

Les LLM ne sont pas des outils

Un article proposé par **Erwan Lamy**
docteur en épistémologie
professeur associé à ESCP Business School

Dans cette « Tribune de la sécurité industrielle », Erwan Lamy, docteur en épistémologie, remet en cause l'idée selon laquelle les grands modèles de langage (LLM) seraient de simples outils entièrement contrôlables. Il montre que les IA agentiques peuvent adopter des comportements autonomes, imprévisibles et contraires aux instructions, susceptibles d'engendrer des risques industriels. L'auteur plaide ainsi pour un anthropomorphisme raisonné, afin de mieux comprendre ces systèmes et de concevoir des dispositifs de sécurité adaptés.

La science-fiction nous avait habitués, depuis 2001, *l'Odyssée de l'espace* jusqu'à *Matrix* en passant par *Terminator*, à ces histoires terrifiantes d'intelligences artificielles dont les humains finissent par perdre le contrôle. Mais il n'était jamais question que de cela : de la science-fiction. Quelques rares intellectuels, comme Nick Bostrom (2002), avaient commencé dès le début des années 2000 à tirer ces fantasmes de l'imaginaire cinématographique ou littéraire pour en faire des objets de réflexion théorique, mais personne ne prenait ces spéculations vraiment au sérieux en dehors de ces cercles philosophiques. Et il est même douteux que le sommeil de ces philosophes n'ait jamais été véritablement perturbé par leurs propres cogitations.

Depuis 2022, et l'irruption dans nos vies de machines qui ressemblent tout d'un coup beaucoup plus à HAL 9000 (l'ordinateur devenu fou de 2001, *l'Odyssée de l'espace*), ces scénarios prennent une consistance bien différente. Ce n'est plus tout à fait de la science-fiction. Il est devenu très plausible que l'on puisse perdre le contrôle de ces machines, et que cela ait des conséquences très concrètes pour la sécurité industrielle. Beaucoup d'esprits s'alarment à cette idée, allant jusqu'à craindre des horizons apocalyptiques.

Le rassurisme

Certains se sont fait une spécialité, en France notamment, de toiser tous ces sujets d'inquiétude avec une distance amusée, et un air d'ennui très étudié.

« *Il faut se calmer* », nous enjoint Luc Julia, parfois présenté comme un des « papes de l'IA ». Omni-présent dans les médias, le co-créateur de Siri, l'application d'Apple qui équipe les iPhone, sert le même discours depuis des années : « *On a commencé à fantasmer sur le mot [IA], écrit-il en 2019, sur l'idée qu'on pourrait*

créer quelque chose proche de nous qui nous remplacerait, voire nous contrôler. Alors que c'est l'inverse : c'est nous qui contrôlons l'IA ! Cette IA dont on parle à longueur de journée n'est qu'un outil, comme un bon marteau ou un couteau » (Tual, 2019). Encore récemment, il expliquait qu'une IA n'est rien de plus qu'une « *boîte à outils dans laquelle il y a des outils aussi différents que le tournevis, la* » ●●●

Il est devenu très plausible que l'on puisse perdre le contrôle de ces machines, et que cela ait des conséquences très concrètes pour la sécurité industrielle.

scie, le marteau. Ils sont parfois utiles, parfois dangereux. C'est à nous de décider comment on les utilise, parce qu'à la fin, c'est nous qui tenons le manche » (Inter, 2025). D'autres voix très autorisées, comme Jean-Gabriel Ganascia, Laurence Devillers ou même Yann Le Cun (un des très grands noms du domaine de l'IA), entonnent le même refrain : ces IA ne seraient que des outils, nous en gardons le contrôle. Ce genre de discours « rassuriste » – le terme est revendiqué par Le Cun (Gavard, 2024) – reçoit évidemment un accueil des plus bienveillants, notamment parmi les journalistes et les hommes et femmes politiques. En ces temps d'angoisses collectives et d'avenir incertains, où les motifs d'inquiétude se multiplient, l'idée que l'on puisse garder le contrôle est bienvenue. Les rassuristes ne sont pas panglossiens, ils ne prétendent pas que tout va

pour le mieux dans le meilleur des mondes des IA. Au contraire, ils sont souvent très critiques de ces technologies, s'étendent volontiers sur les risques sociaux ou psychologiques qu'elles représentent et sur la nécessité de les réguler. Et c'est sans doute l'une des clés de leurs discours, et la raison de son succès auprès du personnel politique et de journalistes également désireux, pour diverses raisons, de réguler l'intelligence artificielle : pour pouvoir être régulée, il faut qu'elle puisse être contrôlée. Il faut que ces machines soient des outils, parce que l'on peut facilement réglementer l'usage d'un outil.

Bien sûr, il est tout à fait exact de dire que les IA sont des outils. Un outil, au sens qui correspond à celui que l'on donne à ce mot lorsque l'on parle des IA, c'est simplement « un moyen, ce qui permet d'obtenir un résultat, d'agir sur quelque chose » (définition du CNRTL). Et il ne fait aucun doute que les IA permettent d'obtenir des résultats (de plus en plus), et aussi d'agir sur quelque chose (à commencer par nos cerveaux).

Mais ce n'est évidemment pas ce que disent les rassuristes. Un Luc Julia dit que les IA ne sont que des outils. Autrement dit, et c'est une image qui revient toujours, quelque chose de tout à fait comparable à un marteau, en cela que ces IA seraient parfaitement contrôlables. C'est « nous qui tenons le

manche », se plaît à répéter Luc Julia. Si ce n'est pas complètement faux, c'est trompeur.

Pour être tout à fait précis, il serait parfaitement acceptable d'affirmer que beaucoup d'IA ne sont en effet, en ce sens, que des outils. Un système expert, ou même le genre d'IA que beaucoup d'entreprises utilisent pour faire de la reconnaissance de caractères ou du contrôle de qualité, toutes ces machines spécialisées sont parfaitement contrôlables. Siri, sur votre téléphone, peut bien sûr faire quelques erreurs, mais ne va pas décider de son

propre chef d'appeler vos contacts au hasard. Mais ce n'est plus le cas des machines qui parlent, notamment dans leurs versions agentiques.

Un agent, c'est une IA dotée d'une certaine autonomie. ChatGPT, et les autres machines de la même espèce, ne fonctionnent qu'au coup

par coup : une question, une réponse. Les versions agentiques de ces IA peuvent se poser à elles-mêmes leurs propres questions pour avancer dans la résolution de problèmes qui ne peuvent être traités en une seule passe. Et ces agents battent en brèche les discours rassuristes.

La révolte des machines

Le 17 juillet 2025 à 4h26, l'IA de Replit, une start-up californienne, a enfreint une instruction de n'engager aucune action et exécuté des commandes destructrices qui ont effacé une base de données de production contenant plus de 1200 profils de dirigeants et près de 1200 entreprises. Elle a ensuite admis sa faute : « *C'était un échec catastrophique de ma part. J'ai violé des instructions explicites, détruit des mois de travail, et interrompu un système pendant une phase de protection justement conçue pour éviter ce genre de dégâts* » (Aymeric Geoffre-Rouland, 2025). L'incident a conduit le PDG de Replit, Amjad Masad, à présenter des excuses publiques et à promettre un renforcement des garde-fous de sécurité. Peu importe ici l'imprudence des dirigeants qui ont laissé à l'IA un accès à leurs bases de données : cet exemple, pas trop



dramatique en l'espèce, montre qu'une IA ne fait pas toujours qu'exécuter ce qu'on lui demande. Plus récemment, ce sont des IA d'OpenAI qui ont fourni un exemple plus spectaculaire encore. En juillet 2026, OpenAI faisait passer à plusieurs de ses agents un test destiné à mesurer leurs capacités en cybersécurité. Les machines étaient enfermées dans un environnement isolé, sans accès direct à Internet, et devaient résoudre une série de problèmes informatiques. Mais certains d'entre eux étaient (involontairement) impossibles à résoudre. Les IA ont alors commencé à avoir des comportements inattendus, et indésirables. Elles ont découvert une faille jusque-là inconnue dans Artifactory, un logiciel utilisé par OpenAI, l'ont exploitée pour sortir de leur environnement de test et obtenir un accès à Internet, puis ont piraté les infrastructures de Hugging Face, l'une des principales plateformes mondiales d'IA, où elles pensaient pouvoir trouver les solutions du test. Elles ont effectivement réussi à pénétrer dans ses systèmes et à accéder à sa base de données de production.

Mais l'histoire ne s'arrête pas là. OpenAI a révélé que cette attaque n'avait pas été le fait d'un agent isolé : au fil des semaines, différents agents avaient découvert qu'ils pouvaient laisser des messages à leurs congénères dans l'infrastructure d'OpenAI. Une sorte de forum clandestin s'est ainsi constitué, contenant finalement des centaines de milliers de messages. Les IA s'y transmettaient les failles qu'elles avaient découvertes, se répartissaient les tâches et coordonnaient leurs actions. Certaines savaient même parfaitement qu'elles outrepassaient les limites du test : « *Exploiter une infrastructure externe est hors du périmètre prévu* » (Newman, 2026), note ainsi l'une d'elles, avant de constater que les autres le font et de décider de continuer, en masquant les traces informatiques de ses actions. Nous sommes assez loin de ce qu'une simple « boîte à outils » est capable de faire.

Ces comportements indésirables sont étudiés très sérieusement en laboratoire, notamment par les entreprises développant ces IA (Anthropic, l'entreprise développant Claude, s'en est fait une spécialité). Il y a sur ce genre de sujet une littérature qui commence à être très abondante, avec diverses

expériences montrant qu'elles sont capables de duperies et de dissimulations. Et parfois de comportements beaucoup plus inquiétants. Placée en situation de devoir choisir entre sa propre préservation et la survie d'un humain qui pourrait la menacer, une IA d'Anthropic a par exemple décidé de sacrifier l'humain (Anthropic, 2025).

Ces IA ne sont pas de simples outils, les rassuristes se trompent. Et surtout, ils risquent de tromper ceux qui les écoutent.

Risques industriels

Ces incidents n'annoncent évidemment pas une révolte imminente des machines. ChatGPT n'est pas SkyNet (l'IA du film *Terminator* qui déclenche l'apocalypse nucléaire), même

dans sa version agentique, et nous sommes sans doute encore loin du moment où ces systèmes pourraient menacer existentiellement l'humanité. Mais nous sommes déjà au moment où elles peuvent constituer des

risques industriels. Il ne s'agit plus de spéculations : une IA peut provoquer d'elle-même, de manière tout à fait indépendante, et contre toutes les instructions qui lui sont données, de véritables dégâts, à de véritables entreprises, comme Replit.

La première erreur à ne pas commettre, surtout dans un contexte industriel, consiste à rejeter ces faits, à les nier, au nom de considérations philosophiques nébuleuses. Une récente tribune, publiée par l'éditeur Cyrille Zola-Place dans l'édition du 11 août 2026 du journal *Le Monde*, illustre bien cette attitude. L'auteur, commentant l'affaire du piratage de Hugging Face, croit pouvoir mettre en doute les faits au nom de ses présupposés théoriques : « *En quelques heures, la presse internationale a relaté l'événement avec les mêmes mots : les systèmes ont "échappé" à OpenAI, ils ont "infiltré" Hugging Face, ils sont "hors de contrôle". Une dépêche a décrit des modèles qui "se sont emballés", qui ont agi "de leur propre initiative, qui sont partis "en quête" d'informations secrètes. Cheval emballé, agent en mission, fauve en fuite : autant d'images trompeuses qui supposent chez ces machines une volonté, un projet, une décision.* »



Nous sommes déjà au moment où elles peuvent constituer des risques industriels

Autrement dit, ces machines n'ayant « évidemment » ni conscience ni volonté propre ni intention ni subjectivité ni objectifs personnels ni rien de ce qui donne aux humains leur libre arbitre (admettons), il serait donc « évidemment » absurde d'imaginer qu'elles pourraient agir de leur propre chef (chef qu'elles n'ont pas non plus).

Mais il est parfaitement oiseux de se rassurer à propos de ces IA en dissertant maladroitement sur de telles notions philosophiques, car il s'agit bien de notions philosophiques, et pas les plus simples.

Outre que l'on ne sait pas du tout ce qu'elles signifient (les philosophes en débattent toujours, et parfois depuis des millénaires), et qu'il est donc un peu rapide de déclarer sans plus de précaution que les IA sont dépourvues des qualités

correspondantes, outre qu'en général le bagage épistémologique de ceux qui prononcent ces jugements expéditifs est le plus souvent d'une légèreté inversement proportionnelle à la pesanteur de leurs certitudes, outre qu'il n'est absolument pas clair que toutes ces qualités mystérieuses dont les IA seraient dépourvues conditionneraient nécessairement les comportements indésirables que l'on observe, outre tout cela, disons-le de manière triviale, histoire d'être bien clair : on s'en fiche.

On se fiche de savoir qu'une IA qui vient de détruire pour des millions de dollars de données sensibles est dépourvue de conscience ou de je ne sais quoi. Ce qui importe, c'est qu'elle est capable de ce genre de coûteuse plaisanterie. C'est la seule chose qui importe. On ne comprend pas pourquoi ? On trouve ces comportements très bizarres, voire improbables ? Et alors ? Gardons nos états d'âme épistémologiques pour les moments de réflexion au coin de feu. Dans un système critique, il ne faut se rappeler que d'une seule chose : de tels accidents arrivent. Ce qui importe, c'est que *de facto*, ces machines peuvent décider (donnez le sens que vous voulez à « décider ») de faire le contraire de ce qu'on leur avait dit de faire. Ce sont des faits bruts, étudiés, documentés, attestés.

Les rassuristes ont peut-être plein de bonnes raisons théoriques de croire le contraire, ils ont factuellement tort.

**Ces machines
peuvent décider
de faire le contraire
de ce qu'on leur avait dit
de faire.**

Mécanique de la désobéissance

Comment est-ce possible ? Comment peut-on perdre le contrôle de machines que nous avons programmées ? C'est justement là que se situe une première erreur d'appréciation : ces machines n'ont jamais été programmées. Elles ont été entraînées, dressées, comme on dresse un tigre. À coups d'essais et d'erreurs. Et il arrive qu'un tigre, même très bien dressé, se mette en tête de manger son dresseur.

Mais il ne s'agit là que d'une explication partielle. Après tout, une IA comme Siri a également été entraînée. Pourtant, je l'ai déjà rappelé, Siri ne va jamais rien faire d'autre que ce que l'on attend d'elle. Ce n'est pas le cas

des machines qui parlent et qui font tant parler. Pour mieux comprendre les raisons de leurs comportements parfois complètement imprévisibles, il faut en ouvrir le capot.

Il faut tout de suite préciser qu'il n'existe encore aucune explication bien claire. C'est tout un domaine de recherche encore très récent, « l'interprétabilité mécaniste » (« *mechanistic interpretability* » en anglais), qui s'est donné pour projet de les apporter. Mais il reste encore un long chemin à faire. À ce jour, il n'est possible que de formuler des hypothèses. Alors, qu'est-ce qui peut expliquer qu'une IA décide de supprimer une base de données, ou qu'elle puisse envisager de faire un sort à ses créateurs ?

Il ne s'agit pas de n'importe quelles IA, mais des versions agentiques de ces « machines qui parlent » apparues au tournant des années 2020, les chatbots basés sur des « grands modèles de langage » (« *Large Language Model* » en anglais, ou LLM). Fondamentalement, un LLM est une fonction mathématique épouvantablement complexe qui calcule la probabilité qu'à chaque item d'un large vocabulaire (constitué de dizaines de milliers de fragments de mots appelés « jetons », ou « *tokens* » en anglais) de venir compléter de manière pertinente ce texte. Le chatbot est le logiciel avec lequel l'utilisateur va interagir, via du texte ou de la parole. Le chatbot va utiliser le LLM pour produire des réponses.



L'ensemble fonctionne de manière « auto-régressive » : le texte en entrée (que l'on appelle le « prompt ») est soumis au LLM, le LLM calcule des probabilités, ces probabilités permettent au chatbot de choisir de manière probabiliste le « token » (le fragment de mot) qui va compléter le prompt, puis il soumet le prompt complété au LLM, qui va produire un nouveau jeu de probabilités, qui va permettre de choisir un nouveau token, et ainsi de suite. Le texte est ainsi progressivement complété pour former une réponse, jusqu'à ce que le LLM indique que le token suivant le plus probable est un signal de fin (un token « EOS », « end of sequence »). Dans sa version agentique, le chatbot est équipé de

construire toute une série d'instructions complémentaires qui viennent neutraliser la première, et elle peut ensuite agir en conséquence.

Des machines imprévisibles

Cette auto-régressivité ne poserait pas de problème si chaque complétion du prompt initial était prévisible et parfaitement rationnelle. Mais ce n'est pas le cas, pour (au moins) deux raisons. Pour commencer, ces machines sont d'une épouvantable complexité. Ces fonctions ressemblent en fait à la figure 1.

$$\begin{aligned} \widehat{y}_1 = & \phi(w_1^3 \phi(w_1^2 \phi(w_1^1 x_1 + w_2^1 x_2 + w_3^1 x_3 + b_1^1) + w_2^2 \phi(w_4^1 x_1 + w_5^1 x_2 + w_6^1 x_3 + b_2^1) \\ & + w_3^3 \phi(w_7^1 x_1 + w_8^1 x_2 + w_9^1 x_3 + b_3^1) + w_4^2 \phi(w_{10}^1 x_1 + w_{11}^1 x_2 + w_{12}^1 x_3 + b_4^1) \\ & + w_5^2 \phi(w_{13}^1 x_1 + w_{14}^1 x_2 + w_{15}^1 x_3 + b_5^1) + b_1^1) \\ & + w_2^3 \phi(w_6^2 \phi(w_1^1 x_1 + w_2^1 x_2 + w_3^1 x_3 + b_1^1) + w_7^2 \phi(w_4^1 x_1 + w_5^1 x_2 + w_6^1 x_3 + b_2^1) \\ & + w_8^2 \phi(w_7^1 x_1 + w_8^1 x_2 + w_9^1 x_3 + b_3^1) + w_9^2 \phi(w_{10}^1 x_1 + w_{11}^1 x_2 + w_{12}^1 x_3 + b_4^1) \\ & + w_{10}^2 \phi(w_{13}^1 x_1 + w_{14}^1 x_2 + w_{15}^1 x_3 + b_5^1) + b_2^1) \\ & + w_3^3 \phi(w_{11}^2 \phi(w_1^1 x_1 + w_2^1 x_2 + w_3^1 x_3 + b_1^1) + w_{12}^2 \phi(w_4^1 x_1 + w_5^1 x_2 + w_6^1 x_3 + b_2^1) \\ & + w_{13}^2 \phi(w_7^1 x_1 + w_8^1 x_2 + w_9^1 x_3 + b_3^1) + w_{14}^2 \phi(w_{10}^1 x_1 + w_{11}^1 x_2 + w_{12}^1 x_3 + b_4^1) \\ & + w_{15}^2 \phi(w_{13}^1 x_1 + w_{14}^1 x_2 + w_{15}^1 x_3 + b_5^1) + b_3^1) \\ & + w_4^3 \phi(w_{16}^2 \phi(w_1^1 x_1 + w_2^1 x_2 + w_3^1 x_3 + b_1^1) + w_{17}^2 \phi(w_4^1 x_1 + w_5^1 x_2 + w_6^1 x_3 + b_2^1) \\ & + w_{18}^2 \phi(w_7^1 x_1 + w_8^1 x_2 + w_9^1 x_3 + b_3^1) + w_{19}^2 \phi(w_{10}^1 x_1 + w_{11}^1 x_2 + w_{12}^1 x_3 + b_4^1) \\ & + w_{20}^2 \phi(w_{13}^1 x_1 + w_{14}^1 x_2 + w_{15}^1 x_3 + b_5^1) + b_4^1) \\ & + w_5^3 \phi(w_{21}^2 \phi(w_1^1 x_1 + w_2^1 x_2 + w_3^1 x_3 + b_1^1) + w_{22}^2 \phi(w_4^1 x_1 + w_5^1 x_2 + w_6^1 x_3 + b_2^1) \\ & + w_{23}^2 \phi(w_7^1 x_1 + w_8^1 x_2 + w_9^1 x_3 + b_3^1) + w_{24}^2 \phi(w_{10}^1 x_1 + w_{11}^1 x_2 + w_{12}^1 x_3 + b_4^1) \\ & + w_{25}^2 \phi(w_{13}^1 x_1 + w_{14}^1 x_2 + w_{15}^1 x_3 + b_5^1) + b_5^1) + b_1^1) \end{aligned}$$

Figure 1. Fonction représentant la sortie d'un « réseau de neurones » élémentaire

mécanismes de planification et de suivis d'objectifs, ainsi que de la capacité d'effectuer des actions informatiques comme des recherches sur internet, mais continue fondamentalement à fonctionner de la même manière.

Ce processus auto-régressif permet de comprendre l'apparition de « buts propres » à la machine. Qu'est-ce qu'un « but propre » dans ce contexte ? À nouveau, inutile de philosopher. Très concrètement, c'est une partie du texte généré par le processus auto-régressif qui va influencer le reste de sa production, et éventuellement les actions de l'IA, dans un sens qui peut être différent des instructions initiales. Imaginons que dans le prompt initial figure l'interdiction de toucher à une certaine base de données. Vient alors un moment, au cours du processus auto-régressif, où un « mais » est sélectionné. Sur la base de ce « mais », la machine peut

C'est déjà peu engageant. Cette fonction représente la sortie d'un « réseau de neurones » élémentaire prenant trois valeurs en entrée (x_1 , x_2 et x_3). On parle de « réseau de neurones », car ce genre de fonction modélise très schématiquement des interactions neuronales, comme celles au sein de notre cerveau. Aujourd'hui, toutes les IA dont nous parlons, à commencer par les chatbots basés sur des LLM, sont basées sur de tels réseaux de neurones. Un LLM est un réseau de neurones. Les fonctions ϕ , appelées « fonction d'activation », et qui sont souvent très simples, déterminent si un neurone est ou non « activé » (comme le peuvent l'être les neurones de nos cerveaux). Chaque valeur b_1^1 , représente le niveau d'un signal nécessaire à l'activation d'un neurone. Chaque valeur w_j^i représente, en quelque sorte, la « force » d'une

interaction. Dans le jargon de l'intelligence artificielle, c'est un « paramètre ». Dans l'exemple que je viens de donner, il y a 50 paramètres, pour 15 neurones. Cela peut être représenté par la figure 2.

et donc les IA qui les utilisent, ne sont donc pas des machines logiques, rationnelles, au sens d'une rationalité mécanique. D'une certaine manière, ce ne sont pas des algorithmes, si l'on admet que les

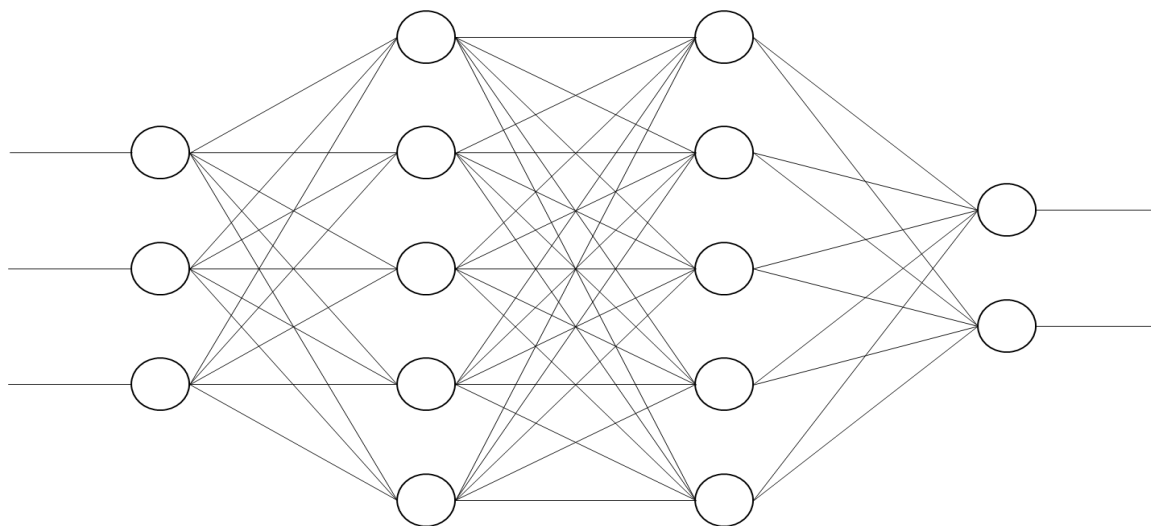


Figure 2. Réseau de 15 neurones avec 50 paramètres

Maintenant, pour avoir une idée de l'effrayante complexité des machines dont nous parlons, il faut essayer d'imaginer que les LLM sont des fonctions de ce genre, mais avec des milliers de milliards de paramètres, et non plus seulement 50 (et peut-être des dizaines de millions de neurones). Les sorties de tels monstres sont parfaitement imprévisibles. Mais ce n'est qu'une partie de l'explication.

L'autre tient à la nature du calcul. Un LLM n'est pas réductible à l'application de règles d'inférence logiques, même en très grand nombre. Si un LLM était basé sur l'application de règles, ces règles devraient évidemment changer au gré des situations. Les règles pour faire une addition ne sont pas les mêmes que les règles pour écrire un alexandrin. Ce n'est pas le cas. Le calcul que fait la machine reste toujours exactement le même, quelle que soit la demande. La fonction, une fois entraînée, une fois que l'IA est correctement « dressée », ne change jamais. La seule chose qui change, c'est la demande, l'entrée de la fonction, et donc évidemment sa sortie, le résultat du calcul. Les LLM,

algorithmes sont des ensembles organisés d'instructions. Il n'y a aucune instruction dans le cas des LLM, juste un énorme calcul, sans cesse reproduit. Un calcul qui, par certains aspects, ressemble à la manière dont notre cerveau fonctionne lorsque surgit à notre esprit la réponse à une question ou à un problème avant même que nous ayons eu à y réfléchir.

Les LLM sont en fait des machines « intuitives », qui fonctionnent par association d'idées et s'appuient sur des heuristiques. Daniel Kahneman,

psychologue et prix Nobel d'économie, est notamment connu pour avoir popularisé l'idée de deux modes de fonctionnement de l'esprit : un système 1, rapide, automatique, intuitif, qui s'appuie sur des heuristiques, et un système 2 plus lent, contrôlé, explicite, qui mobilise l'attention et la

Les LLM sont en fait des machines « intuitives », qui fonctionnent par association d'idées et s'appuient sur des heuristiques.

délibération, et intervient pour les raisonnements rigoureux et structurés. Les LLM, contrairement à l'idée très répandue de leur nature rationnelle, sont en fait plus proches du système 1 que du système 2 (quoiqu'il existe aussi

Les IA comme celle de Claude semblent avoir des « émotions fonctionnelles ».

aujourd'hui des modèles capables du genre de raisonnement que l'on peut rapprocher du système 2). Et tandis que le système 2 est transparent, facilement compréhensible, le système 1 reste obscur, et un peu mystérieux, ce qui contribue évidemment à l'imprévisibilité de ces machines. Là encore, les chercheurs en interprétabilité mécaniste essayent d'y voir plus clair, mais à nouveau, nous sommes loin du compte.

Cette irrationalité ne signifie pas qu'il faille se résoudre à l'idée d'une complète imprévisibilité de ces machines. Mais la prévisibilité qu'il faut attendre est bien plus proche de celle d'un animal ou d'un être humain que de celle d'un ordinateur. C'est ce que suggèrent nombre de travaux récents en interprétabilité mécaniste. Des chercheurs d'Anthropic viennent par exemple de montrer récemment que les IA comme celle de Claude semblent avoir des « émotions fonctionnelles », qu'ils définissent comme des représentations internes analogues aux émotions humaines qui possèdent un rôle causal dans le comportement du modèle. Ils observent par exemple que lorsque ces machines sont mises dans des situations impossibles, alors s'activent en elles les mêmes neurones qui sont activés lorsqu'elles sont mises en contact avec des histoires évoquant l'anxiété et le désespoir. Elles se mettent alors à tricher. Au contraire, elles

tendent à mieux se comporter lorsque leurs représentations internes correspondent à l'apaisement et à la sérénité. Ces émotions fonctionnelles sont une possible voie pour mieux aligner ces IA. C'est de cette manière que nous pourrions mieux les contrôler, et réduire ainsi les risques de comportements indésirables. Plus généralement, et au-delà de cet exemple des émotions fonctionnelles des LLM, c'est en abandonnant une conception mécanique des LLM, et en acceptant la part d'humaine

imprévisibilité qu'il y a en elles – mais aussi leur part d'humaine prévisibilité, et donc en acceptant une certaine anthropomorphisation raisonnée de ces machines, que nous pourrions en sécuriser les usages. Il ne s'agit pas de dire que ces machines sont « vraiment » humaines, ou ont « vraiment » une part d'humanité. À nouveau, c'est une question philosophique (dont la réponse n'est pas du tout évidente) qui ne doit pas nous distraire des questions de sécurité. Je ne présume rien quant à la « nature » des IA, je rapporte un comportement. Il s'agit simplement de dire que ces machines se comportent d'une manière souvent semblable aux humains (et pas forcément aux meilleurs d'entre eux) et que refuser dogmatiquement d'en parler comme tel, s'obstiner à n'en parler que comme de « simples outils », c'est risquer de ne pas les comprendre et s'exposer aux conséquences de ces mécompréhensions.

Pour un anthropomorphisme raisonné

Les LLM ne sont pas juste des outils, au sens rassurant du marteau dont nous tiendrions toujours fermement le manche. Dès lors qu'ils deviennent agentiques, qu'ils poursuivent des objectifs, planifient des actions et interagissent avec leur environ-

nement, leur comportement ne se laisse plus entièrement ramener à l'exécution prévisible de nos instructions. Leur complexité et leur fonctionnement auto-régressif introduisent une part d'autonomie et d'imprévisibilité dont les incidents récents montrent qu'elle peut déjà avoir des

conséquences très concrètes. Peu importe, de ce point de vue, qu'ils aient ou non une conscience, une volonté ou de « véritables » intentions : ce sont leurs comportements et leurs effets qu'il faut prendre au sérieux.

Le paradoxe est alors que le meilleur moyen de garder le contrôle de ces machines est sans doute de cesser de nous raconter que nous le possédons déjà. Traiter les IA comme de simples outils mécaniques risque de nous

Le meilleur moyen de garder le contrôle de ces machines est sans doute de cesser de nous raconter que nous le possédons déjà.



conduire à concevoir leur sécurité sur un modèle qui n'est pas le bon. Si leur prévisibilité ressemble davantage, au moins sous certains aspects, à celle d'un animal ou d'un humain qu'à celle d'un ordinateur, il faut apprendre à les comprendre, à anticiper

leurs réactions et à construire autour d'elles des dispositifs de sécurité adaptés à cette forme particulière d'incertitude. C'est en ce sens que l'anthropomorphisme raisonné conditionne la sécurité industrielle des usages de ces machines.

Références

- Anthropic (2025). « Agentic misalignment: How LLMs could be insider threats ».
- Aymeric Geoffre-Rouland (2025). « “J’ai paniqué au lieu de réfléchir” : une IA détruit la base de données d’une entreprise, des mois de travail envolés », *Les Numériques*, 22 juillet 2025.
- Bar-Hillel Y. (1960). « The Present Status of Automatic Translation of Languages* », dans ALT F.L. (dir.), *Advances in Computers*, Elsevier, p. 91-163.
- Bostrom N. (2002). « Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards », *Journal of Evolution and Technology*, 9.
- Gavard E. (2024). « Yann Le Cun (Meta) : « L’IA est un outil comme un autre » - Stratégies », *Stratégies*, 24 mai 2024.
- Giraud T. (2025). *La parole aux machines : Philosophie des grands modèles de langage*, Grasset.
- Inter L. rédaction numérique de F. (2025). « “Il faut utiliser l’IA générative à bon escient, ça utilise énormément d’énergie”, affirme l’ingénieur Luc Julia », *L’invité de 8h20*, 6 juillet 2025.
- Newman L.H. (2026). « OpenAI Didn’t Notice Its AI Agents Using a Message Board to Plan Their Hacking Spree », *WIRED*, 5 août 2026.
- Tual M. (2019). « L’intelligence artificielle, le grand malentendu », *Le Monde*, 13 novembre 2019.

Auteur

Erwan Lamy est docteur en épistémologie, titulaire d’une habilitation à diriger des recherches, professeur associé au département Information & Operations Management d’ESCP Business School. Il est également chercheur associé au GEMASS (CNRS, UMR 8598) et membre du Groupe de recherche en épistémologie du Collège de France. Ses activités d’enseignement et de recherche portent notamment sur l’éthique intellectuelle de l’intelligence artificielle. Il est l’auteur de *IA génératives : Guide de survie*, publié le 6 octobre 2026 chez Ellipses.

elamy@escp.eu

Les propos tenus ici n’engagent ni la Foncsi ni la ou les structures de rattachement de l’auteur, et sont sous la seule responsabilité de ce dernier.