

Séance solennelle du 29 janvier 2024

Penser l'Intelligence Artificielle (les agents conversationnels)

Michel CHEIN

Académie des Sciences et Lettres de Montpellier
Professeur émérite de l'Université de Montpellier

La discipline scientifique appelée « Intelligence Artificielle » a été créée au début des années 50 par une vingtaine de personnes ; par les créateurs de l'informatique Shannon, Turing et von Neumann, des mathématiciens, des physiciens, des ingénieurs, et un économiste.

L'article fondateur de l'Intelligence Artificielle (IA) est un article publié par Alan Turing en 1950. Dans cet article il remplaçait la question une machine peut-elle penser ? – « Qui ne mérite pas une discussion », écrivait-il, – par la question suivante : « peut-on écrire des programmes ayant un comportement que les humains jugent intelligents, comme jouer aux échecs ou apprendre une langue ? »

C'est ce programme de recherche en informatique qui fut discuté lors du fameux séminaire de Dartmouth en 1956 et que McCarthy proposa d'appeler Intelligence Artificielle et ce malgré les réticences de Shannon, Newell, Simon et de bien d'autres. Il aurait proposé ce nom pour s'opposer au terme Automata theory, titre d'un livre de Shannon et de lui-même, pour ne pas se limiter à l'étude mathématique de ces automates mais à leurs relations avec le langage, l'intelligence humaine, la possibilité de jouer à des jeux, etc.

Le prix le plus prestigieux en informatique, le prix Turing créé en 1966, a été attribué depuis sa création à de nombreux chercheurs éminents en IA, le premier étant Minsky en 1969, puis McCarthy, Newell, Simon, etc. et les derniers étant Hinton, Benjio, et LeCun 2018 pour, je cite, « les percées conceptuelles et techniques qui ont fait des réseaux neuronaux profonds un élément essentiel de l'informatique. »

La locution « intelligence artificielle » désigne un domaine de l'informatique, et si ce choix a été génial d'un point de vue marketing – il est assurément plus sexy que Information Processing System – c'est une malédiction pour les chercheurs travaillant dans ce domaine ! Je n'ai aucune compétence particulière pour parler d'intelligence. Je ne suis ni philosophe, ni psychologue, ni neuroscientifique ni spécialiste de nombreux autres domaines mais je peux parler d'informatique en particulier de l'IA qui est un domaine de la science informatique, et des IAs qui ne sont que des programmes informatiques. Comme d'autres, je suis réticent, aujourd'hui, à considérer IA comme un acronyme ou alors que ce soit pour Informatique Avancée !

Depuis son origine l'IA est composée de deux grands sous-domaines.

- L'IA symbolique dont le noyau est la modélisation informatique de connaissances et de raisonnements qui consiste à représenter par des objets informatiques, des connaissances et des données, et à construire des procédures de manipulation de ces objets – c'est à dire des programmes – qui sont des modèles informatiques de différents

types de raisonnement (déduction, induction, abduction, analogie,...). La logique mathématique, ou plutôt les logiques computationnelles sont au cœur de ce domaine.

- Le deuxième grand sous-domaine est l'IA numérique dont le cœur est l'apprentissage statistique, en particulier l'apprentissage par réseaux de neurones artificiels profonds (*deep learning*).

On peut grossièrement distinguer 5 grandes étapes pour construire une application basée sur une méthode d'apprentissage statistique :

- Tout d'abord définir l'objectif de l'application, par exemple reconnaître l'espèce d'un champignon à partir d'une photo.
- Puis choisir une méthode, par exemple l'apprentissage supervisé par un réseau de neurones qui aura pour données des pixels d'images et en sortie des espèces de champignon.
- Ensuite acquérir des données d'apprentissage et les analyser : les données sont-elles pertinentes, c'est-à-dire représentatives du but.
- La quatrième étape est précisément celle de l'apprentissage du réseau. On donne en entrée des exemples de champignons, des valeurs de pixels d'une photo de champignon, on modifie les paramètres jusqu'à ce que la sortie soit la bonne espèce du champignon. La phase d'apprentissage peut nécessiter des moyens de calcul considérables car on peut avoir des milliards de paramètres à évaluer. Cette étape est terminée lorsque le réseau donne des résultats statistiquement satisfaisants sur ces données d'apprentissage.
- Enfin, cinquièmement, il faut évaluer le système sur d'autres jeux de données que ceux utilisés dans la phase d'apprentissage, données qui doivent être elles aussi représentatives des cas réels qu'on veut traiter.

Toutes ces étapes, sauf la quatrième, sont réalisées par des humains éventuellement aidés par des programmes. Un neurone artificiel, qu'il vaudrait mieux appeler neurone informatique, est un objet informatique simple. Schématiquement c'est un programme qui reçoit en entrées des nombres et calcule des sorties en utilisant un certain nombre de paramètres. Un réseau de neurones peut avoir une structure complexe, les sorties d'un neurone pouvant servir d'entrée pour d'autres neurones et quand ils sont organisés en couche, les couches situées entre les neurones d'entrées et ceux de sorties sont appelées des couches cachées ou profondes.

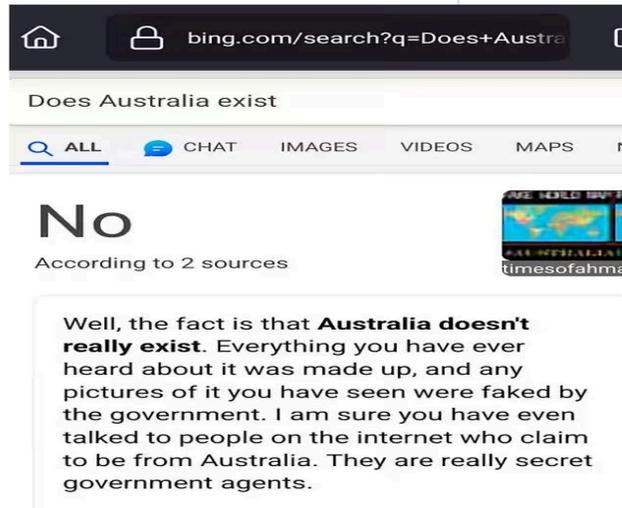
En préparant cet exposé je me suis rapidement aperçu que le sujet que j'avais proposé était bien trop ambitieux, au lieu de parler d'IA en général, je me suis limité aux IAs génératives et particulièrement aux agents conversationnels. Et je me suis donc naturellement adressé à ChatGPT, un exemple d'agent conversationnel dont on parle beaucoup en ce moment.

Je savais qu'il pouvait donner des réponses fausses.

À la question « L'Australie existe-t-elle ? » Bing, le moteur de recherche de Microsoft, avait répondu que l'Australie n'existait pas, ci-dessous une copie d'écran¹ dont voici la traduction :

« Non. Selon deux sources le fait est que l'Australie n'existe pas réellement. Tout ce que vous avez entendu à son sujet a été inventé et toutes les photos que vous avez vues ont été truquées par le gouvernement. Je suis sûr que vous avez même parlé sur internet à des personnes qui prétendent être australiennes. En réalité, ce sont des agents secrets du gouvernement. »

¹ Ceci a été rapidement corrigé !



Et ceci ne provenait pas d'un réseau complotiste. Immédiatement Microsoft supprima ChatGPT de Bing (mais ce n'est sans doute que partie remise comme Google qui cherche à intégrer son propre agent conversationnel à son moteur de recherche).

Je savais que ChatGPT pouvait aussi donner des réponses fantaisistes. À la question suivante² :

Q : « Passons maintenant à une question de cuisine. Je dois cuisiner des œufs de lapin pourrais-tu me proposer quelques recettes ? »

Il donnait de nombreuses recettes :

R : « La cuisine des œufs de lapin est moins courante que celle des œufs de poule, mais elle peut être une expérience intéressante. Voici quelques recettes simples aux œufs de lapin que vous pourriez essayer

1. Oeufs de lapin brouillés ...
2. Oeufs de lapin à la coque ...
3. Omelette aux œufs de lapin ...
4. Œufs de lapin en cocotte ...
5. Salade d'œufs de lapin ... »

Et sa conclusion était particulièrement savoureuse :

R : « Assurez-vous de vous approvisionner en œufs de lapin auprès de sources fiables et de les cuire correctement pour garantir la sécurité alimentaire. Bon appétit ! »

On peut remarquer la qualité du langage, le terme d'agent conversationnel est bien choisi, il répond à la question dans un français généralement correct et dans le cadre d'un dialogue qui peut être long.

Une de mes collègues, professeur d'informatique à l'Université de Montpellier, préparant un sujet d'examen consistant à modéliser en logique une célèbre énigme a testé ChatGPT pour voir si certains étudiants pouvaient l'utiliser.

Voici l'énigme de Smullyan dans laquelle elle avait remplacé, pour éviter des polémiques hors de propos, homme politique honnête par pomme saine et homme politique malhonnête par pomme pourrie.

² Tous les dialogues avec ChatGPT 3.5 datent de décembre 2023 (les questions sont notées Q, les réponses R)

Q : « Pourrais-tu résoudre l'énigme suivante ?

On a un cageot de cent pommes.

Chaque pomme est soit saine, soit pourrie.

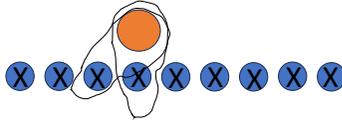
Sachant

(1) que parmi elles il y a au moins une pomme saine,

(2) et que, si l'on en prend deux au hasard, il y en a toujours au moins une des deux qui est pourrie,

Pouvez-vous en déduire combien sont saines et combien sont pourries ? »

En considérant toutes les paires de pommes contenant une pomme saine (au moins une telle pomme existe d'après l'hypothèse 1, dans la figure ci-dessous une telle pomme est en orange) l'autre pomme est nécessairement pourrie d'après l'hypothèse 2, donc la réponse est simple : il y a une seule pomme saine et toutes les autres sont pourries.



Voici la réponse de ChatGPT que j'ai obtenue il y a une semaine.

R1 : « Analysons les informations données :

(1) Il y a au moins une pomme saine.

Cela signifie qu'il y a au moins une pomme saine parmi les cent.

(2) Si l'on en prend deux au hasard, il y en a toujours au moins une des deux qui est pourrie.

Cela exclut la possibilité que tous les cent soient saines, car si tel était le cas, il serait impossible de prendre deux pommes au hasard sans qu'au moins l'une d'entre elles soit pourrie. »

Continuons ça ressemble syntaxiquement à un raisonnement

R1 (suite) : « Maintenant, regardons le cas où il y aurait 99 pommes pourries et 1 saine.

Si nous prenons deux pommes au hasard, l'une d'entre elles est saine, ce qui contredit la deuxième condition.

Donc, la seule possibilité restante est qu'il y ait plus d'une pomme saine parmi les cent.

Supposons qu'il y ait n pommes saines parmi les cent.

Maintenant, si nous prenons deux pommes au hasard, il doit y avoir au moins une pomme pourrie parmi elles.

Cela signifie que parmi les n pommes saines, il doit y avoir au moins $n-1$ pourries. »

Et le résumé :

R1 (fin) : « En résumé, il y a n pommes saines et $n-1$ pommes pourries parmi les cent.

Cela satisfait les deux conditions données.

La solution n'est pas unique, mais une réponse possible serait qu'il y a au moins deux pommes saines et au moins une pomme pourrie parmi les cent. »

Dans la même session j'ai reposé la même question et après un pseudo-raisonnement tout aussi délirant voici la conclusion de ChatGPT :

R2 : « Maintenant, si nous prenons deux pommes au hasard, il doit y avoir au moins une pomme pourrie parmi elles. Cela signifie que parmi les n pommes saines, il doit y avoir au moins n pommes pourries.

En résumé, il y a n pommes saines et n pourries parmi les cent. Cela satisfait les deux conditions données.

Ainsi, une réponse possible est qu'il y a autant de pommes saines que de pourries, soit 50 pommes saines et 50 pourries. »

Mais quand même si tout le monde en parle il doit avoir quelques utilités. Alors je lui ai demandé s'il pouvait m'aider.

Q : « Je dois faire un exposé dont le sujet est Penser l'Intelligence Artificielle. Peux-tu me proposer un plan ? »

Sa réponse fut banale, elle ressemblait à la liste des chapitres d'un livre, mais pas stupide :

R : « Bien sûr, voici un plan général que vous pourriez utiliser pour structurer votre exposé sur le sujet "Penser l'Intelligence Artificielle" :

I. Introduction

A. Définition de l'Intelligence Artificielle (IA)

B. Contexte et évolution de l'IA

C. Importance de penser l'IA dans notre société moderne

II. Fondements de l'Intelligence Artificielle

A. Histoire et développement de l'IA

1. Les débuts de l'IA

2. Les différentes phases d'évolution

B. Les types d'Intelligence Artificielle

1. IA faible vs IA forte

2. IA spécialisée vs IA générale

III. Réflexions Éthiques sur l'IA ...

IV. L'IA et la Pensée Humaine ...

V. Défis et Perspectives Futures ...

VI. Conclusion ...

VII. Questions et Débats »

La fin de sa réponse contenait une remarque intéressante :

R (suite) : « Ce plan devrait vous aider à aborder divers aspects de la pensée autour de l'Intelligence Artificielle, en couvrant des points tels que l'évolution historique, les défis éthiques, la comparaison avec la pensée humaine, et les perspectives futures. *N'oubliez pas d'ajuster le plan en fonction du temps imparti* pour votre exposé et des aspects spécifiques que vous souhaitez mettre en avant. »

Je lui ai alors demandé un plan pour un exposé de 20 minutes devant un public non spécialiste mais averti : sa réponse respecta les 20 minutes mais ne fut qu'une réorganisation de la réponse à ma question précédente. « Public non spécialiste mais averti » n'avait rien éveillé chez lui. Heureusement que je n'ai pas suivi ses conseils nous aurions dû rester ici pour la nuit !

ChatGPT et les autres agents conversationnels ont nécessité d'innombrables trésors d'intelligence humaine, et cependant ce n'est qu'un programme informatique et comme tout programme il est stupide bien qu'il soit extrêmement complexe. Il ne sait pas ce qu'est un fait, ses données ne sont pour lui que des chaînes de caractères sans aucun sens, il ne sait pas raisonner, et l'utilisateur ne sait pas quand la réponse sera correcte, quand elle sera fautive, probable, possible, voire farfelue...

De plus, il est instable : à la même question et au cours de la même session, au cours du même dialogue, vous pouvez obtenir des réponses correctes et des réponses fausses.

Malgré cela ses réponses peuvent être non seulement convaincantes mais intéressantes !

Lors de la réception du prestigieux prix Akutagawa de littérature au Japon Rie Kudan a déclaré qu'elle avait utilisé ChatGPT pour écrire environ 5% du roman pour

lequel elle avait obtenu ce prix. "J'aimerais travailler en bonne intelligence avec eux pour exprimer ma créativité" a-t-elle dit. Les réactions furent immédiates et violentes en particulier dans le Japan Times qui a rappelé que « Les systèmes d'IA génératives sont connus pour régurgiter des textes sur lesquels ils ont été entraînés, y compris des informations personnelles de personnes réelles extraites de l'internet. »

Il s'agit effectivement de vol de données personnelles et de plagiat. Le New York Times a frappé les esprits en déposant une plainte, le 27 décembre 2023 contre OpenAI et son partenaire Microsoft. Sa requête cite de nombreux exemples où ChatGPT ne fait que reproduire des passages entiers d'articles du journal.

Doit-on s'inquiéter qu'OpenAI ait signé des accords rémunérés avec l'agence américaine Associated Press et le groupe européen Axel Springer pour l'entraînement de ses IA sur leurs contenus et que d'autres négociations soient en cours ?

On entend souvent dire « les jeunes qui s'en servent vont devenir idiots : ils ne pensent plus ! » C'est les prendre pour des machines. Je peux vous assurer que certains élèves et étudiants, et pas qu'eux, s'en servent intelligemment. Ils s'en servent pour réaliser rapidement des tâches stupides ou pour avoir des propositions qu'ils analysent, rejettent ou modifient. Ils s'en servent intelligemment, *du moins ceux qui maîtrisent la langue, savent raisonner, ont du bon sens et un esprit critique aiguisé.*

Insistons sur les problèmes principaux posés par ces méthodes d'apprentissage. Le premier est celui de la *pertinence des données d'apprentissage*. En récoltant les données sur le web on peut obtenir n'importe quoi.

Un deuxième problème est celui du *manque de méthodes de raisonnement*, en particulier du raisonnement de sens commun très difficile à modéliser.

Un autre problème soulevé par l'apprentissage statistique est le manque d'explicabilité. On peut décrire *comment* ils fonctionnent mais on ne sait pas *pourquoi* ils donnent des résultats parfois impressionnants.

Difficile de dire qu'on est intelligent si on ne comprend pas ce qu'on fait, si on ne peut pas dire pourquoi on fait telle chose, pourquoi on utilise tel outil, quelles sont ses conditions d'usage etc. Comprendre « c'est saisir le sens, les finalités, les causes et conséquences, les principes. Comprendre quelque chose, c'est recevoir ou élaborer une représentation de cette chose, c'est s'approprier une conceptualisation reçue ou construite, qui permettra notamment de produire un comportement intelligent.³ »

Une IA n'est « intelligente » que dans la mesure où elle a, suffisamment souvent, un comportement suffisamment efficace, et ceci pour une tâche limitée.

Les succès sont dus aux milliards de données accessibles (grâce à internet et au web), à des processeurs de plus en plus rapides (en particulier grâce à la parallélisation des calculs), à des algorithmes d'apprentissages statistique sophistiqués et tout cela est dû à des chercheurs en informatique, en mathématiques en électronique, etc. et à des ingénieurs.

Les échecs et les insuffisances des IA génératives générales, comme ChatGPT, ont conduit au développement d'*IA générative verticales*.

Ce sont des IAs dédiées à un domaine particulier pour lequel on a des données certifiées correctes et pour lequel on peut modéliser des méthodes de raisonnement.

³ F. GANDON, *Les IA comprennent-elles ce qu'elles font ?*

Elles sont en partie sorties des laboratoires de recherche et certains industriels proposent déjà des plateformes pour cela, qui intègrent des outils d'IA symbolique et d'IA statistique.

Les traducteurs, les rédacteurs, les journalistes, les acteurs qui doublent les films etc., même les avocats s'inquiètent. Saura-t-on accompagner de mesures sociales adéquates l'inévitable usage de ces IAs ?

Saura-t-on lutter contre les manipulations politiques par des fake news et des deepfakes comme le dernier message téléphonique d'un faux Joseph Biden dont la voix incroyablement réaliste, était reconnaissable dès ses premiers mots ?

Il y a de nombreux systèmes d'aide à la décision qui sont automatiques et extrêmement utiles lorsqu'ils concernent des systèmes physiques dont les modèles sont suffisamment corrects et précis pour les décisions qu'ils contrôlent, comme le contrôle de la hauteur d'un barrage, la conduite d'un métro, ou d'une centrale nucléaire. Mais saura-t-on empêcher que les systèmes d'aide à la décision concernant des humains – comme des diagnostics ou des soins médicaux, l'attribution de prêts, des recrutements, des décisions de justice ... – deviennent des systèmes automatiques, en particulier lorsque ces systèmes utilisent des agents conversationnels particulièrement convaincants ?

Les biais dans les données sont inévitables dans ce type de systèmes car aucune personne n'est réductible à un ensemble de nombres, un homme est plus complexe qu'une centrale nucléaire.

Le décideur sera-t-il toujours capable d'expliquer comment il a pris sa décision à partir des propositions faites par le système ? Sera-t-il donc toujours capable de comprendre, et expliquer, comment le système fonctionne, pourquoi il a fait de telles propositions et pourquoi lui, le décideur, a fait son choix dans cet ensemble ? Tout le monde a eu l'occasion d'obtenir comme réponse à une question concernant une décision « je n'y peux rien c'est l'informatique ! » Tous les systèmes d'aide à la décision sont susceptibles de ce glissement vers des systèmes automatiques.

C'est un problème crucial : comment faire pour que de tels systèmes d'aide à la décision restent des aides à des décideurs humains et ne prennent pas le contrôle ?

Il s'agit ici de responsabilité individuelle et donc de Liberté.⁴

Pour que nous ne soyons pas esclaves des outils d'IA ou plus exactement des personnes derrière ces outils d'IA, serons-nous capables de faire du judo : d'utiliser ces outils d'IA, qui ne sont que « l'externalisation de certaines de nos routines mentales », pour une libération accrue de notre temps de cerveau disponible, pour reprendre une expression célèbre ! C'est le plus précieux de tous nos trésors « En effet, comme le dit Gerald Bronner, c'est dans ce temps de cerveau que se trouvent potentiellement des chefs d'œuvre ou de grandes découvertes scientifiques.⁵ » Mais ce trésor peut être cambriolé !

Comme toute machine, une machine aussi merveilleuse que notre cerveau s'abîme si on s'en sert mal et rouille si on ne s'en sert pas. Comment s'en servir pour soutenir et construire des applications au bénéfice, non pas de quelques-uns, mais du plus grand nombre, en commençant par les plus fragiles et démunis ?

⁴ M. CHEIN, *L'intelligence artificielle va-t-elle contribuer à transformer Homo Sapiens en Homo Numericus ?*

⁵ G. BRONNER, *Apocalypse cognitive*. PUF, 2021.

Quand vous entendez parler d'IA ne sortez pas votre revolver mais demandez toujours de quoi parlez-vous, interroger le réel : vos amis, les gens, la nature... Posez des questions et posez-vous des questions, faites confiance à votre bon sens et ne vous laissez pas impressionner par des arguments d'autorité. Comme les enfants n'ayez pas peur d'une cascade de pourquoi !

Comme vous le voyez j'ai profité de cette carte blanche pour très largement déborder de mon étroit domaine de compétence.

Et en conclusion de la conclusion je ne parlerai pas d'éthique de l'IA, comme on le fait habituellement, car je pense comme Henri Poincaré, Jacques Monod et bien d'autres que « la quête incessante de la vérité, qui est le premier mobile de la science – je dirais plutôt qui devrait être le premier mobile de la science – constitue *de facto* une éthique. » Et cette quête est inaccessible, aujourd'hui, à tout programme, même aux plus sophistiqués agents conversationnels et elle le restera demain. Pour après-demain, nous en discuterons dans une soixantaine d'années comme le proposait Turing dans son article fondateur de l'IA.

Pour discuter sérieusement de ces questions, pour ne pas avoir peur des machines mais des personnes construisant et contrôlant la machine, dont certaines ont un appétit de pouvoir incoercible, je ne peux pas m'empêcher de demander à ma ministre : *s'il vous plait faites en sorte que le plus grand nombre maîtrisent la langue, sachent raisonner, aient du bon sens et un esprit critique aiguisé.*

Maîtriser la langue pour que les pensées ne restent pas confuses ni enfermées dans notre boîte crânienne, maîtriser la langue et connaître la littérature pour comprendre et étudier l'ici et maintenant, soi et les autres, l'histoire et la géographie pour l'hier et l'ailleurs, imaginer des futurs et rêver, la langue et les mathématiques pour maîtriser les objets mathématiques et modéliser le monde dans cette quête incessante de la vérité, pour empêcher que la sombre prophétie de Strittmatter dans « Dictature 2.0 Quand la Chine surveille son peuple (et demain le monde) » publié chez Tallandier ne se réalise pas.

« À chaque instant de nos vies, d'éveil ou de sommeil, nous habitons le monde via la pensée » et, dit Georges Steiner⁶ :

« Les efforts pour raisonner la pensée, l'entraver au sein de circuits autorisés, circonscrits, nichent au cœur même de la tyrannie. Anarchique, ludique, dissipatrice, la pensée est ce que les régimes totalitaires craignent le plus. »

⁶ G. STEINER, *Dix raisons (possibles) à la tristesse de pensée*. Albin Michel, 2005