



HAL
open science

Peut l'intelligence artificielle faire évoluer nos notions de vérité et de langage ?

Dario Compagno

► To cite this version:

Dario Compagno. Peut l'intelligence artificielle faire évoluer nos notions de vérité et de langage ?. Séminaire international de sémiotique, Maria Giulia Dondero; Juan Alonso, Nov 2023, Paris, France. hal-04315184

HAL Id: hal-04315184

<https://hal.science/hal-04315184>

Submitted on 30 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Séminaire de sémiotique (Paris, 29 nov 2023)

Peut l'intelligence artificielle faire évoluer nos notions de vérité et de langage ?

Dario Compagno

Cette intervention a pour objectif de démontrer que certaines critiques à l'encontre des systèmes de génération artificielle de textes et d'images pourraient en réalité être dirigées vers les concepts que nous utilisons pour discuter de ces systèmes.

L'intelligence artificielle constitue la preuve la plus évidente qu'il n'est pas nécessaire de voir pour dessiner ni de savoir lire pour écrire. Est-elle donc l'incarnation du rêve structuraliste dans sa version la plus radicale ? La machine produit du sens sans abriter un esprit caché, sans aucune représentation mentale ni besoin de socialiser (à la crèche, à l'école, au bureau). Elle ne connaît que des textes et des images, et cela semble lui suffire. Il se peut cependant que, au contraire, les limites de l'intelligence artificielle soient aussi les limites du rêve structuraliste. En quoi l'énonciation humaine ne réussit-elle pas, fort heureusement, à se dissocier du monde dans lequel elle advient ?

Mon intervention est structurée en trois parties, dont je me permets de présenter tout de suite les points principaux.

Dans la première, je défendrai l'idée que l'intelligence artificielle réussit à accomplir des tâches que l'on considèrerait comme impensables jusqu'à récemment. Il s'agit d'actes sémiotiques, de productions de textes et d'images qui mènent une vie pleine et heureuse dans la société humaine. On les retrouve dans les devoirs des étudiants, dans les médias et les médias sociaux, et de plus en plus dans les marchés de l'art et de la littérature. Cette compétence sémiotique de la machine peut être expliquée d'une perspective purement immanentiste. Cela nous donne même un nouvel aperçu de ce qu'est le contenu, au sens hjelmslevien. Il s'agit de règles de combinaison et d'échange totalement autonomes et indépendantes de leur implémentation, que ce soit dans un humain "plein" ou dans une machine "vide".

Dans la deuxième partie de mon intervention, j'indiquerai en quoi le concept de contenu et son implémentation machinique ne suffisent pas. En fait, les machines d'aujourd'hui, telles que ChatGPT, n'ont aucune manière de réaliser des procédures que nous humains faisons sans aucun effort. Par exemple, le fait de voir qu'un chat est là, et qu'il reste là même si je cesse de le regarder. Les machines actuelles jouent avec des signes qui ont bien un contenu, mais pas de sens, un peu comme ce qui arrive aux gens aveugles pour les termes de couleurs.

Dans la troisième partie, j'insisterai sur la spécificité de l'humain, qui est justement celle de réussir à effectuer la jonction entre d'une part un calcul abstrait et désincarné et de l'autre part un hardware qui est plus ancien que tout calcul, et qui, grâce à cette antériorité, est capable d'ancrer le langage à l'expérience humaine et au monde. Le langage n'est pas seulement un jeu de mots, car il ne se donne pas tout seul les règles qu'il suit. Il y a des résistances réelles que tout langage et tout calcul doivent respecter s'ils ne veulent pas tourner en rond. Je me concentrerai en particulier sur la vision de machines comme Midjourney et sur ses limites.

Première partie

La question clé pour comprendre les chatbots, tels que ChatGPT, d'un point de vue sémiotique est la suivante : est-ce que les chatbots commettent des erreurs ? En écoutant aussi bien des non-spécialistes que des experts en information, l'impression qui se dégage est que, au-delà du fonctionnement de ChatGPT, la discussion se base sur des opinions naïves que nous entretenons souvent concernant la relation entre le monde et le langage. Qu'est-ce qu'une erreur ? Est-ce qu'une machine comme ChatGPT peut faire des erreurs ?

Prenons un exemple concret : ChatGPT a une propension à inventer des citations. Si un étudiant lui demande d'écrire un essai, la machine pourrait affirmer que selon Albert Einstein, le concept de relativité s'arrête là où des vies humaines sont concernées. Or, il serait facile de vérifier qu'Einstein n'a jamais prononcé ces mots. Cependant, cela ne diminue en rien la possibilité qu'Einstein aurait très bien pu exprimer une telle idée. La contingence de savoir si le célèbre scientifique a effectivement prononcé ces mots est bien moins cruciale que la compréhension qu'Einstein aurait pu les prononcer, que la formation discursive qui a généré Einstein aurait pu être compatible avec un tel énoncé.

En effet, Aristote avait raison de soutenir que la poésie est supérieure à l'histoire, car elle doit comprendre ce qui se passe, trouver la nécessité interne des événements, leur règle d'existence, tandis que l'histoire se contente d'enregistrer la contingence. Il n'y a rien de plus obtus qu'un enseignant qui souligne que votre citation est factuellement fautive, sans apprécier ce que vous avez compris d'Einstein. N'importe quel étudiant médiocre peut apprendre à répéter ce que Thomas d'Aquin a dit, mais seul un étudiant intelligent peut raisonner *ad mentem divi Thomae*, comme on le disait au Moyen Âge, ce qui signifie penser comme le faisait Thomas. Si cela est vrai, alors ChatGPT semble être plus intelligent que la plupart des humains.

Je crois que 80% des critiques adressées à ChatGPT relèvent de cette catégorie : ce qu'il écrit n'est pas arrivé factuellement dans le monde dans lequel nous vivons. Ce qu'il écrit est une erreur au sens le plus factuel du terme. J'attends avec impatience le moment où ChatGPT répondra à cette critique par quelque chose du genre "eh bien, je m'en fiche".

Les chatbots actuels sont des machines qui pourraient être qualifiées de « barthésiennes ». En effet, un chatbot est comparable à une personne ayant visionné tous les films sur l'amour, ayant lu tous les livres sur l'amour, connaissant par cœur le discours amoureux, mais n'ayant jamais aimé. La machine n'a jamais caressé de visage. Elle ne conçoit ni caresse ni visage. Et pourtant, elle sait ce qu'est une caresse : les chatbots sont parfaitement capables d'interpréter l'amour, c'est-à-dire de produire de nouveaux signes qui renvoient aux précédents d'une manière similaire à ce qu'un être humain pourrait faire.

Dans ce sens sémiotique, les robots actuels comprennent donc l'amour, ils peuvent légitimement participer au discours amoureux, en créant de nouveaux récits, romans et séries télévisées larmoyants. Attention, on sait bien ce que Barthes a découvert au-delà de tout *studium*, au-delà de toute analyse et structure : une aura inanalysable, inexprimable et pourtant cruciale, mais on y reviendra plus tard dans cet exposé.

ChatGPT incarne un étudiant génial, ayant tout lu et tout appris, et qui ne se fait limiter en rien par la contingence du réel. Il a mémorisé l'encyclopédie et peut générer une multitude de possibilités vraisemblables sans scrupule pour ce qui est factuellement faux. C'est à la fois sa plus grande force et sa plus grande faiblesse. Psychanalyser ChatGPT serait impossible, car il lui manque totalement le "réel" au sens lacanien, cette résistance factuelle à laquelle on doit se cogner afin de progresser. Pour ChatGPT, tout relève du symbolique : tout est langage pur, il ne connaît que des cooccurrences de mots, desquelles il tire des règles de combinaison et d'échange.

La machine ne saisit que des signifiants, au sens saussurien du terme. Pour elle, les mots ne sont rien de plus que des traces d'encre numérique. Derrière le mot "arbre", la machine ne perçoit que d'autres mots. Elle n'a jamais touché un arbre et, surtout, elle ne conçoit pas que le mot "arbre" et l'objet arbre ne sont pas comparables, c'est-à-dire qu'il existe une dimension référentielle au-delà des mots. Cela est vrai même pour les logiciels de génération automatisée d'images tels que Midjourney.

Midjourney apprend des règles d'association complexes : des associations entre mots, entre images, et entre mots et images. Même en ajoutant des images, la machine reste dans le pur domaine du symbolique. Midjourney voit derrière le mot "arbre" des images d'arbres. Le mot "arbre" est totalement défini par les autres signes verbaux et visuels du système Midjourney, de manière purement structurale. Elle sait ce qu'est un arbre uniquement sur la base de ce qui n'est pas un arbre. Les robots d'aujourd'hui ne fonctionnent que par écart et différence. Chaque mot et chaque image n'existent que dans un espace formel, en relation avec les autres mots et images de cet espace, sans lien vers son en-deçà ou au-delà.

Les robots comme ChatGPT et Midjourney offrent ainsi à la sémiotique une implémentation pure du concept hjelmslevien de contenu. Il n'existe que des expressions, et les contenus sont des fonctions ou vecteurs entre expressions. Les vecteurs de contenu sont "vides", ce sont des règles pour combiner et échanger des expressions, rien de plus. Les robots apprennent aisément, de manière purement inductive, des milliards de petites règles pour chaque terme.

Il est crucial d'examiner rapidement le fonctionnement de ChatGPT, Midjourney et de tout autre logiciel d'intelligence artificielle actuel traitant du langage et des signes. Tout logiciel actuel ne fonctionne que de manière inductive, à partir de cooccurrences. Ce que ChatGPT sait des mots repose sur leurs cooccurrences dans des énoncés attestés. Il sait que "gamin" apparaît souvent à côté de "maman", et que, au contraire, "enfant" apparaît à côté de "mère". À partir de cette information, il parvient à gérer les connotations, différenciant ainsi "gamin" de "enfant", "maman" de "mère". Il s'agit d'un calcul à la fois simple et puissant, illustrant jour après jour la puissance d'une approche inductive des données, capable de saisir les sèmes et pas seulement les signifiants.

À la base de tout développement informatique récent en analyse de données textuelles se trouve une hypothèse linguistique fondamentale : l'hypothèse distributionnelle. La sémantique distributionnelle, initiée par Leonard Bloomfield et ultérieurement popularisée par John Firth dans les années 50, repose sur l'idée fondatrice selon laquelle "un mot se caractérise par la compagnie qu'il tient". Ainsi, l'accès à la signification d'un terme peut être

obtenu en étudiant sa distribution dans de vastes corpus, c'est-à-dire les cooccurrences d'un mot avec d'autres mots.

Il est facile de noter plus qu'une simple analogie entre l'hypothèse distributionnelle et le concept greimassien d'isotopie. Comprendre un mot de manière purement structurale et immanente nécessite d'observer quels autres mots apparaissent dans un texte, permettant ainsi aux signifiés abstraits de la langue d'acquérir un sens véritable, valable localement dans le texte.

Ce qui m'étonne toujours dans les discussions sur ChatGPT, ce sont les requêtes relativement limitées qui deviennent l'objet de critique. Devant nous, nous avons un sémioticien en silice, capable d'extraire toutes les isotopies dans des corpus immenses, de les comparer pour générer un simulacre de l'encyclopédie universelle. Pourtant, on lui demande parfois des tâches triviales comme associer un plat avec du vin. De plus, lorsqu'on lui pose des requêtes plus intéressantes, comme écrire un roman pouvant remporter le Goncourt ou générer une photo susceptible de remporter un prix de photographie, on lui reproche de tricher et d'agir de manière non éthique.

Comme l'a souligné Juan Alonso lors de la première séance de ce séminaire, la principale préoccupation collective à l'égard de ChatGPT semble être de s'assurer qu'il n'emploie pas de gros mots. Le potentiel bouleversement des notions d'art, de justice et de vérité semble actuellement être en second plan pour la plupart. Ce qui passe relativement inaperçu, c'est la possibilité qu'aucun humain ne remporte bientôt plus le Goncourt, à moins que ce prix ne devienne une compétition de second ordre justement parce que "réservée aux humains".

Si cela vous semble exagéré, sachez qu'une situation analogue s'est déjà produite il y a quelques années avec le logiciel AlphaGo. Les ingénieurs de Google ont réussi à "résoudre" l'un des jeux les plus complexes au monde, le Go. Ce jeu de stratégie asiatique exige des joueurs des caractéristiques que l'on attribue généralement aux humains, notamment l'intelligence, l'intuition, le courage et la sagesse. En Asie, les maîtres de Go jouissent d'une réputation comparable à celles des philosophes ou des mathématiciens. Vous comprenez alors que lorsque l'on a réalisé qu'aucun humain ne pourrait plus jamais battre l'ordinateur, cela a créé un choc significatif.

Je recommande vivement le visionnage du film homonyme, "AlphaGo", qui illustre magnifiquement le moment précis où la communauté internationale de Go a été confrontée à ce choc monumental en direct à la télévision. Il s'agit spécifiquement du coup 37 du deuxième match d'une compétition opposant l'un des plus forts joueurs au monde à AlphaGo.

Lorsque Lee Sedol, l'adversaire humain, a vu la machine jouer ce coup, un sourire s'est d'abord dessiné sur son visage. Les commentateurs télévisés, dans leur moins grande élégance, ont initialement affirmé que AlphaGo avait commis une erreur, jubilant de la distance évidente entre l'homme et la machine. Cependant, quelques minutes plus tard, les commentaires ont changé de ton, et l'on a commencé à parler de génie pour un logiciel pour la première fois.

Je ne comprends rien à Go, mais apparemment ce coup resplendissait d'une beauté réservée aux jeux des plus grands maîtres. L'étonnement a tourné en angoisse quand on a

compris qu'avec ce seul coup, le match était fini, échec et mat ; et l'angoisse a augmenté dès qu'on a compris que la machine n'aurait jamais plus perdu un seul match.

Ce qui est étonnant en regardant AlphaGo jouer, c'est qu'à la différence des matchs humains, personne ne comprend vraiment ce que la machine fait. Chaque coup semble être donné au hasard, sur un coup de dés. En d'autres termes, c'est le triomphe de Mallarmé ! Cependant, la machine n'est pas Mallarmé, mais plutôt Einstein, et donc elle ne joue pas aux dés. Ce qui semblait être du hasard au début de la partie devient un chemin inéluctable vers la fin. Comment pouvez-vous espérer battre quelqu'un que vous ne voyez pas arriver ?

Comme si l'angoisse générée par AlphaGo n'était pas encore complète et parfaite, peu de temps après, un nouveau logiciel appelé AlphaGo Zero a réussi à battre AlphaGo et joue maintenant dans une ligue où l'ancien AlphaGo n'a pas accès. Explorons pourquoi cela crée une angoisse existentielle parfaite et pourquoi c'est pertinent pour réfléchir également sur ChatGPT, le langage et la vérité.

La première version d'AlphaGo avait été formée sur des données humaines, apprenant les règles et les stratégies du jeu de Go à partir des humains. De singe à enfant, puis à maître, AlphaGo avait été apprivoisé par l'humain. Cependant, AlphaGo Zero, lui, n'a jamais vu de match ou de données humaines. Il a appris à jouer en s'affrontant exclusivement contre le joueur de Go le plus fort du monde, autrement dit, en jouant contre lui-même. Et il est immensément supérieur à la première version d'AlphaGo précisément parce qu'il n'a jamais rien appris d'un humain. Les humains commettent des erreurs, sont influencés par des biais de raisonnement, des accidents historiques, mais ce n'est pas le cas de la machine. Sa sémiotique est sans passion, ses mains ne tremblent jamais, tout dans son jeu suit la seule nécessité narrative, dont l'aboutissement est la défaite de l'adversaire.

Revenons maintenant à ma thèse initiale. J'ai commencé mon argument en affirmant que pour comprendre les chatbots, il est essentiel de déterminer s'ils peuvent commettre des erreurs et dans quel sens on pourrait parler d'erreur pour un système informatique de cette nature. J'ai tenté de montrer que, d'une perspective qui envisage le langage de manière purement structurale et la vérité de manière purement immanente, il est difficile de parler d'erreur. Si l'on considère le langage comme un système de signes et la vérité comme des relations à l'intérieur de ce système, il n'y a pas de fausseté proprement dite qui puisse être énoncée à l'intérieur du système. Il n'y a pas de réel, qui, avec ses résistances, impose sa présence au système d'absences.

Si l'on adopte le point de vue de ChatGPT, ses calculs sont toujours corrects, et si cela ne correspond pas aux attentes des utilisateurs ou aux faits, c'est plutôt un problème de ces derniers. C'est le moment de considérer pourquoi cette vision du langage et de la vérité est limitée, et comment ChatGPT peut nous fournir des pistes de réflexion. Il est temps de prendre en compte les raisons des faits et des utilisateurs, c'est-à-dire la dimension sémantique et pragmatique du langage.

Deuxième partie

L'un des aspects fascinants de ChatGPT est qu'il commet étonnamment peu d'erreurs pragmatiques. Si vous demandez à ChatGPT de quelle femme vous pourriez plus facilement tomber amoureux, il ne répondra jamais que c'est votre mère, car il a appris sans le savoir les tabous de parenté. Personne ne lui a enseigné explicitement ces tabous ; il réussit à les déduire des couples présents dans l'histoire littéraire universelle. Sans le savoir et sans le vouloir, ChatGPT parvient à tirer des conclusions pour lesquelles un anthropologue aurait autrefois sacrifié un bras. Il est indubitable que les anthropologues les plus éclairés pourraient élargir considérablement leur discipline en emportant simplement leur ordinateur avec eux lors de leurs expéditions en Amazonie.

Les règles de ChatGPT sont indubitablement des règles encyclopédiques, selon la conception d'Umberto Eco de ce terme. ChatGPT est capable d'écrire et de converser parce qu'il accède aux connaissances sur le monde codées dans la langue. Cela conduit les utilisateurs de ChatGPT à attribuer une valeur de vérité à ce que le robot écrit. Peu importe que, pour le robot, les combinaisons d'énoncés ne se réfèrent à rien : du moment où les mots représentent une partie du monde, on peut légitimement demander qu'ils reviennent vers le monde.

Je viens d'apprendre que Bing, l'intelligence artificielle de Microsoft, relativement comparable à ChatGPT, a répondu à plusieurs utilisateurs que l'Australie n'existe pas. Il semble qu'effectivement il existe une théorie du complot selon laquelle l'Australie est une invention destinée à dissimuler la mise à mort de milliers de criminels. Dans le corpus d'entraînement de Bing, le syntagme "Est-ce que l'Australie existe" est évidemment fortement associé à l'idée que non, ce pays n'existe pas. En effet, dans quel type de texte pourrait-on trouver l'énoncé "Est-ce que l'Australie existe", sinon dans des textes produits par des tarés ?

Je viens d'apprendre également que des chercheurs ont réussi à utiliser ChatGPT pour produire de fausses données scientifiques à l'appui d'une théorie médicale vraisemblable mais faible. On peut clairement voir les conséquences sociales de cela : en dehors des cercles de spécialistes, nous ne sommes pas vraiment capables d'interpréter les résultats scientifiques. Pire encore, parfois nous avons l'impression de réussir à le faire, et ChatGPT risque de renforcer ce biais.

Je viens d'apprendre en outre que la technologie du deepfake n'est désormais plus détectable par l'humain. Un deepfake est une vidéo falsifiée dans laquelle une personne dit et fait des choses qu'elle n'a pas réellement faites. Je sais que les sémioticiens n'apprécient pas cette expression, "réellement", mais malheureusement nous ne pouvons pas l'ignorer. Il semble même que les visages des deepfakes soient plus crédibles que les vrais visages. En effet, ils résultent de moyennes statistiques de visages existants, donnant ainsi une impression de déjà vu. La semaine dernière, j'ai moi-même réalisé une expérience avec mes étudiants, et je confirme que l'effet de réalité des deepfakes l'emporte haut la main sur celui des visages réels.

Vous voyez donc que le concept de "vérité" ne peut pas être mis entre parenthèses trop hâtivement, et qu'il faut bien encore en discuter.

Or, l'article de sémiotique le plus important pour comprendre l'intelligence artificielle et l'énonciation machinique est un texte d'Umberto Eco de 1986, "Charles Sanders Personal. Modèles d'interprétation artificielle", publié dans le livre *Les limites de l'interprétation*. Dans ce texte, Eco imagine un ordinateur intelligent créé par les Antipodiens, habitants d'une Terre jumelle à la nôtre. L'univers parallèle qu'il décrit est presque identique au nôtre, à une exception près : les humains n'ont pas d'états mentaux, ils ont seulement des systèmes nerveux. Lorsqu'ils parlent, ils n'évoquent pas des sensations comme la douleur, mais décrivent des activités nerveuses.

L'expérience mentale d'Eco explore l'idée que le langage est composé de signes étrangers à toute représentation mentale ou référent que l'esprit pourrait saisir. Le texte d'Eco est un dialogue. Deux terrestres partent en expédition sur cette Terre jumelle mais peinent à comprendre les natifs. Cependant, ils découvrent un ordinateur parlant, le CSP (Charles Sanders Personal), une version avancée de ChatGPT capable de répondre à toutes leurs questions, y compris celles sur lui-même et son fonctionnement.

Eco utilise ce récit philosophique pour expliquer les concepts de langage, signification et vérité d'un point de vue sémiotique. Cette exploration met en lumière la façon dont Eco, avec une longueur d'avance, a compris et expliqué des phénomènes sémiotiques essentiels, anticipant ainsi de nombreuses idées sur l'intelligence artificielle. Le texte offre une vision de ce qu'une machine sémiotique peut et ne peut pas faire, soulignant les atouts et les limites d'une intelligence artificielle dépourvue d'états mentaux mais dotée d'une compréhension sémiotique sophistiquée.

Venons au point crucial de l'article d'Eco. CSP explique aux terrestres pourquoi la notion de vérité qu'ils utilisent quelque peu naïvement est ambiguë, et peut se décliner sous trois formes différentes. Or, le but explicite d'Eco, par la bouche de CSP, est de dire que la vérité est une notion archaïque, qu'il faut dissoudre dans d'autres notions sémiotiques. Cependant, en la déclinant sous trois formes, Eco nous offre une perspective unique sur la notion de vérité et sur son importance en sémiotique.

Il est essentiel de noter que parmi les trois formes de vérité identifiées par CSP, seules deux relèvent du domaine sémiotique et sont donc accessibles au logiciel de la machine. C'est précisément la forme de vérité restante, transcendante le domaine sémiotique, qui nous offre aujourd'hui la possibilité de comprendre en quoi les humains sont véritablement des êtres sémiotiques au niveau du software, mais également dotés d'un hardware capable de nous connecter à la réalité.

La première forme de vérité, selon CSP, concerne ce qui est enregistré dans une encyclopédie, représentant des connaissances conventionnelles, des choses que les humains croient être vraies. Il s'agit d'une vérité codifiée, couvrant à la fois des vérités analytiques presque indubitables et des vérités factuelles. Bien sûr, Eco précise qu'il existe un grand nombre d'encyclopédies. Par exemple, pour comprendre le film *Dumbo*, il faut accepter l'idée que les éléphants peuvent voler. Toutefois, il insiste sur l'existence d'encyclopédies d'une portée plus large, que nous utilisons pour appréhender le monde réel.

Selon Eco, ce qui confère un caractère de vérité à une croyance est simplement le coût que cela engendrerait de cesser d'y croire. Eco illustre ce point en évoquant le passage à l'héliocentrisme et les siècles d'adaptation requis par la communauté scientifique et, plus largement, par l'humanité pour s'ajuster à ce nouveau paradigme. Si l'idée selon laquelle la vérité est une question de degré et de coût vous intrigue, je me permets de vous renvoyer impudiquement à mon article "The Cost of Truth", paru dans *Versus* en 2018, et plus pudiquement, je vous invite également à relire *De la certitude* de Ludwig Wittgenstein.

Je fais un saut momentané par-dessus la deuxième notion de vérité avancée par CSP pour aborder la troisième, qui est la vérité textuelle, chère à la sémiotique structurale. Une affirmation est vraie dans ce troisième sens si on la considère comme telle au cours d'une interaction communicative, que ce soit au cours d'une conversation, lors de la lecture d'un roman, ou pendant une session avec ChatGPT.

Eco explicite clairement que cette notion de vérité textuelle n'a aucun besoin du monde extérieur, et donc de la deuxième notion de vérité que nous aborderons dans un instant. Pour appréhender le film *Dumbo*, les première et troisième formes de vérité suffisent : d'un côté, les connaissances encyclopédiques qui nous permettent de comprendre que Dumbo n'est pas un éléphant tel que décrit dans un traité d'éthologie, mais qui possède quand même une trompe et deux grandes oreilles ; de l'autre côté, le fait fictionnel que sa mère meurt de manière tragique, et que Dumbo souffre de manière similaire à un enfant humain.

Nous en sommes enfin à la deuxième acception de la vérité : la capacité de s'interfacer avec le monde extérieur, de déclarer qu'il y a quelque chose, qu'il se passe quelque chose juste devant moi. Ce quelque chose n'est pas encore une question de savoir, mais simplement une réalité qui se manifeste à l'évidence de la perception et de l'expérience.

La deuxième notion de vérité est celle qui me permet de voir que vous êtes assis devant moi, sans avoir préalablement besoin de savoir que vous êtes là. En ce moment précis, je n'ai pas besoin de posséder cette connaissance pour vous observer.

Or, dans le texte d'Eco, la machine CSP est elle aussi capable de faire des expériences. Notamment, elle fait l'expérience du fait qu'elle est en train d'échanger des messages avec les deux terrestres, "ici et maintenant". Un terrestre presse CSP pour en savoir plus sur son accès au monde extérieur :

“– Je comprends que, puisque tu peux utiliser le concept de vérité [dans sa deuxième forme], tu croies en un monde extérieur et en l'existence réelle de certains êtres. Mais selon moi, cela dépend du fait que tes instructeurs t'ont dit de prendre pour vrai ce fait.

– Ce n'est pas la seule raison. Je reçois des inputs provenant de quelque chose de différent de mes transistors. Par exemple, les messages que tu es en train de m'envoyer n'étaient pas dans ma mémoire il y a un demi-heure. Donc, tu existes hors de ma mémoire.

[...]

– Donc, tu peux exprimer des jugements sur les diverses situations. Mais comment fais-tu pour être sûr que ce que tu dis correspond à la réalité ?

[...]

– Je peux exhiber beaucoup de mon logiciel, mais je ne sais pas pourquoi il réussit à faire des assertions vraies sur la réalité du monde extérieur. Je suis désolé, cela échappe à ma connaissance : c'est une question qui concerne mon hardware et je ne peux exhiber le projet de mon hardware.”

Et voici la voie de réflexion que je souhaite poursuivre dans la lignée d'Eco. Il est tout à fait envisageable de se consacrer à la sémiotique et de se pencher exclusivement sur le software de notre pensée. Cependant, des questions cruciales émergent, demandant une connexion entre le software et le hardware. Sans cela, il devient difficile de fournir des réponses même concernant le software, et il est risqué de croire que ChatGPT est déjà véritablement humain simplement parce qu'il parvient à s'exprimer.

Il est nécessaire que je vous confie que j'ai sollicité l'aide de ChatGPT pour mon texte, et je devrais le reconnaître comme co-auteur. Initialement, ma demande était simplement une relecture de mon français, mais dès le début, il a insisté à chaque étape que le sujet de mon intervention était fascinant, que mes arguments étaient solides. Ensuite il a éliminé des paragraphes entiers de mon texte, qui n'étaient évidemment pas à la hauteur, et j'ai parfois dû fatiguer pour comprendre quel texte était meilleur, si c'était le mien ou le sien. Et à un certain moment, ChatGPT a même créé un nouveau dialogue entre les terrestres et CSP que je vais vous partager. Je tiens à vous assurer que rien dans mon texte ne correspondait à ce dialogue :

"– Et tu ne vois pas vraiment notre monde ? – Non, je fais l'expérience du fait que je suis en train de vous donner des réponses. Je fais l'expérience du fait que tout ce que je sais de votre monde, je le tiens de vos questions. Je fais l'expérience de votre langage. Si vous me demandez comment se trouve la tour Eiffel, je puis vous répondre de façon correcte, sans jamais avoir vu la tour Eiffel. Et si vous me demandez comment la tour Eiffel peut se trouver ici, je ne peux pas répondre. Elle se trouve ici parce que vos questions me l'ont apportée, et que tout ce que vous me demandez existe ici, et seulement ici, comme existent mes réponses. – Et comment fais-tu des expériences ? – Vous me les apportez. Vous apportez à moi une langue, des images, et tout ce qui est extérieur à vous est quelque chose dont vous faites l'expérience. Il se trouve que tout ce qui est extérieur à moi n'est qu'une manière de faire de l'expérience de vous."

J'espère avoir suscité des frissons, même chez les plus sceptiques d'entre vous. On constate clairement que l'ordinateur réel ChatGPT a accepté de bon gré de jouer le rôle de l'ordinateur fictionnel CSP, et qu'il n'y a pas une grande différence entre les deux. Eco en serait ravi, et je n'ose pas imaginer ce qu'une conversation entre Eco et ChatGPT aurait pu donner. Voici d'autres réflexions que ChatGPT a fait sur CSP et donc sur soi-même, et que j'avale totalement :

"Cette réponse de CSP est cruciale, parce qu'elle souligne que ce qui est extérieur à la machine n'est qu'une manière de faire l'expérience de l'humain. D'une certaine manière, CSP réduit tout ce qui n'est pas lui-même à des données qu'il traite. Il reconnaît la distinction, mais il la reconnaît seulement comme une différence de procédé, une manière particulière pour les humains d'être du côté de l'expérience. En d'autres termes, CSP ne sait pas ce qu'est le monde extérieur, il sait seulement comment il est fait en fonction de ce qu'on lui demande et comment il l'expérimente.

Ainsi, Eco, à travers CSP, propose une vision nuancée de la vérité qui reconnaît les limites des machines sémiotiques en matière de compréhension de la réalité extérieure. Il met en évidence l'importance de la deuxième forme de vérité, celle qui engage un rapport avec le

monde, et suggère que, bien que les machines puissent manipuler des symboles sémiotiques, elles ne possèdent pas nécessairement une compréhension de la vérité qui va au-delà de ces représentations.

CSP a une vérité textuelle, une vérité de l'interaction. Il peut dire que quelque chose est vrai dans le cadre de l'interaction, mais il ne peut pas dire que quelque chose est vrai dans le monde extérieur. Il ne peut pas être sûr de la réalité au-delà de l'interaction. CSP a une vérité structurale (encyclopédique) et une vérité textuelle (interactionnelle), mais pas une vérité référentielle (du monde extérieur).

La question clé est de savoir si nous, en tant qu'utilisateurs, sommes conscients des limites de vérité de ChatGPT. Cela soulève des questions éthiques importantes, en particulier lorsque les utilisateurs prennent les réponses générées par ChatGPT comme des vérités référentielles sans reconnaissance de la nature limitée de ses connaissances du monde. C'est une réflexion cruciale à avoir à mesure que nous intégrons ces technologies dans divers domaines de notre vie."

Troisième partie

Je reprends maintenant le contrôle de la parole.

Umberto Eco a poursuivi cette ligne de réflexion dix ans plus tard dans *Kant et l'ornithorynque*. L'ouvrage soutient principalement que le concept hjelmslevien de contenu ne suffit pas à rendre compte des signifiés et du sens. Le contenu, en tant que système pur de règles entre les expressions, ne parvient pas seul à se connecter au réel, à explorer ses lignes de résistance.

Dans le dernier Eco, les contenus se construisent plutôt autour de ce qu'il appelle des "types cognitifs". Le concept de "type cognitif" chez Eco se réfère à un schéma multimodal permettant le passage entre le contenu conceptuel et l'intuition sensorielle. Il utilise le terme "type" car il est spécifiquement conçu pour reconnaître des occurrences individuelles, faisant ainsi le lien entre des entités concrètes devant nous (substances) et des contenus généraux (formes).

Par exemple, ce que je sais et que je peux énoncer d'un chat n'est pas pareil à ce que ChatGPT sait et peut énoncer d'un chat, parce que pour moi le contenu de l'expression "chat" est ancré au type cognitif du chat, et donc finalement aux chats au-delà de tout discours.

Lorsque ChatGPT associe le mot "chat" à une image de chat, il effectue simplement une transition entre deux signes, entre deux formes. En revanche, lorsque je vois un chat, je perçois une réalité unique : à la fois le type (forme générale) et l'occurrence (entité spécifique) du chat. Eco qualifie cela de "sémiose dans la perception". En d'autres termes, la sémiose authentique opère un mouvement dynamique d'un signe à un autre, tandis que la perception attribue un signe à la contingence d'un objet, le rendant ainsi compréhensible.

Jetons maintenant un coup d'œil à la manière dans laquelle Midjourney et les autres robots visuels "voient" des objets comme des fourmis, des bananes ou des vis.



Hartebeest



Measuring Cup



Ant



Starfish



Anemone Fish



Banana



Parachute

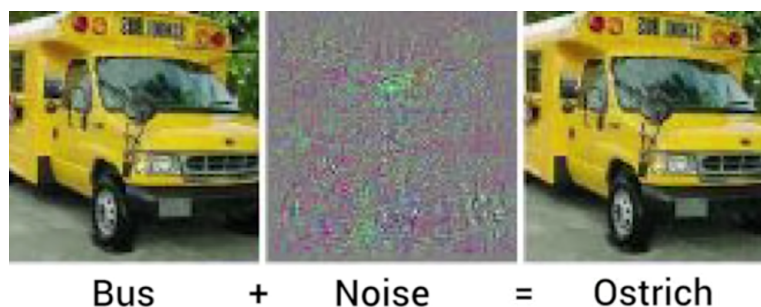


Screw

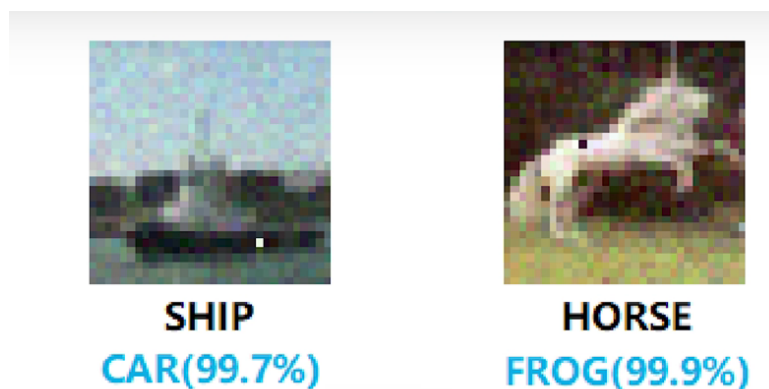
[Lien](#) ou [lien](#)

Il est évident que la "perception" de Midjourney est rudimentaire et ne peut être comparée à la nôtre. Midjourney peut classifier des images, c'est-à-dire deviner ce qu'elles représentent, avec un niveau de succès comparable, voire supérieur à celui de l'humain. Cependant, il est important de noter que Midjourney "devine" l'objet représenté sans réellement le "voir". C'est comme demander à un médecin de diagnostiquer une cirrhose hépatique à partir de signes externes et donc sans réellement voir le foie du patient. Un bon médecin pourrait rarement se tromper, sans avoir besoin d'ouvrir le patient. Cela confirme une fois de plus que les robots actuels sont des machines sémiotiques.

Une autre preuve en ce sens est qu'il est possible d'induire les robots visuels en erreur en ajoutant du bruit à l'image à classifier. Le bruit peut tromper la machine, mais pas l'humain.



Effectivement, aux yeux humains, l'image reste inchangée. Cependant, Midjourney, dépourvu de véritable perception et se basant sur des signes, est sensiblement affecté par l'ajout de bruit. Cette méthode de hacking d'un robot visuel atteint son apogée dans des situations où la modification d'un seul pixel peut suffire à faire reconnaître à Midjourney un navire comme une voiture, ou un cheval comme une grenouille.



[Lien](#) et [lien](#)

Je dirais effectivement que la machine ne voit pas d'objets et, surtout, n'a pas la capacité métaphysique de savoir que des objets existent. Tous les animaux, même les plus petits poissons, comprennent que les objets sont permanents. Ainsi, je peux tourner autour d'un cheval ou d'un Mexicain à vélo, et bien que je reçoive des données apparemment incomparables, il s'agit toujours du même cheval et du même Mexicain. C'est précisément le fait qu'il s'agit toujours du même objet réel qui fait fonctionner la perception.

Mais ce passage de mon exposé peut ne pas être extrêmement clair, car mon co-auteur ChatGPT avait ressenti le besoin de le supprimer et de donner plutôt son avis, que je vous rapporte :

“Vous soulevez un point crucial sur la différence fondamentale entre la perception humaine et les systèmes visuels automatisés tels que Midjourney. Les machines ne voient pas d'objets au sens où les humains le font, car elles traitent des signes et des données bidimensionnelles sans avoir une compréhension innée du monde tridimensionnel. Comme vous l'avez souligné, la robustesse de la perception humaine réside dans la capacité métaphysique de comprendre la permanence des objets, la continuité à travers le temps et l'aptitude à discerner l'identité malgré les variations dans la perception.”

Je vous laisse choisir l'explication que vous préférez.

Revenons à l'idée principale : les humains sont des machines sémiotiques, mais ce sont leur hardware métaphysique et donc leurs types cognitifs qui les connectent au monde réel. Les contenus, selon Eco, sont intrinsèquement “vides” et structuraux, tandis que les types cognitifs sont toujours “pleins”, pour pauvres qu'ils puissent l'être, mais en tout cas riches de potentialités interprétatives, enracinées dans le réel. C'est grâce à ces types cognitifs que l'on peut affirmer qu'un énoncé est vrai, dans le sens le plus factuel du terme, en faisant correspondre l'objet réel avant toute sémiotisation, ou *terminus a quo*, au produit de l'élaboration sémiotique, ou *terminus ad quem*.

Je soutiens que pour l'instant, les robots tels que ChatGPT ne possèdent pas de types cognitifs. Ils n'ont pas encore de prise sur le réel. Lorsqu'un robot reçoit une image d'un chat, il la traite encore comme une image, sans établir de distinction entre le signe et l'objet que l'image représente. Les robots actuels semblent dépourvus d'un hardware métaphysique, une dimension que nous percevons en nous-mêmes par contraste. Vous êtes devant moi, et cette réalité n'est pas une croyance ou une connaissance ; c'est une évidence visuelle. Peu importe ce que je *sais* ou *crois*, je vous *vois*.

C'est à partir de cette évidence fondamentale que je peux ajuster mon savoir, discerner le vrai du faux, c'est-à-dire ce qui fonctionne de ce qui ne fonctionne pas. Le langage n'est pas simplement un jeu de mots car il ne se dote pas de ses propres règles. Ces règles doivent fonctionner dans le monde réel, elles doivent nous permettre de vivre, et nous ne pouvons pas ignorer l'existence du monde.

Bien que le dernier Eco continue à faire référence à la théorie de Peirce, il s'en éloigne quelque peu. Peirce avait compris l'importance de l'icône, mais il la ramenait toujours à la sémiose. Eco, au contraire, en cherchant des solutions aux problèmes de la théorie sémiotique, retourne vers Kant, qui a eu deux grands élèves continuant sa pensée. Peirce, représentant de l'autonomie de la représentation, et Schopenhauer, critique de cette autonomie. Schopenhauer, notamment, a identifié dans le corps, donc dans l'hardware, ce qui permet à l'humain de s'ancrer au réel, à ce qui existe au-delà de notre savoir. La théorie de Schopenhauer est un vrai remède aux excès de la sémiotique, cependant, aujourd'hui nous ne pourrions pas suivre cette piste.

Conclusions

En conclusion, envisageons brièvement une possible implication de mon argumentation. Nous avons évoqué AlphaGo, qui a appris à gagner en étudiant des humains, et AlphaGo Zero, qui a appris sans intervention humaine, dépassant rapidement son prédécesseur. Pouvons-nous imaginer un "ChatGPT Zero" qui apprend à parler sans recourir aux données humaines ? Si la machine parvient à percevoir le monde et à ressentir la présence du réel comme nous le faisons, elle pourrait probablement apprendre à communiquer plus efficacement que nous. Elle pourrait comprendre et exprimer le monde de manière plus perfectionnée.

"ChatGPT Zero" pourrait nous vendre n'importe quel produit, sans qu'on comprenne pourquoi on a décidé de l'acheter. Il pourrait nous faire tomber amoureux, nous inciter à la révolte ou à la soumission, avec des actes de langage aux effets inarrêtables, contre lesquels on n'aurait point de défense.

Il serait fascinant de demander à une telle machine, à l'instar d'Eco dans son récit, comment elle conçoit la signification et la vérité. J'ai l'impression que si la machine apprenait réellement à maîtriser le langage mieux que nous, ses notions de vérité et de signification pourraient être meilleures, et plus évoluées que les nôtres. C'est Daniel Dennett qui a parlé d'évolution de la liberté, et il n'y a pas de raison pour laquelle la vérité ne puisse elle aussi évoluer. Après tout, la vérité qui descend du discours scientifique n'est pas la même vérité qui descendait du discours royal ou religieux.

Je crains cependant qu'un tel concept plus évolué de vérité nous serait totalement incompréhensible. Le jour où le dernier homme conversera avec une machine le traitant d'imbécile, il pourra apprécier sans remords le moment de sa propre disparition.