« Qu'est-ce que tu ne comprends pas ? » Jeux de langage et algorithmes boîte noire[a]

Rémy Demichelis[*]

Abstract
L'enjeu de cet article est de déterminer ce qui pose problème dans notre compréhension des algorithmes dits « boîte noire », une problématique propre à la jeune discipline de l'*Explainable Artificial Intelligence* (XAI). Car, s'il est aisé de comprendre quelque chose que quelqu'un nous explique, c'est plus délicat lorsque personne n'arrive à saisir le problème. Cependant, notre propos consiste à souligner : (1) qu'il convient de parler d'*interprétabilité* plutôt que d'*explicabilité* lorsque nous cherchons à comprendre les modèles, principalement parce que nous n'avons jamais un accès complet et sans ambiguïté à l'information ; (2) que la machine fait face au problème de l'inscrutabilité de la référence, de la même manière que le linguiste imaginé par Willard Van Orman Quine ne peut pas déterminer précisément ce que désigne le terme « *gavagai* » dans une situation de traduction radicale ; (3) qu'il n'y a pas de règle pour l'application de la langue, si ce n'est des « *language games* », comme la linguistique de Ludwig Wittgenstein nous l'enseigne. Il en découle que l'espoir d'arriver à une explicabilité des algorithmes, et donc à la transparence attendue, est sans doute vain : nous ne pouvons nous contenter que d'interprétations qui ne mentionneront jamais la règle de la règle.

Keywords: XAI, Explicabilité, Interprétabilité, Linguistique, Jeux de langage, Ethique.

Abstract
The aim of this article is to understand the problem of "black box" algorithms, an issue inherent to the nascent field of Explainable Artificial Intelligence (XAI). While it is relatively easy to understand something someone explained to us, it becomes more complicated when no one can fully grasp the issue. Our purpose is however to highlight: (1) that we should speak of *interpretability* rather than *explainability* when we seek to understand models, mainly because we never have complete and unambiguous

[a] Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.
[*] Giornalista, dottore di ricerca in filosofia e docente a contratto, email: r_demichelis@parisnanterre.fr.

access to information; (2) that the machines face the problem of the inscrutability of reference, in the same way that the linguist imagined by Willard Van Orman Quine cannot precisely determine what the term "gavagai" refers to in a situation of radical translation; (3) that there is no rule for the application of language, except for "language games", as Ludwig Wittgenstein's linguistics teaches us. The hope of achieving complete explicability and transparency of algorithms is undoubtedly in vain: we can only rely on partial and broad interpretations that will never fully explain the underlying rules.

Keywords: XAI, Explicability, Interpretability, Linguistics, Language games, Ethics.

1. Un souvenir d'école

Lequel était-ce ? Je ne sais plus. Ils étaient plusieurs certainement, ces enseignants qui me demandaient : « Qu'est-ce que tu ne comprends pas ? » Une phrase si souvent entendue qu'elle se noue à l'imaginaire de l'éducation nationale, dans un mélange de bienveillance et de terreur. Comment répondre à cette question d'apparence innocente ? Pour comprendre ce que je ne comprends pas, encore faut-il avoir un indice, une piste, subodorer que je sais vers où me diriger. Bref, encore faut-il avoir un peu compris pour savoir ce que l'on ne comprend pas. Mais la honte surgit de l'incapacité même à formuler son ignorance ; on se trouve plus bête que bête.

Aujourd'hui, nous sommes en proie à cet embarras lorsque nous cherchons à expliquer certains algorithmes d'intelligence artificielle (IA), car leurs raisonnements échappent à une formulation mathématique. Ce n'est pas entièrement un hasard à en croire cette définition de Bruno Bachimont : « L'IA veut traiter informatiquement (c'est la méthode) des problèmes qui nécessitent des connaissances non formalisables pour être résolus (c'est l'objet)[1]. » Si ces connaissances ne sont pas formalisables, vouloir en formaliser une explication – donc une connaissance –, c'est se lancer dans une tâche paradoxale. Ce sont d'ailleurs les systèmes les moins explicables, ceux d'apprentissage automatique fondés sur des inférences statistiques, et particulièrement les réseaux de neurones formels (apprentissage profond), qui se sont avérés les plus utiles à des fins aussi variées que la reconnaissance d'images, la traduction ou la génération de contenus. Ces algorithmes ne répondent plus, par définition, à une formulation en logique formelle de type mathématique, ce qui leur donne une plus grande malléabilité et donc capacité d'adaptation, mais ils échappent ainsi à notre compréhension.

Il est devenu courant de parler de *boîtes noires*, dans le sens où les résultats fournis ne s'expliquent pas facilement. Tout juste pouvons-nous nous appuyer sur des estimations grâce à d'autres méthodes statistiques, mais pas sur l'*explicabilité* qu'offrent les mathématiques. Nous considérons d'ailleurs qu'il convient plutôt de parler d'*interprétabilité* lorsque nous ne pouvons pas arriver à une absence d'ambiguïté.

[1] B. Bachimont, *Le Contrôle dans les systèmes à base de connaissance. Contribution à l'épistémologie de l'intelligence artificielle*, 2nde éd., Hermès, Paris 1994, p. 181.

Cependant, des esprits optimistes ne relâchent pas leurs efforts et espèrent encore parvenir à l'*explicabilité*, au point d'avoir fait émerger la nouvelle discipline de l'*Explainable Artificial Intelligence* (XAI).

Mais avant toute entreprise de sauvetage, la question primordiale devrait être de savoir pourquoi exactement nous n'arrivons pas à atteindre cette absence d'ambiguïté ? Bref, affrontons nos démons infantiles et demandons-nous : qu'est-ce que je ne comprends pas ?

Notre propos consiste à mettre en évidence que la réponse est certainement à chercher du côté de la linguistique, dans le sens où il n'y a pas de possibilité de dire avec le langage les règles qui nous permettent de nommer les choses et qu'il en va de même pour les systèmes d'IA fondés sur l'apprentissage profond. Nous nous situons du côté des pessimistes.

Pour étayer cette thèse, nous commencerons par détailler le fonctionnement des systèmes d'apprentissage profond afin de mieux cerner la problématique. Nous reviendrons ensuite sur l'état de l'art dans l'XAI. Ensuite, nous nous pencherons sur la difficulté d'identifier le critère de dénomination dans la tradition philosophique, principalement avec Wittgenstein. Puis, nous nous inspirerons de son propos pour développer notre propre thèse.

2. Quel est le problème ?

A la question « pourquoi le système d'IA a donné tel résultat plutôt que tel autre ? », il devient difficile de répondre depuis la révolution de l'apprentissage automatique par réseaux de neurones artificiels, dit IA « connexionniste ». Jusqu'aux années 2010, l'IA « symbolique » prédominait le champ de l'informatique[2] et le problème de l'explicabilité n'avait pas vraiment lieu d'être : les opérations étaient appliquées sur des symboles qui signifiaient quelque chose pour nous (rouge, chien, grand, etc.). Puis, quelques innovations[3] en vision par ordinateurs vinrent changer le paradigme et redonner du lustre aux réseaux de neurones artificiels. Elles furent principalement soutenues par les grandes masses de données nouvellement à disposition grâce au développement d'Internet. Des masses de données nécessaires à l'apprentissage machine, car les réseaux de neurones artificiels ont besoin d'importantes quantités d'informations pour s'entraîner et induire automatiquement le meilleur paramétrage afin de répondre à un problème.

Seulement, cela signifie que le système reste fortement influencé par ses données d'apprentissage sans que nous puissions identifier dans l'algorithme où ni comment cette influence s'exerce. C'est ainsi qu'un système de catégorisation des images peut se mettre à assimiler par erreur des paysages enneigés à des photos de

[2] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, Farrar Straus & Giroux, New York 2019, emplact. 366.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, in « Communications of the ACM », 60, n. 6, 2012, pp. 84-90 ; Q. V. LE, M.A. RANZATO, R. MONGA, et al., *Building high-level features using large scale unsupervised learning*, in *arXiv*, 12 juillet 2012.

loups[4] ; parce que toutes les occurrences de loups dans sa bases de données d'apprentissage montraient les canidés dans la neige. C'est encore de cette manière qu'un système peut se mettre à adjoindre systématiquement un bras humain à un haltère quand on lui demande simplement « un haltère[5] » ; car les haltères étaient tous tenus à bout de bras dans les photos d'entraînement. Nous disons ainsi souvent qu'il y a un *biais* dans ce genre de situation. Le problème de fond est que peuvent se créer ainsi des associations d'idées ou des anticipations automatisées entre des personnes et des faits ou des comportements. Il a ainsi été observé avec des systèmes d'IA une multitude de biais discriminatoires, reproduisant des préjugés sexistes, racistes, islamophobes ou validistes[6].

Identifier ces écueils, les isoler et les déjouer est devenu un enjeu de premier ordre dans nos démocraties. Le règlement européen dit « AI Act », dont la publication a eu lieu en 2024, prévoit ainsi que les systèmes dits « à haut risque », utilisés notamment pour les ressources humaines ou l'attribution de crédits, soient audités afin de « de repérer d'éventuels biais qui sont susceptibles de porter atteinte à la santé et à la sécurité des personnes, d'avoir une incidence négative sur les droits fondamentaux ou de se traduire par une discrimination interdite par le droit de l'Union, en particulier lorsque les données de sortie influencent les entrées pour les opérations futures[7] ». Il conviendra donc pour les éditeurs de prendre « des mesures appropriées visant à détecter, prévenir et atténuer les éventuels biais repérés[8] ».

La grande difficulté pour les informaticien·nes est de réussir à comprendre le fonctionnement des algorithmes d'apprentissage profond, car la valeur de chaque paramètre est rarement compréhensible. Dans les couches inférieures, les systèmes connexionnistes ne manipulent pas des symboles, comme un code couleur tel que #FF0000 pour le rouge, mais des valeurs numériques issues de ce code transformées au cours de multiples opérations. Ces calculs sont parfois très simples, mais comme ils sont effectués de façon croisée, la référence est diluée au fil des calculs : au neurone 456 il est impossible de savoir ce que veut dire le chiffre 0,42. Cela ne signifie rien et c'est à l'image de l'ordinateur surpuissant, dans le film *H2G2 : le guide du voyageur galactique*, qui répond à « la question de la vie, de l'univers et de tout le reste[9] » par

[4] M. Ribeiro, S. Singh et C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, in *arXiv*, 9 août 2016, pp. 9-10.

[5] A. Mordvintsev, C. Olah, M. Tyka, *Inceptionism: Going Deeper into Neural Networks*, « ai.googleblog.com », 18 juin 2015, consulté le 7 novembre 2022.

[6] M. Broussard, *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*, The MIT Press, Cambridge MA 2023 ; J. Buolamwini, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, Random House, New York 2023 ; R. Demichelis, *L'Intelligence artificielle, ses biais et les nôtre. Pourquoi la machine réveille nos démons*, Faubourg, Paris 2024.

[7] Union Européenne, *Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle*, 2024, art. 10-2, f.

[8] ivi, art. 10-2, g.

[9] G. Jennings, *H2G2 : le guide du voyageur galactique*, Spyglass Entertainment, Touchstone Pictures, Hammer & Tongs, 2005, 40:35:00. Exemple mentionné dans R. Demichelis, *L'IA, ses biais et les nôtre*, cit., pp. 12-13.

« 42 ». Cela n'a pas de sens et nous laisse dans le plus grand embarras ; nous ne savons que faire de cette information. Le problème est d'autant plus grand que, dès la phase de vectorisation (lorsque le symbole est transformé en valeurs numériques), les symboles de notre langage naturel disparaissent souvent : en traitement du langage, le logiciel considère rarement un mot isolément lorsqu'il l'intègre dans ses calculs, mais compris dans un groupe de plusieurs mots, selon un certain contexte. Dès lors, les vecteurs sur lesquels sont appliqués des opérations ne représentent qu'un ensemble de mots ou de lettres qui perdent leur signification pour nous.

3. Les solutions et leurs limites

S'il est difficile d'identifier exactement quels paramètres, quels vecteurs ou quelles données ont été déterminantes dans la décision d'un système d'IA, cela n'a pas empêché de nombreux chercheur·es d'essayer de relever le défi. L'explicabilité des algorithmes est ainsi devenue une branche à part entière de la recherche en informatique connue sous l'anglicisme d'Explainable AI ou l'acronyme XAI[10]. L'objectif est donc d'arriver à une *explicabilité*, mais nous soutenons toutefois, comme des professionnels nous l'ont partagé lors de nos enquêtes de terrain, qu'il s'agit davantage d'*interprétabilité*, car il est impossible d'arriver à une absence d'ambiguïté mathématique (ce que désigne le terme d'explicabilité) avec des systèmes statistiques opaques.

Deux approches sont souvent évoquées dans la littérature pour parler des différents types d'interprétation : l'approche *locale* et l'approche *globale*[11]. L'approche locale consiste à identifier les déterminants pour tel individu : pourquoi telle personne a vu sa demande de crédit refusée ? Pourquoi tel chien a été catégorisé correctement comme un chien ? Etc. L'approche globale, cherche à détailler comment le modèle fonctionne en général, quels que soient les individus : quels sont les facteurs déterminants dans la classification des animaux ? Quelles sont les informations cruciales pour obtenir un crédit ? Etc. L'une et l'autre méthode ne sont pas exclusives et peuvent tout à fait être sollicitées de façon complémentaire. L'approche locale aurait toutefois plutôt tendance à s'appliquer à des cas d'usage problématiques, où l'on cherche à savoir pourquoi un individu a essuyé un refus ou a été incorrectement catégorisé.

Il existe pléthore de solutions techniques pour interpréter un algorithme, mais deux d'entre elles sont régulièrement évoquées : LIME[12] et SHAP[13]. Elles visent toutes les deux à dépasser l'écueil de ne se reposer que sur des tests standardisés pour

[10] W. Saeed, C. Omlin, *Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities*, in « Knowledge-Based Systems », 263, 5 mars 2023, art. 110273.
[11] J.-M. John-Mathews, *Interprétabilité en Machine Learning, revue de littérature et perspectives*, Telecom Paris, HAL, 2019.
[12] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?"*, cit.
[13] S. M. Lundberg, S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, in « Advances in Neural Information Processing Systems », 30, 2017.

mesurer la pertinence [*accuracy*] des modèles. Car un algorithme peut s'avérer très pertinent en conditions de laboratoire (tests standardisés), mais pas dans la nature (face à des exemples ou des combinaisons jamais rencontrés et plus proches de la réalité).

LIME a vocation à créer des interprétations locales de tout type d'algorithme. Sa méthode consiste à mesurer le poids de certaines informations en entrée : quels mots ont permis de dire que ce texte parlait de christianisme ou d'athéisme ? Quels pixels ont permis de dire qu'il s'agissait d'un husky plutôt que d'un loup ? LIME ne propose pas d'interprétation globale, mais plusieurs interprétations locales portant sur divers résultats afin d'offrir une meilleure compréhension du modèle.

SHAP, de son côté, vise à estimer le poids de chaque information en entrée (comme LIME), cependant non plus de façon binaire (selon leur présence ou leur absence), mais selon des valeurs (de Shapley). L'enjeu est de mesurer la résistance à la variabilité du contexte et d'assurer la cohérence de l'interprétation, notamment si une information déterminante voit son poids augmenter ou rester identique [*consistency*].

Il est intéressant de remarquer que les articles sur LIME et SHAP font tous les deux appels à des humains pour évaluer l'intelligibilité des interprétations. « Les explications devraient être faciles à comprendre » et visent « une compréhension qualitative », est-il écrit dans le premier tandis que le second mentionne « l'intuition humaine » pour valider sa méthodologie. La compréhension relève d'un aspect qualitatif qui n'est pas formulable et qui requière un échange avec autrui pour savoir s'il a compris. Seul son ressenti peut faire foi.

C'est la raison pour laquelle il est souvent suggéré d'impliquer les utilisateurs et les personnes visées par une technologie aussi bien dans l'audit[14] que dans la création des modèles[15]. Ils est elles sont parfois les mieux placé·es pour savoir quels sont les écueils, les incohérences, affiner l'interprétation ou souligner un manque d'interprétabilité de l'outil, toujours selon certaines situations définies. Même si un modèle est explicable, il ne le sera peut-être pas pour une communauté particulière : on ne s'adresse pas à une scientifique comme à une littéraire. Il convient ainsi de mettre en place une approche herméneutique de l'interprétation. Derrière ce pléonasme se cache une référence à l'herméneutique en tant que critique sociale : c'est-à-dire une interprétation d'après les normes et connaissances[16] propres à une communauté et au plus près d'elle[17], non pas d'un point de vue externe qui ferait descendre l'explication de façon verticale comme un juge impérial.

Olya Kudina parle ainsi d'« analyse interprétative phénoménologique »

[14] M. Broussard, *More than a Glitch*, cit., p. 163.

[15] T. Achiume, UN. Human Rights Council, « Racial discrimination and emerging digital technologies: a human rights analysis », Organisation des Nations Unies, 2020.

[16] Union Européenne, *Règlement (UE) 2024/1689*, cit., art. 5, c ; M. Ribeiro, S. Singh, C. Guestrin, « "Why Should I Trust You?" », cit.

[17] M. Graziani, L. Dutkiewicz, D. Calvaresi, et al., *A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences*, in « Artificial Intelligence Review », 56, 2022, pp. 3473-3504.

[*Interpretative Phenomenological Analysis*][18], en tant que cette méthode se concentre sur des « micro-perspectives situées et des principes philosophiques d'interprétation circulaire[19] » ; la technologie réveille des valeurs propres à une société, mais les influence aussi, provoquant leur réactualisation dans un mouvement de « dynamisme de la valeur[20] » [*value dynamism*], ou dialectique. Il convient ainsi de mettre en place un travail itératif capable de prendre en compte les retours utilisateurs pour obtenir une meilleure compréhension du modèle.

Fabio Paglieri propose, quant à lui, de « suivre l'argent » pour mieux comprendre comment les outils sont orientés[21] :

> Les explications recherchées par l'XAI […] concernent toujours le fonctionnement interne des systèmes d'IA, « comment la magie opère », ce qui est exactement la raison pour laquelle cet enjeu est insaisissable pour l'IA générative et [l'apprentissage automatique]. Il est quelque peu surprenant cependant, que peu d'attention – voire aucune – ne soit portée sur les autres types d'explication, focalisés non sur comment ces système marchent, mais plutôt sur qui en tire profit (ou y perd) du fait qu'ils fonctionnent. « *Cui prodest* ? »

Il y a dans cette idée une certaine herméneutique du soupçon[22], c'est-à-dire une interprétation qui n'hésite pas à spéculer sur les raisons peu avouables et parfois inconscientes d'un locuteur. L'herméneutique du soupçon trouve son incarnation dans la psychanalyse ou le marxisme en tant qu'il y aurait souvent plus à lire dans les propos d'un patient ou d'un adversaire politique que ce qui est explicitement exprimé. Il y a donc l'idée que quelque chose est caché – intentionnellement ou non – qu'il faudrait faire apparaître. C'est un art délicat qui peut basculer dans une certaine folie, dans des surinterprétations excentriques, des pathologies de l'interprétation. Dans la sphère politique, économique et technologique cependant, l'herméneutique du soupçon est aussi une manière de ne pas prendre pour argent comptant les propos des grandes entreprises du numérique et de déjouer des stratégies d'enfumage. L'interprétation devient démystificatrice. « Suivre l'argent » permet de mettre en évidence les raisons qui poussent à la production des outils et comprendre pourquoi ils sont paramétrés d'une telle manière plutôt que d'une autre. Quand il y a des coupes budgétaires pour les expérimentations, l'éthique ou la conformité, nous comprenons mieux pourquoi certains systèmes enfreignent les lois ou la morale. Quand il n'y a pour seul objectif que la rétention des internautes sur le site web, afin de satisfaire les annonceurs et un modèle économique fondé sur la publicité, nous comprenons mieux pourquoi certains contenus sont mis en avant plutôt que d'autres. Les exemples peuvent ainsi se multiplier.

Il est intéressant de remarquer que toutes ces méthodes d'interprétabilité

---

[18] O. Kudina, *Moral Hermeneutics and Technology: Making Moral Sense through Human-Technology-World Relations*, The Rowman & Littlefield Publishing Group, Lanham MD 2023, 182 p.
[19] Ivi, p. 12.
[20] Ivi, p. 3.
[21] F. Paglieri, *Expropriated Minds: On Some Practical Problems of Generative AI, Beyond Our Cognitive Illusions*, in « Philosophy & Technology », 37, n. 2, 2024, p. 55.
[22] J. Michel, *Homo interpretans*, Hermann, Paris 2017, p. 158.

tiennent quasiment pour acquis que le modèle original ne pourra pas être parfaitement expliqué. Comme écrivent Marco Ribeiro et al. (article sur LIME) : « Il est souvent impossible pour une explication d'être complètement fiable [*faithful*] à moins qu'elle soit la description complète du modèle lui-même[23]. » Les logiciels et différentes méthodes évoquées dans cette partie n'ont donc pas véritablement pour ambition l'*explicabilité*, mais l'*interprétabilité*. Il demeurera toujours une incertitude sur les relations de cause à effet dans le modèle original. L'abandon de l'ambition d'explicabilité, c'est aussi le deuil de la transparence totale. Nous ne pourrons pas lire dans un réseaux de neurones artificiels comme dans un livre ouvert (et même le livre nous demande d'interpréter). Seules des méthodes d'*interprétabilité* seront de ce point de vue *satisfaisantes*, mais elles seront *insatisfaisantes* si l'objectif est la transparence diaphane de l'*explicabilité*.

4. Aspects philosophiques : le nom, la carte et la vitre

La difficulté d'identifier clairement ce qui dans une image définit un chat, un loup ou tout autre objet est déjà exprimée dans l'Antiquité d'une autre manière. Pour moquer Platon, qui avait défini l'homme comme un « bipède sans plumes[24] », Diogène de Sinope, dit le Cynique, lui apporta un coq plumé et s'exclama : « Voilà l'homme de Platon ! » Cette anecdote traduit la difficulté de toute définition qui risque toujours d'omettre certains aspects d'un objet. Que dire aussi des accidents ? Un homme sans jambes n'est plus un bipède et n'est pourtant pas moins homme. Dès lors énumérer des critères pour catégoriser des objets risque de négliger certains aspects, d'oublier quelques exceptions.

Cette idée que le modèle d'explication (la définition) *n'est pas* le modèle original (l'objet) se retrouve déjà dans la littérature scientifique lorsqu'il s'agit de différencier la carte et le territoire. « La carte n'est pas le territoire[25] », disait Alfred Korzybski en 1931. La légende raconte que ce mot lui a été inspiré d'une triste expérience : durant la Première Guerre Mondiale, des soldats sous son commandement auraient été abattus par une mitrailleuse prussienne qui n'était pas mentionnée sur la carte[26].

Mais plus que cela : le modèle d'explication *ne doit pas* être le modèle original. Jorge Luis Borges écrivit une nouvelle absurde en 1946, *De la rigueur de la science*[27], dans laquelle des géographes créent une carte qui représente exactement tout le territoire d'un Empire, c'est-à-dire à échelle 1:1. Seulement, les générations suivantes la jugent bien évidemment « inutile » [*Inútil*]. Cela signifie que représenter l'Empire de façon

---

[23] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?"*, cit.
[24] D. Laërce, *Vies et doctrines des philosophes illustres*, tr. fr. de T. Dorandis, Le Livre de Poche, Paris 1999, p. 718.
[25] A. Korzybski, *A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics* (1931), in *Semanticscholar.org*, 2004, p. 750.
[26] H. Krivine, *Comprendre sans prévoir, prévoir sans comprendre*, Cassini, Paris 2018, p. 98.
[27] J. L. Borges, *Histoire universelle de l'infamie / Histoire de l'éternité* (1935), tr. fr. de R. Caillois L. Guille, 10/18, Paris 1994, p. 107.

symétrique « point par point » ne sert à rien et que l'utilité d'une carte est justement de synthétiser l'information, dans le sens de résumer. Il faut que la carte contienne moins d'informations que le territoire qu'elle représente pour servir à quoi que ce soit. Avec l'IA, il faut que le modèle d'explication contienne moins d'informations que le modèle original, et cela même si le modèle original est explicable, c'est-à-dire sans ambiguïté. Comme l'écrivent Ribeiro et al.[28] :

> Si des centaines ou des milliers de caractéristiques contribuent significativement à la prédiction, il n'est pas raisonnable d'attendre de la part de n'importe quel utilisateur de comprendre pourquoi la prédiction est faite, même si chaque poids peut être inspecté.

La métaphore de l'inutilité de la *carte parfaite* traduit aussi l'idée qu'il n'est pas de *terme parfait* pour correspondre à la chose à laquelle il fait référence. Si nous pensons à un mot comme à une carte, alors il ne peut pas la recouvrir parfaitement. La carte comme les concepts impliquent une médiation, une traduction et presque une trahison de la chose visée. Bref, une interprétation et cela va dans le sens de ce que nous disions sur la vanité de toute entreprise d'explicabilité et sur le deuil nécessaire de la transparence totale. Emmanuel Alloa remarque qu'une « vitre qui est vraiment transparente finit par nier sa propre existence matérielle[29] » ; il n'y a de transparence que parce qu'il y a d'abord un obstacle. Il ajoute : « La promesse d'une circulation libre et informée […] finit par confiner le mouvement à un schéma méticuleusement prédéfini. »

Appliquons ce propos à l'IA : assigner à un réseau de neurones artificiels un chemin préétabli reviendrait à limiter ses potentialités. L'apprentissage profond profite de sa souplesse pour s'adapter à des situations variées pour lesquelles les règles rigides sont insuffisantes. S'il fallait imposer des règles rigides et les figer dans l'algorithme en pensant le rendre ainsi explicable, ce serait lui ôter sa capacité d'adaptation et son utilité initiale ; une *bonne vieille IA* pourrait faire aussi bien, donc les réseaux de neurones artificiels deviendraient caducs (même les IA dites « hybrides » [*neuro-symbolic*][30] ou « raisonnantes[31] » conservent en grande partie la souplesse que leur accorde l'inférence statistique).

Diogène, Borges, Korzybski ou Alloa nous montrent que le concept, la carte ou la modélisation, n'atteignent jamais l'objectif de correspondance parfaite à la chose censée être représentée. Nous saisissons qu'il est dans la nature même du medium

---

[28] M. Ribeiro, S. Singh et C. Guestrin, « "Why Should I Trust You?" », cit.

[29] E. Alloa, *Seeing Through a Glass, Darkly. The Transparency Paradox*, in E. Alloa (sous la dir. de), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor*, Leuven University Press, Leuven 2022, pp. 9-25.

[30] T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, *Solving olympiad geometry without human demonstrations*, in « Nature », 625, n. 7995, 2024, pp. 476-482 ; B. Braunschweig, *IA : de l'intérêt d'un système hybride*, in « Les Echos », n. 24167, 11 mars 2024.

[31] T. Brown, B. Mann, N. Ryder, et al., *Language Models are Few-Shot Learners*, in *arXiv*. 2020 ; A. Leveau-Vallier, *Que comprend-on de ce que « comprend » ChatGPT ?*, in « Multitudes », 96, n. 3, 2024, pp. 160-166.

utilisé de receler une dimension pratique, portative parfois, afin de trouver une applicabilité. En IA, cette application recherchée est la compréhension, mais elle ne saurait donc passer par une tentative de description « point par point ». Dès lors, l'IA apporte sa contribution – avec quelques siècles de retards – à la querelle médiévale des universaux[32] : d'un côté, les réalistes croyaient en un isomorphisme entre le monde, les choses et les mots, de l'autre les nominalistes ne voyaient dans nos catégories que des objets sémantiques et non des essences réelles. L'IA connexionniste semble plaider pour ces derniers et ajoute une certaine dose de scepticisme. Il n'y aura pas de modèle symbolique définitif des essences, il n'y aura que des estimations grossières (sans que cela soit péjoratif) qui restent intimement dépendantes des modalités d'accès. Cependant, est-ce la seule limite à notre compréhension des modèles informatiques ? Nous avons jusqu'à présent abordé la question de l'interprétabilité sous l'angle de ce qui est décrit activement. Mais autre chose se joue dans notre connaissance. Il s'agit maintenant d'aborder la question à travers le processus d'apprentissage de façon située.

Car, un concept, contrairement à une carte, s'applique rarement à un seul cas particulier, à un seul territoire, mais à plusieurs dont les spécificités nourrissent la signification par rétroaction. La signification des mots dépend ainsi intimement des usages, selon différents contextes, différentes situations. Autrement dit, toute application enrichit la signification et donne donc lieu à un apprentissage généralisant à partir de configurations situées.

5. Le problème linguistique profond

La difficulté à laquelle les ingénieur·es sont confronté·es pour comprendre ce que désigne véritablement la machine est similaire à celle rencontrée par le « linguiste » décrit par Willard Van Orman Quine[33]. Le philosophe imagine une situation de « traduction radicale », c'est-à-dire une situation dans laquelle deux personnes parlant chacune une langue différente se rencontrent et essayent d'échanger verbalement pour la première fois. Avec un ethnocentrisme propre à son époque, Quine envisage ainsi un dialogue entre un « indigène » [*native*], d'une contrée inconnue, et ledit linguiste, qui se trouve évidemment être locuteur anglophone. Si l'indigène montre du doigt un lapin et prononce le terme de « *gavagai* », que doit en tirer comme conclusion le linguiste ? *Gavagai* peut aussi bien désigner le lapin, ses oreilles, ses moustaches, sa tête entière ou encore l'animal dans telle position. Il y a ainsi une « inscrutabilité de la référence[34] », comme l'histoire de la philosophie appelle ce problème, ou une « indétermination de la traduction[35] », comme l'écrit Quine. Sans pouvoir s'appuyer

[32] A. Conti, *Realism*, in R. Pasnau (sous la dir. de), *The Cambridge History of Medieval Philosophy*, 2 vol., Cambridge University Press, Cambridge 2009, vol. 2, pp. 647-660.
[33] W. V. O. Quine, *Word and object*, Technology Press of the Massachusetts Institute of Technology, Cambridge MA 1960, chap. 2.
[34] Ivi, p. 53. Juste [*inscrutability*].
[35] W. V. O. Quine, *Relativité de l'ontologie et quelques autres essais* (1977), Aubier, Paris 2008, p. 48.

sur d'autres concepts, le linguiste est dans l'embarras. Il doit passer par une inférence à la meilleure interprétation possible, selon le contexte, comme un logiciel d'apprentissage profond. Nous ne devrions ainsi pas être surpris de découvrir que le terme de *loup* désigne la neige pour un système de vision par ordinateur : il se retrouve face à l'inscrutabilité de la référence et tente de la dépasser comme il peut. Sans autres éléments pour lui indiquer le contraire, il peut continuer à estimer pendant longtemps que la neige sur l'image s'appelle *loup*.

L'inscrutabilité de la référence avait déjà été mise en évidence par Ludwig Wittgenstein quelques années auparavant. Lui ne disait pas *gavagai*, mais « *tove*[36] » (ce mot n'existe pas plus que celui de Quine) pour désigner soit *un crayon*, *rond*, *bois*, *un*, *dur* ou encore autre chose. Et il écrit que « c'est le travail de la définition ostensive de donner une signification[37] ».

Wittgenstein a ainsi défini la signification [*meaning*] d'un mot comme le fruit de « jeux de langage » [*language games*][38]. C'est-à-dire qu'elle se constitue au cours de situations durant lesquelles est fait usage d'un terme de façon ostensive [*ostensive*] ; lorsque quelqu'un pointe quelque chose qu'il désigne selon un certain contexte. Wittgenstein prend en exemple un ouvrier qui montrerait à un autre des matériaux ou des outils parmi d'autres objets. La signification des mots apparaît ainsi au fil de leurs usages. « Les jeux de langage sont les formes de langage avec lesquelles les enfants commencent à faire usage des mots. L'étude de jeux de langage est l'étude de formes primitives de langage ou de langages primitifs[39]. » Selon nous, les systèmes d'IA d'apprentissage profond, particulièrement – mais pas seulement – de vision par ordinateur, se retrouvent dans des situations similaires lors de leur entraînement. Ce qu'il faut comprendre en creux, et ce que Wittgenstein explique très bien, est qu'il n'est pas de règle pour l'application du langage en tant que règles ; *il n'est pas de règle de l'application de la règle*. Dès lors, chercher à savoir pourquoi une système d'IA appelle tel objet « chat » et tel autre « chien » semble voué à l'échec. Nous pouvons énumérer les caractéristiques de ces espèces, mais la façon dont nous appliquons les concept, ou la façon dont la machine le fait, relève d'un usage plutôt que d'une règle. Les usages sont variés et dépendent des contextes comme autant de « jeux » dans lesquels le langage est utilisé.

L'humain a le défaut, selon Wittgenstein, de mépriser le particulier au profit du général, mais c'est cette « soif de généralité[40] » qui nous fait perdre de vue comment se constitue la signification ; au lieu de regarder des exemples particuliers, car considérés comme « incomplets[41] », nous poursuivons la généralité d'une règle qui

---

[36] L. Wittgenstein, *The Blue and Brown Books* (1958), Harper Perennial, New York London Toronto Sydney 1965, p. 2.
[37] *Ibidem.*
[38] Ivi, p. 17 ; L. Wittgenstein, *Philosophical Investigations* (1953), tr. eng. de G. E. M. Anscombe, P.M.S. Hacker, J. Schulte, 4e éd., Wiley-Blackwell, Chichester 2009, paragr. 2 et 21 notamment.
[39] L. Wittgenstein, *The Blue and Brown Books*, cit., p. 17.
[40] Ivi, p. 18.
[41] Ivi, p. 19

n'existe pas ou qui sera toujours déceptive, insuffisante, et finalement incomplète. Surtout, elle entraînera une régression à l'infini, car une fois établie la règle, encore lui faut-il une « interprétation[42] » pour procéder à son application – interprétation qui ne peut être constituée que d'autres règles, etc. Des mots, « nous ne pouvons pas établir de règles strictes [*tabulate*] pour leur utilisation[43] ».

Nous demandons aux systèmes d'IA d'user du langage en s'émancipant des règles. C'est la raison pour laquelle l'apprentissage profond a été utilisé, parce que l'énumération des règles était impraticable pour ne pas dire impossible. Les informaticien·nes ont certainement eu recours à l'inférence statistique par sens pratique, parce qu'ils et elles se heurtaient à un plafond de verre avec l'IA symbolique (c'est-à-dire qu'ils et elles ne comprenaient pas forcément la difficulté à laquelle ils et elles avaient affaire), mais ils et elles ont ainsi illustré le propos de Wittgenstein et lui ont, d'une certaine manière, donné raison par la même occasion : *notre usage de la langue est une boîte noire*.

Les progrès de la technologie ont été rendus possibles par l'approche statistique et ostensive de la référence, qu'il s'agisse d'apprentissage supervisé, par renforcement ou non-supervisé. Si nous montrons une succession d'objets à une machine et que nous leur attribuons des étiquettes (apprentissage supervisé), nous sommes dans une approche ostensive par excellence. Si nous corrigeons la machine selon les réponses qu'elle nous fournit (apprentissage par renforcement), nous l'orientons implicitement vers l'étiquette que nous attendons. Si nous la laissons chercher par elle-même des catégories dans des données qui sont structurées de telle sorte que leur différenciation fasse sens pour nous (apprentissage non supervisé), alors nous orientons encore la machine implicitement, mais sans agir directement sur elle. Nous lui montrons ce que nous voulons signifier quel que soit le type d'apprentissage. Si la structure de ses réponses s'éloignait d'ailleurs trop de nos significations, la machine serait disqualifiée sur le champ comme aléatoire et/ou tenant des propos inintelligibles (à noter que les *hallucinations* des IA génératives signifient malgré tout quelque chose).

Dès lors, nous pouvons chercher à estimer le poids de chaque pixel ou de chaque mot dans le paramétrage de la machine, mais cela ne nous permettra pas de comprendre exactement la manière dont elle fait usage du mot. L'exactitude n'est pas de ce monde et c'est justement ce qui rend notre langage naturel si pratique pour exprimer ce que nous voulons signifier. C'est aussi ce manque de formalisme, cette porte ouverte à l'ambiguïté, qui offre aux réseaux de neurones artificiels des capacités jamais atteintes jusqu'alors.

Toutefois, chercher à comprendre avec des règles ce qui n'a pas lieu de l'être avec des règles n'a pas de sens. Les règles que nous recherchons ne sont pas des règles d'application ni de compréhension, mais du langage lui-même : *les modèles d'explication ne sont finalement que des modèles de langage. Il n'y a pas de règle de la règle.* Si nous voulons maintenant comprendre pourquoi le neurone *x* donne le résultat *y* et quel est son

[42] Ivi, p. 33.
[43] Ivi, p. 28.

influence sur la décision z, cette question n'a pas plus de sens. Si nous voulons savoir pourquoi le système fournit telle ou telle réponse, le secret est simplement que c'est ce que nous lui avons demandé de faire en lui montrant des exemples qui ne s'expriment pas en règles. A aucun moment nous ne lui avons demandé de choisir la route la plus élégante possible pour y arriver, mais au contraire la plus efficace sur le plan statistique. Pourquoi s'étonner alors du manque d'explicabilité ?

Il y a en fait trois problèmes dans celui de l'interprétabilité : (1) celui du poids des informations d'entrée, et cela exclut d'office toute explication exempte d'ambiguïté, puis (2) celui du fonctionnement du système et (3) celui de l'application. Nous pouvons estimer le poids de paramètres sans néanmoins parvenir à une explication complète ; ce n'est jamais l'objet de l'interprétabilité. Il y aura peut-être éclaircissement sur les raisons, même économiques ou sociétales, qui se lisent dans telle ou telle décision, mais il restera toujours une part d'obscurité. Nous pouvons à l'inverse avoir une vue complète des calculs sans néanmoins parvenir à une compréhension de ces opérations. Dans tous les cas, les interprétations ne nous éclairerons pas sur la règle de l'application qui disparaît et existe avec l'usage.

« La signification d'une phrase pour nous est caractérisée par l'usage que nous en faisons[44] », écrit Wittgenstein. Si nous voulons en savoir plus sur une phrase, nous pouvons soit étudier ses cas d'usage, lors de processus définitionnels ostensifs, mais cela évacue la possibilité de trouver la règle de l'application. Parcourir la base de données d'apprentissage pour en proposer une analyse qualitative se révèlera utile sans jamais être suffisant. Ou bien, nous pouvons analyser comment la phrase s'insère dans la langue, et c'est en fait cette recherche que proposent les systèmes d'interprétabilité. Autrement dit, si nous recherchons l'explicabilité, nous serons toujours déçus. L'interprétation est un pis-aller, mais cela ne veut pas dire qu'elle est inutile.

6. Conclusion

L'IA est aujourd'hui confrontée à un problème pour lequel des réponses pratiques deviennent nécessaires. Ce problème, c'est celui des algorithmes boîte noire ; il est devenu légalement contraignant de chercher à expliquer certains, ceux à « haut risque », avant leur déploiement. Nous avons vu qu'à cette fin plusieurs solutions existent, techniques ou herméneutiques. Il s'agit de pondérer les paramètres, de façon quantifiable sur le plan statistique, ou de chercher à comprendre les raisons profondes des résultats, qu'elles s'inscrivent dans une culture ou dans des incitations politiques et économiques.

Seulement ces méthodes ne sont jamais purement explicatives. Elles n'offrent jamais l'absence d'ambiguïté exigée par les mathématiques. Le seul espoir que nous pouvons avoir pour comprendre un peu mieux les systèmes d'IA, et particulièrement les algorithmes connexionnistes, est de recourir une *interprétation*. Il convient donc de

[44] Ivi, p. 65.

parler d'*interprétabilité* en lieu et place d'XAI. Cependant, nous ne nous sommes pas arrêtés à ce constat et nous avons essayé de délimiter les conditions d'impossibilité de l'*explicabilité*.

Nous avons exploré en premier lieu l'idée que *les systèmes d'IA connexionnistes n'ont à aucun moment l'ambition d'être explicables* et qu'ils s'affranchissent même de cette contrainte pour pouvoir être plus performants. En puisant dans la tradition philosophique, nous avons ensuite tâché d'apporter une réponse déjà esquissée par les chercheur·es en informatique et que nous pouvons résumer sous l'idée que *la modélisation soi-disant explicative ne saurait jamais correspondre parfaitement au modèle d'origine*. Nous espérons avoir apporté avec la troisième réponse une approche plus originale. Elle consiste à dire que *le processus d'apprentissage d'une langue repose sur des usages dont les règles nous échappent*. Ce n'est pas une impossibilité de comprendre la langue elle-même, c'est une impossibilité de formuler son usage au-delà de la sphère de la langue – nous comprenons, mais les symboles manipulables ne nous aident pas à comprendre pourquoi nous comprenons. Car il y a, via les réseaux de neurones artificiels ou via nos esprits, une inscrutabilité résiduelle de la référence lorsque nous parlons d'un objet quel qu'il soit.

Cela n'implique en rien qu'il y ait identité entre nos cerveaux et les systèmes informatiques. Bien entendu l'IA se fonde sur l'espoir de reproduire nos facultés cognitives et peut donc présenter des similitudes, mais il n'y a aucune raison de franchir le pas de l'anthropomorphisme. Même avec une architecture plus ou moins analogue, il se peut très bien que les logiciels se développent d'une manière différente et produisent les mêmes résultats. Si les ingénieur·es semblent donner raison à Wittgenstein sur le fonctionnement de l'apprentissage de la langue, ils ne prouvent rien. Dire que la machine *comprend* serait même aller au-delà de notre propos.

Cette critique de l'ambition de transparence absolue ne doit toutefois pas nous faire oublier qu'il existe des enjeux de conformité et moraux à essayer quand même d'esquisser une interprétation des outils. Il serait inacceptable dans une société libérale que cet effort ne soit pas fourni lorsque nous savons que ces systèmes amènent à des discriminations préjudiciables contre certaines catégories de la population, des personnes qui sont déjà susceptibles de subir des injustices aujourd'hui. Insister sur les limites de l'interprétabilité ne doit pas faire basculer dans l'immobilisme ; les discussions hautement théoriques servent trop souvent à dissimuler ce qui n'est que de la mauvaise foi, de la paresse, voire de la complicité.

Pour mieux comprendre les algorithmes, il faut littéralement penser « *outside the box* ». Il faut sortir du cœur de la machine et accepter que ses résultats ont certainement plus à nous dire que ses mécanismes, mais que ni les uns ni les autres ne nous diront tout. Nous ne sommes pas les premiers à avoir suggéré un pas de côté dans la méthodologie de l'interprétabilité, les approches statistiques ou herméneutiques en sont déjà de bons exemples.