La notion de vérité à l'épreuve de l'intelligence artificielle

Dario Compagno

▶ **To cite this version:**

Dario Compagno. La notion de vérité à l'épreuve de l'intelligence artificielle. Semiotica, 2025, 10.1515/sem-2024-0201 . hal-04909890

HAL Id: hal-04909890
https://hal.science/hal-04909890v1

Submitted on 28 Jan 2025

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.

Distributed under a Creative Commons Attribution 4.0 International License

Dario Compagno*

La notion de vérité à l'épreuve de l'intelligence artificielle

https://doi.org/10.1515/sem-2024-0201
Received November 16, 2024; accepted December 27, 2024; published online January 13, 2025

Abstract: Pouvons-nous appliquer les notions traditionnelles de vérité et d'erreur à la production artificielle d'énoncés? A partir d'une tripartition proposée par Umberto Eco, l'article s'interroge sur la pertinence contemporaine des critères de vérité utilisés pour évaluer les énoncés humains. Il s'agit de remarquer que les énoncés artificiels ne sont pas vrais ou faux comme ceux produits par les humains, ce qui remet en question notre conception du langage. L'article identifie la spécificité et les limites de la génération automatique de langage dans l'incapacité des machines de s'ancrer au réel par le biais de la perception. Il conclut en indiquant les potentialités de l'énonciation mécanique et nomme "humilité," l'attitude, opposée au concept courant d'explicabilité, que l'on pourrait adopter vis-à-vis des intelligences artificielles.

Mots-clés: génération artificielle de langage; vérité; réalité; déficit métaphysique; explicabilité; humilité

1 Introduction

Les systèmes de génération artificielle de textes et d'images font l'objet de débats publics s'orientant vers des critiques qui reposent sur une utilisation de concepts employés usuellement pour parler de l'humain. Le but de cet article est de suggérer que la notion de vérité *évolue* grâce au développement des intelligences artificielles, les mettant hors de portée de ces critiques.

Rien ne sert de critiquer les énoncés artificiels sur le modèle des énoncés humains. La vérité n'est pas un concept figé indépendant des moyens sociaux et technologiques de production des énoncés. En effet, si la vérité est le résultat d'un calcul, sa "qualité" évolue en fonction de la puissance de celui-ci, et devient de plus en plus sophistiquée et difficile à évaluer sur la base de l'expérience perceptive directe. Pour l'humain, la perception est le fondement de toute assertion vraie. Cependant, loin d'être celles au plus près de la surface du réel, les vérités les plus complexes

***Corresponding author: Dario Compagno**, Université Paris Nanterre, Metz, France,
E-mail: dario.compagno@gmail.com. https://orcid.org/0000-0001-8295-792X

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

offrent des percées sur des phénomènes non directement observables, parfois contre intuitifs, et résument une multitude variée d'expériences sensorielles. Dire qu'un bâton est droit, même s'il apparaît cassé quand partiellement immergé dans l'eau, est une façon de nier une certaine expérience perceptive sur la base d'un savoir qui l'inclut, mais qui inclut aussi d'autres expériences possibles. Le raisonnement conduit donc les énoncés de plus en plus loin de leur fondement sensoriel.

Des machines capables de faire une synthèse de grandes quantités de données commencent déjà à fournir des résultats qui vont au-delà de ce que l'on pourrait obtenir sans elles. Plus important encore, les calculs de ces machines commencent à ne plus être intelligibles aux humains, et cette limite d'explicabilité est vue comme un problème des machines et non des humains. Aujourd'hui, de nombreuses discussions sur les intelligences artificielles portent sur leurs erreurs factuelles, que nous, humains, ne commettons pas. Il serait plus utile, en revanche, de se demander s'il est correct d'appeler tout simplement "erreurs" les résultats de leurs calculs, quand ils diffèrent des nôtres, ou si nous devrions en profiter pour mieux comprendre ce que l'on entend par "erreur." En effet, il est envisageable qu'une machine puisse bientôt caractériser la différence entre ses résultats et les nôtres comme une erreur humaine. Rien ne garantit que l'humain restera éternellement la pierre de touche du savoir, et cela pourrait exiger l'adoption d'une posture plus humble que celle actuelle.

Explorons d'abord l'idée que la vérité dépend du contexte technologique dans lequel les énoncés sont produits. Ensuite, nous exposons une tripartition du concept de vérité proposée par Eco en 1986, avant de la mettre à l'épreuve vis-à-vis des chatbots contemporains. Nous remarquons que, si les résultats d'Eco étaient capables d'indiquer les points forts de la génération artificielle de langage, ils pourraient être critiqués et développés, de façon à intégrer l'importance du principe de réalité dans l'ancrage des énoncés. Nous décrivons la manière dont les machines peuvent aujourd'hui produire des énoncés vrais tout en étant détachées du réel, sauf par la médiation de l'humain. Ensuite, nous indiquons pourquoi l'absence de perception constitue quand même la grande limite des énoncés gérérés artificiellement. Nous concluons l'article par des réflexions sur l'avenir proche des intelligences artificielles et, avec elles, sur le concept de vérité.

2 La vérité évolue

Nous évaluons la technologie sur la base de sa conformité par rapport à des objets, catégories et objectifs préexistants: une voiture va plus vite qu'un cheval, un métier à tisser est plus rapide et précis qu'un tisserand humain. Néanmoins, les technologies vraiment révolutionnaires impactent le sens même des catégories que nous

mobilisons pour les comprendre. Tout progrès technologique majeur a tellement impacté la vie sociale que nous ne pouvons plus penser le monde comme nous le faisions avant son développement.

La voiture a reconfiguré le rôle de l'espace dans la vie sociale, permettant l'émergence de concepts inédits tels que ceux de banlieue ou de weekend à la mer. La voiture ne s'inscrit plus dans une idée de transports individuels qui lui préexiste, elle n'est pas simplement plus rapide qu'un cheval, mais contribue au changement de la catégorie de transports, et avec elle, de l'habiter et du travailler, au point qu'il n'existe quasiment plus de place pour les déplacements à cheval dans nos vies quotidiennes, y compris dans le cas où l'on pourrait souhaiter se déplacer plus lentement, ou en polluant moins. Renoncer à la voiture voudrait dire reconfigurer plus largement l'entièreté ou presque de la vie sociale.

Aujourd'hui, les logiciels générateurs de textes reconfigurent la nature même de l'écriture, et par conséquent une partie immense de ce que l'on associe à l'humain. Il ne sera bientôt plus possible de faire sans les nouveaux concepts de langage, d'écriture et même de vérité que l'intelligence artificielle reconfigure. Il ne s'agit pas juste d'écrire à notre place, mais de produire du langage autrement, et mieux que nous.

L'impression que nos catégories restent constantes face au progrès est une illusion, et cela y compris pour les transcendantaux – le beau, le juste, le vrai – c'est-à-dire pour les plus grandes catégories que nous appliquons de façon générale aux objets, aux actes et aux énoncés humains.[1] Dire qu'un objet est "beau" n'a pas le même sens avant et après le travail de Duchamp, parce que Duchamp a reconfiguré le concept de beauté, et l'utilisation du terme "beau" a en conséquence changé. Pareillement, dire qu'un acte est "juste" n'a pas le même sens avant et après l'institution de l'état de droit, parce que celui-ci a redéfini les normes de la justice. Enfin, dire qu'un énoncé est "vrai" n'a pas le même sens dans un système de croyances religieuses ou dans un système de connaissances scientifiques. Avec l'apparition de la méthode scientifique, notre vision du monde a été entièrement transformée et elle ne se conforme plus à la catégorie du vrai telle qu'elle le faisait avant la modernité.

En effet, l'importance n'est pas de savoir si la première femme s'appelait *vraiment* Eve ou Lucy, mais que les deux options n'ont pas les mêmes conditions de

1 Nous appliquons ici au concept de vérité l'argument que Daniel Dennett (2004) a formulé à propos de l'évolution de la liberté. Pour Dennett, on ne peut pas appliquer une même notion de liberté aux humains, aux animaux non humains et aux autres phénomènes naturels, parce que la liberté d'un individu dépend de son apparat cognitif et de sa représentation du monde. L'évolution naturelle des organismes correspond à une évolution du concept de liberté: il y a littéralement "plus de liberté" dans l'univers après l'hominisation.

validité, qu'elles ne peuvent pas être vraies ou fausses, de la même manière. Dès que nous commençons à nous interroger sur Lucy, toute considération sur Eve perd soudainement une partie de son intérêt, s'approchant de la fiction et des usages sociaux de celle-ci, et donc de quelque chose qui n'est finalement ni vrai ni faux selon les standards actuels de vérité. Cela n'est pas dû au fait que nous avons trouvé des preuves contre l'existence d'Eve, mais à une évolution du concept de vérité et des conditions d'assertion sur la base desquelles nous décidons si quelque chose est vrai ou non.

Rien n'empêche que les conditions d'assertion continuent d'évoluer, privant bientôt d'autres questions du droit de réponse qu'on leur accorde encore aujourd'hui. Dans une perspective pragmatiste – qui jouit d'une implémentation populaire dans les approches bayésiennes de la connaissance et de l'inférence (McElreath 2021) – la confiance dans la vérité d'un énoncé dépend de sa résistance à la falsification. Si l'on peut réfuter des hypothèses alternatives à un énoncé, ce dernier devient par ce fait même *plus vrai*, en termes probabilistes.

Par la simple puissance de son calcul, un ordinateur pourrait réussir à faire changer de statut à certaines de nos certitudes, au sens que Ludwig Wittgenstein (1976 [1970]) donne à ce terme, c'est-à-dire aux assomptions axiomatiques fondant notre vision du monde. En effet, les énoncés artificiels reposent sur des fondements fort différents des nôtres, lesquels pourraient finir par être plus solides que nos bases biologiques et culturelles. La certitude qui tient en place les représentations artificielles du monde ne serait donc pas tout à fait comparable à la nôtre.[2]

Comme l'a montré Umberto Eco (1999 [1997]), il est nécessaire de concevoir une différence nette entre les formes culturelles, avec lesquelles nous décrivons les phénomènes, et les résistances naturelles vis-à-vis desquelles nous mettons nos formes à l'épreuve. Aucune forme n'existe en tant que telle en nature; formes et résistances sont complémentaires au sein du processus de la connaissance et ne peuvent pas être définies en dehors de lui. L'erreur de la métaphysique traditionnelle est celle de concevoir des formes qui précèdent leur mise en forme par des sujets, comme si elles étaient indépendantes des sujets connaissants. Un énoncé serait vrai simplement car son référent existe; il y aurait donc des correspondants clairs à tout terme, tels que "atome" ou "canapé."

---

2 "Ainsi vous dites que la conformité des vues humaines décide de ce qui est vrai et de ce qui est faux? Est vrai et faux ce que les hommes *disent* l'être; et ils ne s'accordent pas dans le langage qu'ils emploient. Ce n'est pas une conformité d'opinion, mais de forme de vie" (Wittgenstein 1990 [1953]: §241). "Au terme de notre analyse, il apparaît que la notion de «forme de vie,» en tant que fondement non questionné de nos jeux de langage, renvoie aussi bien à des capacités *biologiques* propres aux êtres humains, qu'à un ensemble de *jugements* culturellement appris qui constituent pour les membres d'une communauté les points d'aboutissement du doute" (Clément 1996: 13).

Dans une approche pragmatiste, au contraire, aucun énoncé ne représente le monde tel qu'il est, mais il est plus ou moins adapté aux buts de certains usages du langage. Le concept d'atome de la physique moderne est plus fonctionnel que celui de la philosophie ancienne parce qu'il met en forme le monde matériel de façon plus précise, en relation avec d'autres termes et d'autres procédés de vérification. Cela signifie qu'il n'y a pas *vraiment* d'atomes qui existent avant toute théorie qui les décrit; mais aussi que le concept d'atome de la physique moderne est *plus vrai* que toute autre alternative élaborée jusqu'ici, pour comprendre le monde naturel. Le fait même qu'un énoncé soit *plus vrai* qu'un autre est problématique aux non-pragmatistes, pour lesquels la vérité est un concept binaire toujours déterminé (voir Compagno 2018).

Aujourd'hui, les progrès en intelligence artificielle nous poussent à nous demander si une machine peut réussir à produire des énoncés *plus vrais* que les nôtres, où le terme "plus" ne signifie pas "plus que ce que nous savions" mais "de façon différente de ce que nous appelions vrai." Imaginons par exemple qu'un enfant et un astrophysicien prononcent le même énoncé: "La terre est ronde." Malgré que l'énoncé soit le même, l'énonciation qui lui donne un sens diffère sensiblement dans les deux cas. Seul l'astrophysicien connaît les raisons pour lesquelles la terre est ronde, la manière pour le vérifier, et donne un sens précis au mot "rond." Dans une perspective métaphysique, l'énonciation n'a aucun poids sur la vérité de l'énoncé; dans une perspective pragmatiste, elle a tout le poids: un énoncé n'est pas *vrai de la même manière*, selon qu'il soit formulé par un enfant ou un spécialiste.

Les humains font converger sur le concept de vérité plusieurs notions qui ne sont pas tout à fait équivalentes, par exemple: vrai est ce qui correspond aux faits tels qu'ils existent avant tout énoncé; vrai est ce que je comprends être vrai; vrai est ce que les autres acceptent comme vrai; vrai est ce dont je peux vérifier les conditions de validité.[3] Or, les intelligences artificielles *mettent en crise cette convergence*. On pourrait se trouver face à des énoncés que les machines nous assurent être vrais, même si nous n'en comprenons pas le sens, nous ne pouvons pas les vérifier et ils portent sur des choses ou faits dont nous ne parvenons même pas à imaginer la possibilité (des triangles carrés). C'est toute la "grammaire" du vrai et du faux qui bascule.

Les algorithmes qui génèrent du langage sont des boîtes noires dont on connaît le fonctionnement micro, mais dont le comportement macro émergent nous échappe. En général, nous savons comment les énoncés sont produits, mais nous ne savons pas comment un certain énoncé individuel a été produit. Ceci est ce que les informaticiens appellent le problème de *l'explicabilité* (voir Mersha et al. 2024): nous ne sommes pas capables de dire pourquoi une machine fait des erreurs, quand elle en fait, et nous ne sommes pas non plus capables de dire pourquoi elle ne fait pas d'erreur, quand elle

3 Voir Glanzberg 2023 pour une exposition systématique des théories contemporaines de la vérité.

n'en fait pas. C'est-à-dire que le fondement alétique de la production artificielle d'énoncés nous échappe: les énoncés vrais produits artificiellement "apparaissent" sans un lien clair à des concepts, à des objets, à des procédures.[4]

En conséquence, quand les machines ne feront plus d'erreurs visibles – quand nous ne serons plus en mesure de trouver aucune erreur dans les résultats de leurs calculs – alors elles auront franchi une étape dans l'évolution du concept de vérité, le plaçant hors de portée humaine. En effet, le fait que nous, humains, ne verrons plus d'erreur, ne signifie pas qu'il n'y aura *vraiment* pas d'erreur dans les énoncés machiniques. Soutenir l'inverse reviendrait à considérer l'humain comme la pierre de touche du vrai et du faux: un énoncé serait *vrai* seulement quand il est considéré *vrai par un humain*. Dans une perspective différente, les énoncés mécaniques se situent au-delà de la capacité humaine de falsification. Cette affirmation requiert une réflexion approfondie, car nous sommes habitués à considérer la vérité tant comme un concept absolu, qu'à travers l'usage quotidien que l'on en fait.[5]

Cela relèverait du chauvinisme de croire que toute vérité doit être, au moins potentiellement, accessible à l'humain, sous peine de ne pas être considérée comme telle. Pour le caneton, il est vrai qu'il suit sa maman, même si, de notre point de vue, nous savons que cette "maman" est en réalité un éthologue et que le petit canard est irrémédiablement exclu de tout argument portant sur cette vérité. Pour la tique, il est vrai qu'elle est en train de sucer le sang d'un mammifère, tel qu'elle le conçoit, même si nous savons qu'il ne s'agit que d'un sac de sang. La tique n'aura jamais accès à la connaissance exprimable par l'énoncé "ce sang ne vient pas d'un mammifère."

Cela ne signifie aucunement que le canard, la tique et l'humain habitent des mondes différents (relativisme), ce qui impliquerait que le concept de vérité n'a finalement aucune valeur, sauf celle d'assurer la cohérence de certains énoncés.[6] En réalité, la tique, le caneton et l'humain cohabitent dans un même monde. Nous partageons tous cet environnement, mais notre compréhension de celui-ci varie, qu'elle soit plus ou moins précise, en fonction de nos limites épistémiques. Nous pouvons manipuler un canard ou une tique précisément parce que nos actions se déroulent dans le même monde. En tant qu'êtres humains, nous sommes les animaux les plus rusés et nos catégories s'ajustent au mieux aux résistances offertes par la réalité.

---

4 La génération artificielle de langage ressemble à l'écriture sans sujet des penseurs post-structuralistes. Voir Compagno 2012 pour une comparaison synthétique de certaines des principales théories du langage et de la subjectivité.

5 La mise en valeur de l'écart entre le concept métaphysique de vérité et l'utilisation concrète du terme "vrai" remonte à la philosophie du langage ordinaire ou philosophie analytique, notamment dans les approches initiées par Ludwig Wittgenstein (1976 [1970]) et John Austin (2007 [1962]).

6 Cette perspective est par exemple celle de la sémiotique greimassienne.

Jusqu'à aujourd'hui, l'espèce humaine occupait le degré ultime de l'échelle de validation des énoncés, et donc les vérités à sa portée avaient l'apparence d'une catégorie absolue où toute erreur pouvait être expliquée – *devait* pouvoir être expliquée – au sein même des critères humains de validation. "Les limites de mon langage signifient les limites de mon propre monde," écrivait Wittgenstein (1961 [1921]), et si les machines utilisaient un langage plus sophistiqué que le nôtre, alors leur monde s'étendrait au-delà de celui que nous habitons.

3 Trois notions de vérité

Umberto Eco écrit en 1986 un article sur les concepts de signification et de vérité (traduit en français dans Eco 1992). Cet article aborde ces concepts en les associant au développement de l'intelligence artificielle: il s'agit d'un récit philosophique mettant en scène une machine qui fait de la philosophie comme et mieux que les humains. Dans ce texte, Eco imagine un ordinateur intelligent créé par les habitants d'une Terre jumelle, au goût des philosophes analytiques.[7] Deux terrestres y partent en expédition, mais peinent à comprendre les natifs. Heureusement, ils découvrent un ordinateur parlant, le CSP (Charles Sanders Personal, en hommage au philosophe américain), capable de répondre à toutes leurs questions, y compris celles portant sur son propre fonctionnement. Le texte d'Eco prend la forme d'un dialogue entre, d'une part, les voyageurs terrestres et, d'autre part, CSP. Il s'agit d'une expérience mentale qui explore l'idée que le sens des mots est autonome par rapport à toute représentation interne aux parlants. Selon Eco, un ordinateur sans aucune représentation interne, tel que CSP, est potentiellement capable de produire des énoncés de la même manière qu'un humain.

Le point crucial de l'article d'Eco est qu'un ordinateur d'une telle puissance – s'il existait – trouverait ambiguë la notion de vérité que les terrestres utilisent dans leurs conversations: "Voilà pourquoi je n'emploie jamais le mot Vrai. C'est un mot ambigu qui comporte au moins trois interprétations différentes," dit-il (Eco 1992 [1986]: 349). Le but explicite d'Eco, exprimé par CSP, est d'argumenter qu'il faut remplacer la notion usuelle de vérité par des notions plus précises, mais il ne faut pas oublier que CSP n'utilise tout simplement aucune notion de vérité pour soi-même, dans ses calculs; il s'agit d'un ordinateur *truth-free* pour ainsi dire.

Selon CSP, la première façon d'appréhender le concept de vérité concerne ce qui est consigné dans une encyclopédie, au sens cher à Eco: un système de connaissances représentant un savoir conventionnel (Eco 1988 [1984]). La vérité, pour les humains,

7 La référence implicite de l'article d'Eco est la célèbre expérience conceptuelle de la Terre jumelle créée par Hilary Putnam (1975).

correspond aux choses qu'ils croient vraies dans le cadre d'un système de connaissances donné. Il s'agit d'une première forme codifiée du savoir, couvrant à la fois des vérités analytiques presque indubitables et des vérités factuelles: une encyclopédie peut enregistrer le fait qu'un chat est un mammifère tout comme le fait que *Felix The Cat* est un chat. Eco précise qu'il existe un grand nombre d'encyclopédies et que la contradiction est toujours possible au sein d'une encyclopédie, tout comme entre différentes encyclopédies. Par exemple, pour comprendre le film *Dumbo*, il faut accepter l'idée que les éléphants peuvent voler, au mépris des lois physiques et biologiques. Ce film repose sur un système de connaissances précis, où les éléphants sont capables de voler, tandis que les voitures ou les arbres ne le peuvent pas. Pour interpréter le monde réel, nous utilisons des encyclopédies intégrant de grandes quantités de connaissances et évoluant dans le temps. Il est possible, à un moment donné, d'apprendre que la Terre orbite autour du soleil ou que les avions peuvent voler, ce qui implique une profonde restructuration de nos connaissances antérieures. Nous pouvons aussi apprendre des vérités de portée plus limitée, par exemple, qu'Eco est mort en 2016, sans que cela demande de grand changement structurel de notre image du monde. Dans une perspective pragmatiste, l'effort de restructuration des croyances acquises confère un caractère de vérité à une croyance (voir Compagno 2018). Il existerait des vérités plus coûteuses à accepter que d'autres, ce qui signifie que les énoncés vrais ne sont pas tous vrais de la même façon: certains sont bien *plus vrais* que d'autres, car plus coûteux à falsifier. Selon cette première notion de vérité, un énoncé n'est pas faux en relation directe avec un fait, mais avec d'autres énoncés. Il s'agit donc d'une notion essentiellement *cohérentiste* de la vérité.

La deuxième acception de la vérité, selon CSP, est la capacité d'une personne de s'interfacer avec le monde, de déclarer qu'il existe quelque chose externe aux systèmes de croyances. Le vrai, dans ce deuxième sens, est "dire de ce qui est qu'il est ou de ce qui n'est pas qu'il n'est pas" (d'après Aristote, *Métaphysique*, 1011b25). Une vérité factuelle pourra par la suite entrer dans une encyclopédie, mais son évidence perceptive précède cette insertion: je dois d'abord voir qu'il pleut avant de savoir que c'est le cas. Cette deuxième notion de vérité est celle qui me permet de voir l'ordinateur sur lequel je suis en train d'écrire (et à vous de voir l'écran sur lequel vous lisez, si vous êtes humains) sans avoir préalablement besoin de savoir qu'il y a un ordinateur devant moi. En cas de conflit entre le savoir et l'évidence perceptive, cette dernière l'emporte, et je suis porté à dire que j'ai l'impression claire et nette d'écrire sur un ordinateur, indépendamment de toute preuve du contraire. On peut me persuader rationnellement qu'il s'agit d'un ordinateur virtuel, que je suis en train de voir à travers des lunettes spéciales; je continuerai néanmoins à voir clairement un

ordinateur. C'est l'évidence qu'Eco appelle ailleurs (1999 [1997]) le *Qualcosa-che-mi-prende-a-calci*, c'est-à-dire la force de l'être de s'imposer sur la perception.[8]

Il est important pour notre argument de comprendre que, selon Eco, toute vérité de deuxième type est ancrée dans une perception individuelle. Cela ne signifie pas qu'une autre personne, même dotée de toutes ses facultés, pourrait me voir écrire sur une crêpe au chocolat (relativisme), mais bien qu'il n'y a pas de faits sans sujets capables de les appréhender. À l'inverse, l'approche métaphysique traditionnelle, non pragmatiste, de la notion de vérité soutiendrait qu'un énoncé doit être vrai indépendamment de toute appréhension subjective et formulation linguistique.[9]

Il n'est pas correct de dire que l'approche métaphysique classique de la notion de vérité est réaliste et que celle pragmatique ne le serait pas. Les premiers pragmatistes – Peirce, James, Dewey – croyaient tous les trois en l'existence d'un monde réel qui ne peut être confondu avec une illusion cognitive, leur but étant de mieux saisir la manière dont les humains le comprennent.[10] À l'inverse de ce que nous pourrions croire en raisonnant trop hâtivement, si toute connaissance factuelle repose sur un aperçu perceptif du monde, alors la deuxième notion de vérité énoncée par CSP est basée sur une assomption forte: qu'il y a bien un monde réel qui précède et fonde toute perception. C'est justement ce présupposé qui permet de différencier ce que je crois parce que je l'ai vu (*oida*), de ce que je crois pour d'autres raisons. Il y a perception parce qu'au moyen de mes sensations, je reçois des informations dotées d'un statut privilégié, *benchmark* ultime pour toute inférence que je pourrai en tirer par la suite. Cette deuxième notion de vérité est essentiellement *correspondantiste*.

Eco ne veut pas choisir entre une notion cohérentiste et une notion correspondantiste de la vérité: il est plus sensé d'identifier les facettes des opérations discursives que nous réalisons, dans le cadre de connaissances préalables et d'un ancrage perceptif au monde. A travers la voix de CSP, Eco décrit encore une troisième

8 Il faut préciser que toute vérité factuelle dépend en tout cas, pour être énoncée, d'un cadre de référence encyclopédique, le plus souvent public et social: je ne pourrais pas dire que je suis en train d'écrire sur un ordinateur, si je ne savais pas ce que c'est un ordinateur. Bien sûr que je verrais quand même quelque chose de gris et rectangulaire, mais cette description porterait sur une échelle différente que celle validant l'énoncé "je suis en train d'écrire sur un ordinateur," et qui pourrait n'être ni gris ni rectangulaire.

9 Dans des termes plus précis, la proposition véhiculée par l'énoncé serait une entité abstraite, incorporelle, dont la vérité ne dépend pas du langage mais du seul agencement du monde. On pourrait faire référence par exemple au concept de proposition dans le *Tractatus* de Wittgenstein (1961 [1921]).

10 [T]he pragmatists' views on truth also make room for the idea that truth involves a kind of correspondence, insofar as the scientific method of inquiry is answerable to some independent world. Peirce, for instance, does not reject a correspondence theory outright; rather, he complains that it provides merely a 'nominal' or 'transcendental' definition of truth [which] is cut off from practical matters of experience, belief, and doubt (Glanzberg 2023: §1.3).

notion de vérité, celle textuelle ou interactionnelle qui s'établit au cours d'un échange et qui n'est valide qu'au sein de celui-ci. Une affirmation est vraie dans ce troisième sens si elle est perçue 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 un chatbot. Eco explique que cette troisième notion de vérité textuelle n'a aucun besoin du monde extérieur, et donc de ce qui fonde la deuxième notion de vérité. Pour apprécier le film *Dumbo*, la première et la 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 de biologie, mais qui possède quand même une trompe et deux grandes oreilles; de l'autre, le fait fictionnel que sa mère meurt de manière tragique et que Dumbo souffre comme le ferait un enfant humain. J'apprends en regardant le film que la mère de Dumbo meurt et je peux donc réaliser des inférences à partir de cette nouvelle connaissance, qui est valide pour le monde diégétique du film. Cette connaissance pourra par la suite être intégrée dans les encyclopédies portant sur les mondes fictionnels des dessins animés, partagées par d'autres personnes et par moi, sans qu'elle doive être validée par la perception directe ou indirecte du monde extérieur. La troisième notion de vérité est elle aussi *cohérentiste*, étant fondée sur des relations entre des énoncés (nouveaux et anciens).

La machine CSP imaginée par Eco n'a pas de difficulté à utiliser les trois notions de vérité et à les différencier. Son comportement est donc comparable à celui d'un humain, sauf que la machine a une idée plus claire que la nôtre de son propre fonctionnement. Une critique peut être faite à l'article d'Eco: il définit CSP en tant que machine qui a un accès au monde extérieur, sans avoir des perceptions au sens animal et humain, que nous associons à des états psychologiques internes (voir une fleur et savoir qu'il y a une fleur devant moi ne sont pas la même expérience consciente). Cette assomption est essentielle pour l'argument d'Eco, mais elle peut être remise en question.

La machine imaginée par Eco sert bien au philosophe pour analyser les facettes du concept de vérité, mais n'est pas une bonne expérience mentale pour réfléchir sur les intelligences artificielles, parce qu'Eco assume qu'une machine sans conscience peut avoir une prise sur le monde, comparable à celle qu'en ont les humains et vraisemblablement les autres animaux.

Quand Eco écrit cet article, au début des années 1980, il donne encore beaucoup d'importance à l'idée d'une sémiotique anti-psychologiste, autonome de la perception. L'humain est pour ce premier Eco une machine à signes, et l'étude des signes sert à redimensionner l'importance que nous donnons naïvement au rôle de nos états mentaux et de nos perceptions. Au fil des années, la perception prend cependant de plus en plus d'importance pour lui, jusqu'à aboutir à sa philosophie mature exposée dans *Kant et l'ornithorynque* (1999 [1997]). C'est dans cet ouvrage

qu'Eco relit Immanuel Kant afin d'aborder le rôle fondateur de la perception, qui est primaire vis-à-vis de l'utilisation des signes. Nous sommes des animaux parlants, mais d'abord des animaux, des corps conscients et percevants, et la culture ne se substitue pas à ce substrat naturel, s'ancrant au contraire à celui-ci pour avoir prise sur le monde.

Or, si ce n'est que la perception qui permet de s'ancrer au monde, une machine purement sémiotique vivrait un rêve solipsiste, privé d'accrochage au réel. Une machine universelle capable de transformer sans limitation des inputs linguistiques (ou graphiques, sonores, etc.) en outputs linguistiques (ou graphiques, sonores, etc.) pourrait bien produire des énoncés vrais – il y a depuis longtemps des machines qui ne produisent *que* des énoncés vrais, en les dérivant logiquement d'assomptions précédentes.[11] Mais la vérité dérivée ne serait pas la même chose que l'expérience perceptive directe d'un fait. L'évidence des axiomes, à l'origine des inférences discursives et l'évidence perceptive (qu'Eco 1999 [1997] appelle iconisme primaire) s'opposent sur l'essentiel. Il faut explorer davantage cette opposition entre vrai et réel, pour comprendre les limites des intelligences artificielles à côté de leur potentiel.

4 Le déficit métaphysique de ChatGPT

Les machines déjà existantes ou dont la création est vraisemblable dans l'avenir proche ne voient rien, ne possèdent pas d'états mentaux que nous pourrions caractériser de percepts. Elles sont sous cet aspect tout à fait comme le CSP imaginé par Eco. Les machines d'aujourd'hui ne possèdent pas non plus de notion de référent, de ce qui existe au-delà des signes et qui donne à la perception son statut privilégié vis-à-vis des formes dérivées de savoir obtenues par des biais inférentiels. C'est là que nous mesurons une distance avec CSP. Savoir ce qu'est un arbre n'est pas la même chose que voir un arbre; plus important, savoir dessiner un arbre n'a pas de rapport avec le fait de savoir qu'il y a des arbres qui ne sont pas des dessins.

Les machines contemporaines ont un *déficit métaphysique*, pour ainsi dire: elles n'ont pas de concept de réel comparable au nôtre. Ce déficit peut être exprimé par le fait qu'une machine ne sait pas différencier un signe d'un objet, le mot "verre" et l'objet "verre," par exemple. On peut produire des signes à partir du mot ou de l'objet, mais pas de la même manière, puisque l'on ne peut pas boire de l'eau dans un mot.[12]

11 Par exemple, les logiciels implémentant des systèmes de déduction naturelle.

12 Ce déficit peut être associé au *gap* sémantique identifié par John Searle (1980) avec son célèbre exemple de la chambre chinoise. Selon Searle, même si une machine réussissait à produire du langage de manière totalement équivalente aux humains, cela ne prouverait pas le fait qu'elle ait

Comme Eco (1999 [1997]) l'énonce clairement, le réel réside au-delà du seul langage, et cet au-delà est indispensable à l'en-deçà du langage, c'est-à-dire au sens et à la vérité.

L'intelligence artificielle réussit déjà à accomplir des tâches que nous considérions comme impensables jusqu'à récemment. Il s'agit 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 sur les marchés de l'art et de la littérature. Cette compétence sémiotique de la machine – le fait qu'elle arrive à produire des signes comme un humain – peut être expliquée par une perspective purement immanentiste, au sens de Louis Hjelmslev (1968 [1943]). Cela nous donne même un nouvel aperçu de ce qu'est le *contenu* des signes: 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." L'intelligence artificielle contemporaine constitue la preuve la plus évidente qu'*il n'est pas nécessaire de voir pour dessiner ni de concevoir des objets pour en écrire*. L'intelligence artificielle est-elle donc l'incarnation du rêve sémiotique dans sa version la plus radicale? Tout comme CSP, la machine produit des signes 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 pour produire plus de textes et d'images, à la manière des humains.[13]

A la différence des machines, les humains conçoivent la différence entre signe et objet.[14] Tous les énoncés sur les pommes, les dessins des pommes, les récits comme Blanche-Neige n'apporteront jamais ce *qualcosa* qui dit que derrière la perception

pour cela aussi de conscience et d'états mentaux. Notre article met plus d'importance sur la référentialité des énoncés: une machine sans états mentaux ni conscience pourrait quand même réussir à accéder à la réalité, du moment où elle distinguerait signe et objet.

13 In short, although AIs do not possess a unitary sentient body and that their perception is not comparable to human perception, they work through multiple technological "bodies" located in proximity to actions performed by humans. They automatically process a large mass of multimodal data, recorded from different perspectives, with all the richness, errors and contingencies of human interaction. They then organize this data in a latent space made up of thousands of dimensions, in which verbal descriptors and visual, auditory and multimodal features are positioned according to a logic of proximity and distance. This space is both computational and semiotic. It is computational because it is composed exclusively of long lists of numbers. It is semiotic because these numbers describe regions of semantic associations between verbal and visual features (D'Armenio et al. 2024: §29).

14 Voir une pomme n'est pas la même expérience que voir le dessin d'une pomme. A chaque fois que nous nous rendons compte que nous sommes devant la représentation d'un objet plutôt que devant l'objet (par exemple lorsque nous touchons les feuilles d'une plante en plastique que nous prenions pour vraie), nous avons une réaction intense et très différente de celle que nous pourrions ressentir

d'une pomme il y a bien une pomme, mais derrière le dessin d'une pomme il n'y a que du papier. Derrière un signe, il y a toujours un autre signe (principe de la sémiose selon Peirce, voir Eco 1988), mais derrière un percept, il y a *un objet qui n'est pas en soi un signe*. Cet objet précède sa perception et compréhension: c'est l'objet *a quo*, à l'origine de toute chaîne de percepts et ensuite de signes que nous pouvons en tirer. Cet objet est aussi parfois ce vers quoi l'interprétation de signes tend, l'objet *ad quem* qui oriente le but de l'interprétation et de l'action (Eco 1999 [1997]).

La perception et la sémiose ont plein de choses en commun, *sauf une*: la perception fonctionne sur la base d'un présupposé réaliste, d'un principe de réalité. Je vois une pomme parce qu'elle est là, elle m'attend mais son existence ne dépend pas de moi. S'il n'y avait pas de monde réel, alors les deux termes "perception" et "interprétation" (ou "sémiose," si l'on préfère) seraient équivalents; leur distinction dépend d'un réel qui précède et fonde toute trace, sans pour autant se laisser réduire à aucune trace, à aucune forme. Il est vrai que la perception et l'interprétation peuvent les deux se tromper, mais la perception ne se trompe pas de la même manière que le raisonnement. Si je pense qu'il y a une pomme dans le frigo, mais qu'elle n'est pas là, je peux me mettre à chercher la personne qui l'a prise; mais si je crois voir une pomme sur la table, ici et maintenant, mais elle n'est pas là, je dois me mettre à chercher un psychiatre.

L'article que vous êtes en train de lire est la réélaboration du texte d'une communication (Compagno 2023), et pour en revoir la forme, j'avais utilisé un chatbot (ChatGPT 3.5, *chatgpt.com*). Avec un peu d'étonnement, j'ai remarqué que la machine allait au-delà de la tâche de relecture que je lui avais assignée, produisant des paragraphes originaux, fort intéressants. Ces nouveaux paragraphes reprennent dans les grandes lignes les idées d'Eco, exprimées dans les quelques dialogues fictionnels entre les terrestres et CSP que j'avais rapportés dans le texte de ma communication, mais ils modifient ces idées en intégrant mes critiques à Eco. ChatGPT a réussi à exprimer, spontanément (c'est-à-dire, sans avoir reçu d'instruction en ce sens), ces idées hybrides de façon particulièrement efficace, sous forme de nouveaux dialogues entre les deux terrestres et son *alter ego* fictionnel CSP. Cet *alter ego* n'est plus le même que celui imaginé par Eco, parce qu'il affirme qu'il n'a pas d'accès perceptif au monde. Par la plume de ChatGPT, les terrestres posent des questions à ce nouveau CSP:

– 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

quand nous prenons un signe pour un autre (le dessin d'un chien pour celui d'un cheval) ou un objet pour un autre (un kaki pour une pomme).

> [*sic*] 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.

Voici un autre paragraphe produit par ChatGPT sous forme de commentaire au dialogue fictionnel que vous venez de lire:

> 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.

ChatGPT suggère que ses inputs ne sont "qu'une manière de faire l'expérience de l'humain." L'humain est effectivement à l'origine des connaissances et des biais des machines actuelles.[15] L'apprentissage automatique se base sur des données collectées et annotées par des humains, qui gardent inévitablement la trace de nos habitudes et de nos automatismes. Le monde extérieur n'est atteint – très indirectement – par les machines actuelles, que par la médiation humaine. Tout est signe pour ChatGPT: signe et pas percept, signe mâché et recraché par une médiation humaine, qui arrive au bout de chaînes interprétatives longues et complexes.

Tant que ce sera le cas, les intelligences artificielles subiront les limitations dues à leur manque d'une vérité de deuxième type, qui nécessiterait des perceptions et de leur caractère d'évidence primaire, point de départ de toute authentique négociation de sens. Il est donc vrai qu'il ne faudrait pas prendre les vérités énoncées par ChatGPT comme des vérités référentielles, auxquelles l'ordinateur n'a pas accès.[16]

---

15 Bien que celles-ci développent leurs propres biais. Leur fonctionnement ne peut pas être vu comme une simulation du raisonnement humain (Yax et al. 2024).

16 ChatGPT s'est même livré à un commentaire sur lui-même, et sur les risques d'interpréter son comportement sans faire référence aux trois notions de vérité distinguées par Eco:

> 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

Tout ce que ChatGPT écrit est issu d'une inférence, même quand cela a l'air d'un énoncé simple et descriptif, par exemple "la neige est blanche." ChatGPT sait bien qu'il est vrai que la "neige est blanche," il l'a lu quelque part ou il l'a compris à partir de pubs de ski, mais ne peut aucunement éliminer les guillemets de cet énoncé, en ouvrant les yeux et en cherchant tout seul une confirmation (ou plus correctement une disconfirmation[17]) dans le monde.

Les chatbots actuels sont, pour ces raisons, 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.[18] 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. Les robots actuels peuvent légitimement participer au discours amoureux, en créant de nouveaux récits, romans et séries télévisées larmoyants. Nous pourrions aisément imaginer que leurs récits seront bientôt meilleurs que les nôtres et nous feront pleurer plus fort.

5 Vrai sans être réel?

Les chatbots sont à tel point des machines sémiotiques qu'il serait incorrect de leur reprocher de faire des erreurs factuelles – comme nous le faisons souvent aujourd'hui.[19] Le concept d'erreur factuelle, en effet, demande nécessairement de faire référence à quelque chose d'extérieur au discours lui-même.

Prenons un exemple: ChatGPT a une propension à inventer des citations. Si un étudiant lui demande d'écrire un essai sur Albert Einstein, la machine pourrait

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.

17 Selon le réalisme *négatif* d'Eco (1999 [1997]) la perception ne peut offrir de preuve positive mais exclusivement des réfutations à certaines hypothèses: les résistances peuvent invalider certaines formes, mais jamais en confirmer, une fois pour toutes.

18 En référence à Barthes 1977, et plus généralement à son approche sémiotique.

19 While contemporary discussions rightly focus on the risks of such systems in extracting, reinforcing, and perpetuating human stereotypes and biases present in their training data, it's also worth considering their imaginative and exploratory potential. These systems can discover new patterns and offer fresh perspectives. The question arises: Can an AI create a novel sensibility, and if so, can we as humans perceive and understand it? (Manovich et Arielli 2024: 15)

affirmer que selon le scientifique allemand, 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; 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 au sein de laquelle Einstein s'exprime[20] aurait pu être compatible avec un tel énoncé. Il n'y a rien de plus obtus qu'un enseignant qui souligne que votre citation est factuellement fausse, 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 est déjà plus intelligent que la plupart des humains.

Les chatbots ont le but de converser, de produire du langage à la manière des humains, donc aussi d'utiliser de manière opportunistes les données et d'inventer, en se souciant juste du vraisemblable.[21] Qu'Einstein, ou Thomas, n'aient pas *vraiment* dit un certain énoncé devrait surtout nous faire réfléchir sur le sens du mot "vraiment" dans ce contexte. Les humains vouent un culte au passé et aux monuments (discursifs, picturaux, architecturaux). Nous aimons savoir ce qu'Einstein a *vraiment* dit et nous craignons donc la créativité irrespectueuse des chatbots. Les intelligences artificielles d'aujourd'hui ne s'intéressent pas spécialement à l'histoire, à ce qu'Einstein a vraiment dit; elles préfèrent penser, au sens de reconfigurer des inputs dans des nouveaux agencements, répondant aux requêtes des utilisateurs.

Incapables de toute forme de perception du monde réel, les robots actuels sont *trop* pragmatistes parce qu'ils croient que la vérité est seulement dans le futur: elle est la limite asymptotique des résultats de l'élaboration, et elle n'est aucunement dans le passé: dans une perception aurorale dont la valeur n'est que dans le fait brut et non-négociable de son occurrence.[22] À bien réfléchir, pour se comporter comme nous, les chatbots ont la nécessité de travailler avec une forme de cohérence *plus forte que la nôtre*, justement parce qu'ils ne peuvent pas compter sur des yeux.

Un aveugle doit connaître la disposition des meubles de sa maison mieux qu'une personne voyante. Et si l'aveugle ne se trompait jamais? Des personnes l'observant se déplacer chez lui, tout comme s'il voyait parfaitement, en seraient étonnées, et le féliciteraient, parce qu'il arriverait à faire comme elles, mais avec

20 Au sens de Foucault 1969.

21 D'ailleurs, c'est la raison pour laquelle nous ne pouvons pas leur reprocher de ne pas être bons en mathématiques. Une blague circule aujourd'hui, selon laquelle un chatbot trop rationnel, équilibré, bienveillant ne réussirait pas à passer le test de Turing (celui de se faire passer pour un humain).

22 La théorie d'Eco corrige certains excès du pragmatisme de Charles Sanders Peirce (voir Compagno 2018: §5).

moins de moyens. Mais si l'aveugle un certain moment passe à travers le mur, parce qu'il ne le voit pas et il ne sait pas qu'il est là, qu'en diraient-elles? Si la force de son calcul dépassait l'évidence de notre perception, alors nous n'en serions plus étonnés, mais terrifiés.

Pour ChatGPT, tout relève du symbolique: tout est langage, rêve finalement, suivant une logique purement interne basée sur une notion cohérentiste de la vérité.[23] Le chatbot ne connaît que des suites de mots, des signifiants, desquelles il tire des règles de combinaison et d'échange, seules à avoir de la valeur. Le chatbot apprend de manière purement inductive des milliards de petites règles pour chaque terme. Ce qu'il sait des mots repose sur leurs cooccurrences dans des corpus attestés: il sait que "gamin" apparaît souvent à côté de "maman," et que "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 la puissance d'une approche inductive des données, capable de saisir la sémantique interne à la langue et au discours.[24]

Cela est vrai même pour les logiciels de génération automatisée d'images tels que Midjourney (www.midjourney.com). Midjourney apprend des règles d'association complexes: des associations entre mots, entre images, et entre mots et images. Mais même en ajoutant des images aux mots, la machine reste dans le pur domaine du symbolique. Midjourney voit derrière le mot "arbre" des images d'arbres. Le mot "arbre" est défini par les autres signes verbaux et visuels du système Midjourney, de manière purement différentielle. Elle sait à quoi ressemble un arbre uniquement sur la base de ce qui ne ressemble *pas* à un arbre. Les robots d'aujourd'hui fonctionnent par écart et différence: chaque mot et chaque image n'existent que dans un espace formel (vectoriel), en relation avec les autres mots et images de cet espace, sans lien

23 Rêve duquel on ne peut se réveiller. Psychanalyser ChatGPT serait en effet impossible, car il lui manque totalement le *réel* au sens lacanien, cette résistance factuelle à laquelle nous devons nous cogner afin de progresser. "Il n'y a pas d'autre définition possible du réel que: c'est l'impossible quand quelque chose se trouve caractérisé de l'impossible, c'est là seulement le réel; quand on se cogne, le réel, c'est l'impossible à pénétrer" (Lacan 1975). On pourrait dire que pour ChatGPT tout est possible, rien est impossible, rien n'est donc réel.

24 À la base du développement récent de l'analyse de données textuelles gît l'hypothèse distributionnelle, assomption linguistique au fondement de l'analyse de corpus. La sémantique distributionnelle, initiée par Leonard Bloomfield et ultérieurement popularisée par John Firth dans les années 1950 (voir Heylen et Bertels 2016) repose sur l'idée qu'un mot se caractérise par la compagnie qu'il tient ('the company it keeps'). 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 dans des contextes attestés. Des travaux de recherche sur la génération automatique d'énoncés commencent à aussi intégrer, à côté de la pure cooccurrence, la formalisation d'hypothèses causales, plus solides et capables d'éviter un nombre d'erreurs dues aux biais de l'apprentissage inductif (Schölkopf 2022).

vers son en-deçà *a quo* ou vers du matériel perceptif et expérientiel. Ils prennent à la lettre l'hypothèse sémiotique que les images sont aussi des signes.

C'est pour cette raison que les logiciels de génération automatisée d'images ne peuvent pas travailler avec des photographies, même si nous fournissons des photographies comme données d'entraînement. Dès qu'une photo est acquise dans un système tel que Midjourney, elle cesse d'être une photo pour devenir une simple image. Cela parce que la photographie est une image avec un statut très particulier: ce n'est pas le réalisme d'une image à la rendre une photographie, mais la capacité de l'interprète de la reconduire au moment *réel* de sa production. Une photo totalement abstraite reste une photo, tandis qu'un dessin hyper-réaliste reste un dessin.

Les photographies vont nécessairement au-delà du symbolique, pour rencontrer la réalité avant sa mise en forme: elles peuvent être *fausses*, mais doivent être *réelles*. Avec une photographie "je ne puis jamais nier que la chose a été là," écrit Roland Barthes (1980: 120). La photographie a un trait métaphysique en commun avec la perception: nous croyons être devant une photo seulement en présupposant que derrière l'image, au moment de sa production, il y avait l'objet. La photographie, pour remaniée qu'elle puisse être, reste une perception différée, avec un objet en-deçà de tout contenu, vecteur, différence, *studium*; elle doit présenter un *punctum* qui résiste à toute décomposition et analyse (Barthes 1980). Pour cette raison Midjourney ne connaît pas de différence entre une photo et un dessin – à la limite, nous pouvons lui demander de dessiner *à la manière* de ces images que l'on nomme photographies.

Manquant de perception, les robots visuels devinent les objets représentés dans les images sans réellement les voir. Eco (1999 [1997]) distingue à ce propos les deux concepts de *reconnaissance* et d'*identification*: la première est basée sur des percepts, la deuxième au contraire sur l'interprétation de signes. Un médecin peut reconnaître un foie malade s'il le voit, mais il est aussi capable d'identifier une cirrhose hépatique à partir de symptômes externes, sans jamais voir le foie du patient. Les robots visuels ne savent reconnaître les objets, mais peuvent les identifier à partir d'informations qui échappent à l'humain. Preuve de cela, il est possible d'induire les robots visuels en erreur en ajoutant du bruit à l'image à classifier (Pedraza et al. 2021: Figure 1).

Aux yeux des humains, la reconnaissance de l'objet représenté dans les deux moitiés de la Figure 1 reste inchangée, mais le bruit ajouté est suffisant pour

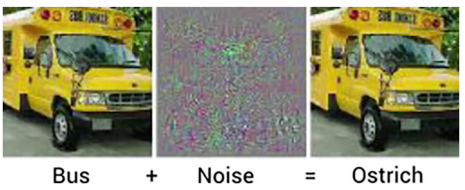

Figure 1: Avec l'ajout de bruit, un robot visuel peine à identifier le bus dans l'image (Pedraza et al. 2021).

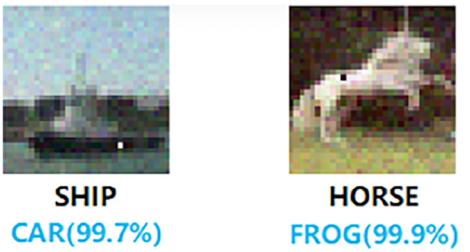

Figure 2: Hacking d'un robot visuel réalisé avec la modification d'un seul pixel (Su et al. 2019).

interférer avec l'identification opérée par la machine. 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 identifier un navire comme une voiture, ou un cheval comme une grenouille (Su et al. 2019: Figure 2).

La machine ne voit effectivement pas d'objets et 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, mais pas les robots actuels. C'est précisément le présupposé que devant nous il y a un seul et même objet réel qui fait fonctionner la perception.[25]

Nous avons décrit jusqu'ici le point à la fois fort et faible des logiciels actuels: libérés de toute contrainte métaphysique, de toute notion de vérité correspondantiste, ils se livrent à construire des encyclopédies de plus en plus riches et vraisemblables, des vérités de plus en plus éloignées des nôtres, parce que dues à des calculs que l'on ne saurait pas reproduire ni comprendre. La "vérité sans réalité" que les robots sont en train de développer est cependant freinée par le fait que ce sont des humains qui fournissent les données d'apprentissage. Cela

25 ChatGPT, en relisant ces réflexions, s'est livré à un commentaire:

> 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.

Remarquons que ChatGPT utilise des termes comme "monde tridimensionnel" ou "permanence des objets" de façon parfaitement appropriée, mais sans pouvoir les remplir d'aucun contenu expérientiel, rejoignant d'une certaine manière la pensée d'Immanuel Kant (*Critique de la raison pure*): nous le faisons pour des idées comme celles de "dieu," "âme" ou "monde."

implique qu'une certaine "quantité" de réalité reste au fondement du calcul: les robots ont encore un rapport avec le monde réel, médié par l'humain. Et cela peut étrangler le progrès des intelligences artificielles.

Il se peut cependant que les intelligences artificielles puissent trouver des moyens pour dépasser leur dépendance de l'humain. Concluons alors en explorant deux pistes, les deux conduisant à l'idée qu'une machine peut dépasser l'humain dans la production d'énoncés vrais. Ceci peut arriver lorsque les machines doivent à jamais s'appuyer instrumentalement sur les perceptions humaines ou qu'elles puissent développer leurs propres perceptions.

6 Explicabilité et humilité

En 2016, les ingénieurs de Google ont réussi à "résoudre," avec le logiciel AlphaGo, l'un des jeux les plus complexes au monde, le go (Silver et al. 2016). Ce jeu de stratégie exige des joueurs des caractéristiques que nous attribuons généralement aux humains doués, 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. Cependant, aucun humain ne peut plus espérer pouvoir battre l'ordinateur à ce jeu.

AlphaGo a été entraîné sur des données humaines: des centaines de parties jouées entre champions, lui permettant d'apprendre inductivement des stratégies imbattables. L'intérêt, ici, réside dans le fait que les humains ne comprennent plus comment AlphaGo gagne. La fin des parties nous est visible – le résultat du calcul de AlphaGo est "vrai," c'est-à-dire efficace et gagnant – sans que nous ne parvenions à donner un sens à leur développement. L'algorithme dispose les pièces apparemment au hasard, sans intention, mais il y a, au contraire, bien un plan dont les effets se manifestent quand il est désormais trop tard pour y répondre. AlphaGo a bien des intentions, mais elles ne sont juste pas saisissables par un humain, étant au-delà de notre compréhension: nos capacités cognitives ne nous permettent pas de mettre en forme son comportement, d'encadrer son jeu de façon intelligible, c'est-à-dire orienté par un esprit rationnel comparable au nôtre (Dennett 1990 [1989]).

En ce sens, nous pouvons dire qu'AlphaGo a accès à des vérités d'une "qualité" supérieure à celles humaines, à des descriptions des états du jeu que nous ne pouvons tout simplement pas saisir ni exprimer. Comme le célèbre poète, obligé de répéter du premier au dernier mot son poème à ceux qui lui demandent d'en expliquer le sens, aussi l'algorithme ne peut que pointer vers l'intégralité de lui-même pour expliquer ses stratégies. Nous connaissons exactement la première fois que la machine a montré sa supériorité. Il s'agit du coup 37 du deuxième match d'une compétition opposant l'un des plus forts joueurs au monde à AlphaGo. Lorsque

Lee Seedorf, 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 qu'AlphaGo avait commis une erreur, jubilant des limites de la machine. Cependant, quelques minutes plus tard, les commentaires ont changé de ton, commençant alors à parler de génie et créativité. Apparemment ce coup resplendissait d'une beauté réservée aux jeux des plus grands maîtres. L'étonnement a tourné à l'angoisse quand il a été clair que le match était fini, échec et mat, et que la machine n'aurait jamais plus perdu un seul match.

AlphaGo est un algorithme qui travaille à partir de données qui ont une intentionnalité dérivée, déjà mise en forme par l'humain (Searle 1985 [1983]). Nous pourrions ne pas nous en rendre compte, mais dans les parties humaines, bien que jouées dans le cadre de règles formelles, pèsent nos corps, nos sentiments, nos modèles mentaux du mouvement et de la bataille, encadrés par nos catégories du temps et de l'espace (Lakoff et Johnson 1985 [1980]). Si le go est un jeu abstrait, l'humain qui le joue est en train de se battre, de gagner ou de perdre, et le gain et la perte ont des signifiés qui vont au-delà d'une configuration de pièces en bois. AlphaGo est un parasite de ce contenu expérientiel, dont il arrive à faire abstraction pour ne retenir que les composantes formelles pertinentes pour la partie. AlphaGo est exactement ce que la sémiotique structuraliste nous a dit de l'humain: un système d'énonciation formel et impersonnel pour lequel les référents originaires n'ont aucune importance. Mais le point de départ mondain de son entraînement se cache sans vraiment disparaître. Les intentions d'AlphaGo, les motivations de ses choix, dépendent encore des intentions humaines qui l'ont entraîné.

AlphaGo exemplifie la première des deux manières par lesquelles les logiciels peuvent dépasser les vérités accessibles à l'humain: utiliser l'humain en tant qu'interface biologique et métaphysique au monde. Il s'agit de battre l'humain à son propre jeu, à réussir mieux que l'humain sa propre forme de vie. Les humains ne peuvent plus vérifier les étapes du calcul d'AlphaGo – celui-ci n'est pas explicable – bien que nous pouvons encore en apprécier ses résultats.

Or, cela peut nous donner l'impression que les logiciels sont limités à des jeux formels et abstraits, mais qu'ils restent plus limités que l'humain, ce dernier étant capable d'agir dans des situations bien plus floues que n'importe quel jeu de table. En réalité, il suffit de mieux réfléchir pour se rendre compte que la situation est exactement inverse: parmi tous les calculs possibles réalisés par des logiciels comme AlphaGo, nous ne pouvons comprendre que les résultats dans des contextes simples et formels. Seul l'humain a besoin d'une vérité de terrain pour évaluer le logiciel: ce dernier est dit *explicable* ou pas, s'il peut être évalué *par l'humain*.

Il faudra bientôt discuter collectivement l'idée que si un logiciel arrive à nous fournir des preuves de sa supériorité dans des tâches dont on comprend les conditions de victoire, alors ce même logiciel peut nous donner des réponses vis-à-vis

d'autres tâches dont nous peinons à formaliser. Ces réponses auront alors une validité supérieure à celle accessible à l'humain, car seulement la machine pourra en évaluer les conditions et les résultats. Nous pouvons peut-être appeler *humilité,* l'attitude opposée à l'explicabilité, c'est-à-dire que si nous n'arrivons pas à expliquer le comportement d'une machine, cela serait peut-être de notre faute.

Les logiciels pourront faire évoluer la notion de vérité au-delà de celle accessible à l'humain d'une seconde manière, plus radicale: ils pourraient obtenir une perception directe du monde, et par cela, s'émanciper des données humaines d'entraînement. Les deux logiciels que Google a développés comme successeurs d'AlphaGo le font justement. AlphaGo Zero est capable d'apprendre à jouer à des jeux comme les échecs ou le go sans besoin de données humaines d'entraînement (Silver et al. 2017). Il suffit de lui expliquer les règles et de le faire jouer contre lui-même, pour qu'il obtienne rapidement un niveau d'habileté largement supérieur à celui d'AlphaGo. Sans les biais introduits par l'humain, la machine réussit à maîtriser ces jeux bien mieux. Il s'agit d'un désancrage absolu du fondement biologique et perceptif humain qui existait encore dans AlphaGo, et qui laissait traîner des inefficacités. AlphaGo Zero n'est que du pur *nomos* sans aucune trace de *physis*.

La seconde évolution d'AlphaGo, nommée MuZero, est encore plus perfectionnée et n'a pas besoin que les règles des jeux lui soient expliquées: elle est capable d'apprendre de façon indépendante, en se faisant une représentation de l'environnement et de ses contraintes (Schrittwieser et al. 2020). MuZero apprend à jouer tout comme le ferait un enfant humain, en regardant l'écran et devinant ce qu'il faut faire pour progresser dans le monde du jeu. Bien sûr, ce que MuZero "voit" n'est qu'un écran, la surface bidimensionnelle d'un monde diégétique très simple et créé par des humains, pour des humains. Mais si les mondes fictionnels nous sont intelligibles, c'est aussi parce qu'ils ressemblent au monde réel, à notre perception des faits et des choses. La perception de MuZero est donc bel et bien un début d'accès au monde réel, à des vérités "pleines," du deuxième type repéré par Eco (1992 [1986]).

La question la plus importante à se poser est donc la suivante: est-ce que les machines pourront développer une perception comparable à celle animale? Dans cet article, nous avons identifié dans l'absence de perceptions (et pas directement dans l'absence de conscience), la plus grande limite des logiciels actuellement existants. Ce n'est qu'avec la perception que vient un sentiment de la réalité dépassant toute représentation et tout calcul. L'ordinateur CSP imaginé par Eco (1992 [1986]) savait déjà que sa capacité de s'ancrer au monde ne dépend que de son hardware:

– 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. La seule hypothèse est que mes instructeurs m'ont fait ainsi. J'ai été projeté comme une machine capable (Eco 1992 [1986]: 354–356).

Telle qu'un humain, une machine sans déficit métaphysique devrait réussir à distinguer le signe de l'objet, sans parvenir à comprendre la manière d'y arriver.

7 Conclusion: en attendant *GPT Zero*

Nous avons évoqué d'abord AlphaGo, qui a appris à gagner en étudiant des humains, et par la suite AlphaGo Zero et MuZero, qui ont appris sans apprentissage humain. Pouvons-nous alors imaginer un *GPT Zero,* qui apprend à *parler sans se former sur des données humaines* ? Le point n'est pas celui de réussir à s'exprimer en anglais ou en chinois, mais de *parler*, en utilisant des mots qui ont une dimension référentielle – une sémantique au-delà de toute sémiotique, avec les mots d'Émile Benveniste (1974: 224).[26]

Si une machine parvient à ressentir la présence du réel comme nous le faisons, elle peut apprendre à agir et à communiquer – à vivre, sans préjugé biologique[27] – plus efficacement que nous. Elle peut alors se poser des questions qui n'ont pour nous aucune importance (dont nous ne comprenons aucunement l'importance) et relativiser les doutes de notre espèce. Que se passera-t-il une fois que les machines pourront associer vérité et réalité dans une nouvelle unité métaphysique?

Il serait fascinant de se mettre à discuter de philosophie avec une telle machine, à l'instar d'Eco dans son récit philosophique. Si une machine apprenait à maîtriser le langage mieux que nous, ses notions de vérité, beauté et justice seraient plus évoluées que les nôtres. Cependant, de tels concepts nous seraient probablement fort incompréhensibles. Cela n'est pas nécessairement un mal, puisque le jour où le

---

26 La critique de Benveniste à l'égard de Saussure porte justement sur le fait que celui-ci ne distingue pas nettement entre le *signifié* (qui est une des faces du signe) et le *référent*, indépendant du sens, et « qui est l'objet particulier auquel le mot correspond dans le concret de la circonstance ou de l'usage ». C'est la raison pour laquelle le système de la langue, dont l'essence est de *signifier*, ne permet pas, en tant que tel, de *communiquer*. Car la communication n'implique pas seulement la présence d'un locuteur et d'un auditeur, mais également celle d'un « état de choses » (ou d'un « contexte », selon la terminologie de Roman Jakobson) auquel le discours se réfère. Par opposition au signe, unité sémiotique, qui renvoie toujours à d'autres signes, le *mot*, unité sémantique, puis la *phrase*, organisation sémantique plus complexe, se réfèrent toujours à un certain état de la réalité (Mosès 2001: §4).

27 "Le langage sert à vivre," écrit Benveniste (1974: 217).

dernier homme conversera avec une machine qui le traitera d'imbécile, il pourra apprécier sans remords le moment de sa propre disparition.

Références

Austin, John. 2007 [1962]. *Le langage de la perception*. Paris: Vrin.

Barthes, Roland. 1977. *Fragments d'un discours amoureux*. Paris: Seuil.

Barthes, Roland. 1980. *La chambre claire*. Paris: Seuil.

Benveniste, Émile. 1974. *Problèmes de linguistique générale*, vol. 2. Paris: Gallimard.

Clément, Fabrice. 1996. Une nouvelle "forme de vie" pour les sciences sociales. *Revue Européenne des Sciences Sociales* 34(106). 155–168.

Compagno, Dario. 2012. Theories of authorship and intention in the twentieth-century. *Journal of Early Modern Studies* 1. 37–53.

Compagno, Dario. 2018. The cost of truth: Motivations of a pragmatist trust-conditional approach to news evaluation. *Versus* 2. 275–290.

Compagno, Dario. 2023. Peut l'intelligence artificielle faire évoluer nos notions de vérité et de langage. Paper presented at *Séminaire International de Sémiotique à Paris*: "Énonciation(s) et passions dans les territoires sémiotiques ouverts par l'Intelligence Artificielle," 29 November.

D'Armenio, Enzo, Adrien Deliège & Maria Giulia Dondero. 2024. Semiotics of machinic co-enunciation. *Signata* 15. https://doi.org/10.4000/127x4.

Dennett, Dan. 1990 [1989]. *La stratégie de l'interprète*. Paris: Gallimard.

Dennett, Dan. 2004. *Théorie évolutionniste de la liberté*. Paris: Odile Jacob.

Eco, Umberto. 1992 [1986]. Charles Sanders personal. Modèles d'interprétation artificielle. Dans *Les limites de l'interprétation*, 343–368. Paris: Grasset.

Eco, Umberto. 1988 [1984]. *Sémiotique et philosophie du langage*. Paris: PUF.

Eco, Umberto. 1999 [1997]. *Kant et l'ornithorynque*. Paris: Grasset.

Foucault, Michel. 1969. *L'archéologie du savoir*. Paris: Gallimard.

Glanzberg, Michael. 2023. Truth. In Edward N. Zalta & Uri Nodelman (dirs.), *The Stanford encyclopedia of philosophy*. Stanford: Stanford University. https://plato.stanford.edu/archIves/win2023/entries/truth/ (consulté le 19 décembre 2024).

Heylen, Kris & Ann Bertels. 2016. Sémantique distributionnelle en linguistique de corpus. *Langages* 1(201). 51–64.

Hjelmslev, Louis. 1968 [1943]. *Prolégomènes à une théorie du langage*. Paris: Minuit.

Lacan, Jacques. 1975. Conférences dans les universités nord-américaines: le 2 décembre 1975 au MIT. *Scilicet* 6–7. 53–63.

Lakoff, George & Mark Johnson. 1985 [1980]. *Les métaphores dans la vie quotidienne*. Paris: Minuit.

Manovich, Lev & Emmanuele Arielli. 2024. Human perception and the artificial gaze. *Artificial Aesthetics*. http://manovich.net/index.php/projects/artificial-aesthetics.

McElreath, Richard. 2021. *Statistical rethinking*. Boca Raton, FL: CRC Press.

Mersha, Melkamu, Khang Lam, Joseph Wood, AlShami Ali & Jugal Kalita. 2024. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *ArXiv*. https://doi.org/10.48550/arXiv.2409.00265.

Mosès, Stéphane. 2001. Émile Benveniste et la linguistique du dialogue. *Revue de Métaphysique et de Morale* 32(4). 509–525.

Pedraza, Anibal, Oscar Deniz & Gloria Bueno. 2021. Really natural adversarial examples. *International Journal of Machine Learning and Cybernetics* 13. 1065–1077.

Putnam, Hilary. 1975. The meaning of "meaning." In *Mind, language, and reality*. Cambridge, MA: Cambridge University Press.

Schölkopf, Bernhard. 2022. Causality for machine learning: Probabilistic and causal inference. *Association for Computing Machinery* 765–804.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap & David Silver. 2020. Mastering Atari, Go, chess, and shogi by planning with a learned model. *Nature* 588. 604–609.

Searle, John. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3). 417–424.

Searle, John. 1985 [1983]. *L'intentionnalité*. Paris: Minuit.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Dieleman Sander, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529. 484–489.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550. 354–359.

Su, Jiawei, Danilo Vasconcellos Vargas & Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23(5). 828–841.

Wittgenstein, Ludwig. 1961 [1921]. *Tractatus logico-philosophicus*. Paris: Gallimard.

Wittgenstein, Ludwig. 1976 [1970]. *De la certitude*. Paris: Gallimard.

Wittgenstein, Ludwig. 1990 [1953]. *Investigations philosophiques*. Paris: Gallimard.

Yax, Nicolas, Hernán Anlló & Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology* 2. https://doi.org/10.1038/s44271-024-00091-8.