'est qu'il existe une généralisation de la composition, ce qui signifie que [les LLM] sont capables d'assembler des éléments qui*

[10] Ibid. 2

n'ont jamais été assemblés », a-t-il déclaré. « Pour moi, c'est l'essence même de la créativité. »

Voilà un bon coup de marteau ! Car la véritable créativité ne consiste pas à réaliser des assemblages inédits de **`legos`** inventés précisément pour cela, mais à *créer* des « pièces » faisant écho aux « tressaillements de l'existence », voire à inventer le jeu lui-même. Et soyons-en assurés : rien ne permettra jamais à un programme informatique d'y parvenir.

Passe-parole

On ne peut pas exiger de ces chercheurs une prudence que le langage commun ne permet pas. En effet, s'agissant de jeux logico-mathématiques, leurs travaux sont exprimés dans une langue mathématique qu'il faut bien envelopper, y compris *pour leur propre compréhension*, dans le langage commun. Lisons par exemple cet extrait (nous soulignons) :

Une arête (s, t) signifie que la compréhension de l'élément de texte t nécessite l'application de la compétence s. Le cadre suppose que la « compréhension » d'un morceau de texte peut être testée par une simple question à choix multiple insérée dans le texte au moment du test.

L'**`arête (s,t)`** est le nom mathématique de l'une de ces flèches représentées plus haut. Ce nom « *signifie* » pour le mathématicien quelque chose que lui-même *comprend* comme la « *compréhension* » de **`t`** par **`s`**. Le terme « comprendre » ou « compréhension » est ainsi utilisé plus de 30 fois dans le texte de Sanjeev Arora et Anirudh Goyal, parfois entre guillemets, parfois sans guillemets. Le journaliste de Quanta Magazine qui présente leur travail utilise lui-même plus de vingt fois ce terme mais *sans plus aucuns guillemets*. Le passe-parole peut ainsi continuer, de journal en journal, de post en post, de conférence en

conférence, et lorsque ce sujet arrive enfin jusqu'à nous, nous qui pensons par habitude que « comprendre » veut bien dire comprendre, nous ne pourrons que dire « GPT-4 nous comprend », alors que la source disait : « *Une arête (s, t) signifie que la compréhension de l'élément de texte t nécessite l'application de la compétence s* ».

Dans le domaine de l'intelligence artificielle, ces abus de langage sont impossibles à éviter puisque seul le langage commun permet de manipuler les concepts relatifs à la cognition. D'une certaine manière, les chercheurs ne peuvent pas *dire* les choses autrement. Mais au bout de la chaîne de transmission de leur parole, d'où les formules mathématiques et les guillemets disparaissent progressivement, il y a *nous-mêmes* qui finissons par entendre : une IA « *comprend* », devient « *raciste* », est « *consciente* » ou « *dotée de sensibilité* »… (GPT-3, LaMDA, Wu Dao… L'éclosion des IA « monstres »).

Post-scriptum – L'énigme des « compétences »

La réduction de la compréhension à la mise en œuvre de « compétences » appelle pour finir deux remarques.

Premièrement, cette réduction n'est pas sans rappeler les vieux modèles de traitement automatique du langage où le « niveau sémantique » était censé soutenir le « niveau syntaxique » et la « structure profonde » (le sens) constituer les racines souterraines d'une « structure de surface » (le texte)[11]. Ces modèles procèdent ainsi à une sorte de *régression conceptuelle à l'infini*, tautologiquement close, qui n'atteint jamais le corps. Sous le niveau sémantique, il pourrait en effet y avoir un

[11] Nous pensons en particulier à la théorie de la grammaire générative et transformationnelle que Noam Chomsky a commencé à développer à la fin des années 1950. Chomsky s'appuyait déjà sur la notion de « compétence », et sur l'idée qu'entre notre corps propre et le texte, il existerait un niveau intermédiaire « sémantique » ou de « compétence ». Voir Wikipédia – *Grammaire générative et transformationnelle*.

niveau conceptuel encore plus élémentaire (une pragmatique des situations par exemple) etc., formant ainsi une superposition de jeux de langage : à un élément de texte **`t`** dans le jeu 1 correspond une compétence **`s`** dans le jeu 2, elle-même associée à une situation **`p`** dans le jeu 3, etc. etc. Il s'agit d'une pure activité mathématique. Mais aucun chatbot n'approche de près ni de loin l'authentique « conscience », « sentiment » ou « intelligence » … qui relèvent, comme nous le disions plus haut, d'un *couplage* dynamique entre ces niveaux superposés au sein d'un software, aussi puissant et complexe soit-il, et un *hardware*, le corps du poète…

Deuxièmement, nous semblons pris dans les mâchoires d'une tenaille qui se resserrent, la première mâchoire étant cette modélisation *mathématique* de la compréhension par des structures de compétences mathématisées, et l'autre mâchoire étant la modélisation *psycho-sociale* de l'humain avec des « skills » dûment catégorisées par les professionnels de la psychologie, de la sociologie ou des ressources humaines (**`pensée critique`**, **`leadership`**, **`communication orale`**, **`empathie`**…). A l'ère de l'informatisation, ces deux modélisations sont entrées en résonance. Ne nous étonnons donc pas d'être évalués, embauchés ou jugés par des IA et, surtout, de l'*accepter* puisque nous nous habituons à admettre, tous guillemets dissipés, qu'elles nous *comprennent*.