Cahier Azur sur l'Intelligence artificielle

Préambule

L'Intelligence artificielle (IA) est en train de devenir une technologie incontournable dans de nombreuses applications (santé, défense, énergie, transport, finances, langage,...). Après avoir été longtemps l'objet de recherches en laboratoires, ces applications de plus en plus diversifiées commencent à diffuser largement et à être connues d'un public dépassant celui des spécialistes du domaine. Le caractère disruptif de l'IA et les possibilités immenses que cette technologie laisse entrevoir, associé à un aspect mystérieux entretenu par certains médias, suscitent souvent des craintes sur les risques que pourrait entraîner un développement incontrôlé de l'IA.

Le colloque F2S

Outre des rappels scientifiques indispensables et quelques exemples d'application, le colloque F2S du 11 mai 2023 sur l'IA organisé par la Fédération française de Sociétés Scientifiques (F2S) dont la SEE est un membre actif avec ses sociétés partenaires : SFP (physique), SFO (optique) et SFV (techniques du vide) s'est conclu par une réflexion éthique sur les risques potentiels liés à l'utilisation de l'IA tout en tentant de démystifier les peurs infondées.

Le cahier azur de ce numéro REE 2023-4 est issu de certaines des interventions prononcées lors du colloque F2S sur l'IA qui s'était fixé pour objectif de focaliser les interventions sur le triptyque : Recherche, Applications et Questions éthiques associées, et on retrouvera ces trois préoccupations présentes dans le cahier qui suit.

Contenu du cahier

Sur les onze chercheurs et ingénieurs intervenus le 11 mai dernier devant un public nombreux et attentif, six auteurs ont bien voulu transcrire leur présentation dans le cahier qui suit sous forme d'un résumé plus ou moins détaillé, voire d'une contribution plus étoffée. Qu'ils en soient tous ici remerciés.

Après une introduction de **Michèle Sebag, du LISN, Université Paris Saclay** qui nous livre une description historique de l'évolution de l'IA depuis la création du concept jusqu'aux derniers développements permis par les techniques d'apprentissage profond et l'arrivée de l'IA générative, les articles qui suivent traitent successivement :

- des possibilités offertes par l'IA générative en recherche mathématique (**Marylou Gabrié, IPP/Ecole polytechnique**) ;

- d'applications de l'IA aux projets de véhicules autonomisés (**Fabien Moutarde, Mines Paris-PSL**),

- d'applications dans le domaine de la Défense : l'IA pour les radars (**Cyrille Enderli, Thales**) et l'IA au service du commandement militaire (**Gérard de Boisboissel, CReC St-Cyr**) ;

- enfin les aspects éthiques associés au développement des robots conversationnels (**Serena Villata, I3S-INRIA**).

NOTA : L'ensemble des 11 vidéos des interventions au colloque IA est consultable sur le site de F2S : https//www.f2s-asso.fr/Evénements F2S/Journée Science et Progrès 2023.

Alain Brenac,
Comité REE et vice-président de F2S

Intelligence artificielle : évolutions, révolutions et dangers

Introduction

Michèle Sebag

AO/TAU, CNRS - INRIA – LISN
Université Paris-Saclay, France
sebag@lri.fr

Reprenant les grandes lignes de l'exposé donné à la F2S le 11 mai 2023, ce document souhaite situer les enjeux de l'Intelligence artificielle et leurs évolutions sur les 70 dernières années.

Les prémices : 1950 – 2000

Comme chacun sait, le terme *Artificial Intelligence* (IA) a été forgé par John Mc-Carthy en 1956 dans le cadre de la conférence d'été de Dartmouth, réunie pour explorer la conjecture que « *tous les aspects de l'apprentissage ou de l'intelligence* (pouvaient) *en principe être décrits assez précisément pour qu'une machine puisse les simuler* ». Six ans auparavant, Alan Turing s'était également posé la question de savoir si les machines pouvaient penser [1]. Il arrive à la question de l'apprentissage des machines par un détour frappant : il estime la taille d'un programme permettant de reproduire une partie de l'intelligence humaine, le confronte à sa propre capacité de programmeur (une page par jour), et en déduit que la création *ex nihilo* d'un tel programme par des programmeurs humains est hors de portée. Une seule solution : que la machine prenne en charge au moins en partie la construction de ce programme, ce qui implique qu'elle soit capable d'apprendre. La méthode qu'il envisage est fortement inspirée de la théorie du behaviorisme de Skinner : *on pourrait en principe réaliser une telle machine au moyen de deux signaux, correspondant respectivement à une récompense (ou plaisir) et à une punition (ou douleur)*. Ces deux visions de l'Intelligence artificielle vont donner lieu à des programmes de recherche très différents, selon que le but est d'arriver à l'intelligence de la machine en s'inspirant des voies humaines, ou par n'importe quel moyen.

Du côté de Turing, l'intelligence de la machine est établie lorsqu'un observateur extérieur ne sait plus faire la différence entre l'humain et la machine simulant l'humain – c'est le test de Turing. Mais ce test peut échouer ou réussir pour de mauvaises raisons (les connaissances ; les usages ; le sens commun...), ce qui ne donnait pas à l'époque de critères rigoureux et ne permettait pas de fonder une discipline scientifique.

Du côté de McCarthy, l'intelligence de la machine est établie lorsque celle-ci se montre capable de résoudre des problèmes logiques et d'obtenir de bonnes performances dans le cadre des jeux ; une réalisation emblématique de ce courant est le *General Problem Solver* de Herbert Simon, Cliff Shaw et Allen Newel (1959). Cette voie aboutit à de premiers résultats brillants (trouver une nouvelle démonstration d'un cas d'égalité des triangles ; trouver des erreurs dans le Principia Mathematica de Whitehead et Russell) mais avec des rendements décroissants : chaque progrès du système demande autant d'efforts que ce qu'il a fallu pour en arriver là. A un certain point, cette voie parait sans avenir et c'est le premier hiver de l'IA.

Plusieurs voies sont explorées allant de la programmation logique aux systèmes experts puis aux réseaux neuronaux de première génération. Les hivers et les printemps de l'IA se succèdent, avec deux constantes : tout d'abord, les difficultés majeures ne sont pas là où on les attendait ; par exemple il est plus difficile d'apporter le plateau du petit déjeuner au lit, que d'exceller aux échecs. En second lieu, les promesses sont pharaoniques, éveillant des espoirs démesurés, suivies de déceptions et du discrédit de la discipline (hiver, et assèchement des financements de l'IA). Vers la fin des années 80-90, l'IA n'est pas prise fort au sérieux par les bons scientifiques, ni par les mathématiciens ni par les informaticiens.

1 Computing Machinery and Intelligence, A. Turing, 1950

L'irrésistible ascension de l'IA : 2000 - 2020

La phase actuelle de vogue de l'IA – portée aux nues par les médias et représentant un enjeu de souveraineté majeur pour les Etats-Unis, l'Europe, la Chine, etc. – commence en 2006, avec l'essor de l'apprentissage profond ou *Deep Learning*, dû à Yann Le Cun, Yoshua Bengio, Geoffrey Hinton, qui reçoivent le prix Turing en 2018 (et bien d'autres : Léon Bottou,…). Cet essor de l'IA est rendu possible par deux faits nouveaux : l'abondance des données (notamment grâce au Web) et, dans une moindre mesure, l'abondance des moyens de calcul. L'abondance des données conduit à formaliser l'IA comme un problème d'optimisation : savoir reproduire la réponse des experts sur les données connues (i.e. pour lesquelles la réponse de l'expert est connue ; on parle de données d'entraînement). Ce critère d'apprentissage fait aussi intervenir certaines garanties, permettant d'assurer qu'en moyenne, les réponses obtenues sur des données nouvelles seront de même qualité que celles de l'expert. Ce problème d'optimisation, généralement de grande taille et non convexe, peut être attaqué avec succès grâce à l'augmentation des moyens de calcul. En résumé, l'IA part des données vues et étiquetées par l'expert, et apprend une solution optimale, ou modèle, permettant d'étiqueter de nouvelles données comme l'expert l'aurait fait.

L'abondance des données et des moyens de calcul conduit à réapprécier des approches anciennes, typiquement les réseaux neuronaux. En dépit de leurs succès expérimentaux, ceux-ci étaient tombés en désaffection : du point de vue de la théorie, il y avait peu de résultats utiles[2] ; du point de vue des résultats empiriques, les experts des réseaux neuronaux obtenaient des résultats exceptionnels, mais les non-experts n'arrivaient pas à les reproduire [3]. Pour toutes ces raisons, les réseaux neuronaux étaient dominés depuis les années 90 par les machines à vecteurs support (SVM, aussi appelés systèmes à vastes marges) ou apprentissage à noyaux, dont le cadre théorique est solide, élégant et fertile [4].

> **" L'abondance des données conduit à formaliser l'IA comme un problème d'optimisation : savoir reproduire la réponse des experts sur les données connues (i.e. pour lesquelles la réponse de l'expert est connue ; on parle de données d'entraînement). "**

Que s'est-il passé en 2006 ? Un atelier mémorable s'est tenu à la suite de la conférence majeure de l'apprentissage, NIPS/NeurIPS, pour faire passer un message nouveau. « Avons-nous tourné le dos aux objectifs et promesses de l'IA ? Les succès que nous remportons reposent *in fine* sur d'énormes efforts humains, pour obtenir une bonne représentation des données ». Il est de fait que les succès des SVM passent par la conception de bons noyaux, adaptés au domaine de l'application ; dans les domaines complexes, la conception, l'étude théorique et la validation d'un noyau approprié pouvaient constituer une thèse. « L'acquisition de connaissances passe par la construction graduelle de notions abstraites, jusqu'au moment où on dispose des concepts qui permettent de bien représenter le problème considéré. C'est ce que fait un réseau profond : chaque couche neuronale est construite sur la précédente. » En résumé, *la construction d'une bonne représentation n'est pas une étape préliminaire : elle fait partie de l'apprentissage.* Il était connu depuis la fin des années 80 que les réseaux profonds permettaient de représenter des concepts complexes, de manière exponentiellement plus compacte qu'un réseau superficiel ; cependant, les réseaux profonds n'étaient pas utilisés, parce qu'on ne savait pas résoudre les problèmes d'optimisation dans cet espace. C'est ce verrou algorithmique qui a été levé.

En 2012, les réseaux profonds sont reconnus à l'occasion du concours international ImageNet Large Scale Visual Recognition Challenge, dont l'objectif est d'identifier automatiquement les objets représentés dans des images collectées sur le Web. Les meilleures performances au niveau mondial étaient autour de 26 %, s'améliorant de 1 à 2 % par an, quand le réseau profond AlexNet de A. Krizhevsky, I. Sutskever et G. Hinton prend part à la compétition et se classe premier, avec plus de 10 % de progrès comparé à la seconde approche. Il s'agit d'un saut technologique. L'année suivante, les huit premières approches seront fondées sur les réseaux profonds.

De tels bonds de performance, dans les domaines de la vision, de la parole et des jeux (voir ci-dessous), expliquent la place de plus en plus grande prise par l'apprentissage profond dans les conférences internationales en IA et en apprentissage. Pour les médias et le grand public, l'apprentissage profond apparaît comme un deus ex machina, capable d'aligner en série des records de performance. A côté de la vision et de la parole, les résultats de l'apprentissage ●●●

2 Ainsi un théorème fondamental est que toute fonction de carré intégrable peut être approximée sur un compact par un réseau neuronal ; mais ce résultat n'est pas constructif.

3 En pratique, la performance repose sur l'ajustement de paramètres opaques, maîtrisé par peu d'experts

4 A training algorithm for optimal margin classifiers, B. Boser, I. Guyon & V. Vapnik, 1992.

profond font aussi la une sur le front des jeux, en commençant par les jeux sur ordinateur[5]. Le tour de force est que le système d'apprentissage par renforcement en question excelle à plusieurs jeux (Pong, Space Invaders, ...), alors que les IA championnes de l'ère précédente étaient dédiées à un seul jeu (TD-Gammon pour le backgammon ; Deep Blue pour les échecs). AlphaGo, premier système à battre un champion humain au jeu de Go (2014, suivi d'AlphaGo Zero en 2017 ; DeepMind) est développé selon les mêmes principes. Il s'agit du seul jeu de Go, certes, mais c'était l'un des défis majeurs de l'IA en raison de sa complexité et de la taille de l'espace du jeu, et les experts ne pensaient pas voir une IA l'emporter sur les champions humains avant plusieurs décennies. Ces succès sont transposables dans des domaines complexes : AlphaFold (2021, DeepMind) s'attaque à la prédiction de la structure 3D des protéines (une étape cruciale pour le domaine de la chimie organique et la conception de médicaments) et remporte les concours internationaux du domaine.

Comme nous l'avons dit, l'apprentissage est formalisé comme un problème d'optimisation, dont la solution garantit de manière simplifiée qu'elle obtiendra les mêmes résultats que l'expert sur les données d'entraînement, et qu'elle obtiendra des résultats de qualité similaire sur des données issues de la même distribution. La difficulté statistique et algorithmique est de formuler le critère d'optimisation : celui-ci doit fournir de bonnes garanties statistiques, mais aussi permettre une optimisation efficace sur des espaces de modèles grands et complexes (un réseau profond comprend couramment de quelques centaines de millions à un milliard de paramètres, et la taille moyenne des réseaux augmente – nous y reviendrons).

Dans les cercles de l'apprentissage, une loi empirique est actée : les données battent les algorithmes, et il est donc plus efficace de chercher à accumuler des données qui seront exploitées par un algorithme simple, que de construire un algorithme complexe, qui exploitera des données moins nombreuses. En simplifiant, la complexité introduite par les concepteurs dans un algorithme ne capture pas la complexité du réel.

La primauté des données inspire un objectif nouveau : être capable de générer des données ressemblant du mieux possible aux données réelles. Il s'agit d'apprentissage génératif. En somme, l'objectif est de réaliser un mécanisme d'échantillonnage, qui mime la probabilité des données du monde réel (en évitant la solution triviale qui consisterait à générer encore et toujours les données d'entraînement). La difficulté est de réaliser un tel mécanisme d'échantillonnage dans des espaces de très grande dimension.

Au cœur de la révolution du *Deep Learning*, une autre révolution apparaît en 2014. Cette révolution s'attaque à la question de savoir si la distribution du mécanisme d'échantillonnage et celle du monde réel sont égales. La première réponse repose sur des tests statistiques. La seconde repose sur l'apprentissage lui-même : si on cherche à discriminer les données issues du mécanisme d'échantillonnage et celles du monde réel, et si on y arrive, alors ces deux distributions sont nécessairement différentes.

A la vision d'une unique « intelligence artificielle » cherchant à générer des données réalistes, se substitue ainsi la vision d'une paire d'IA en compétition : la première (le générateur) cherche à générer des données, la seconde (le discriminateur) cherche à discriminer les données générées des données réelles ; les deux IA sont antagonistes et apprennent de manière jointe : la première réussit si la seconde échoue. Ce mécanisme contradictoire correspond à embarquer un test de Turing dans l'IA génératrice ; le point clé est que ce test de Turing est réalisé de façon autonome. Les succès de l'apprentissage génératif contradictoire font rêver, particulièrement dans le domaine de l'image : les images générées sont nouvelles, réalistes, créatives... avec quelques anomalies ici ou là, nous y reviendrons.

Fondamentalement, le domaine avance, mais les difficultés consistent toujours à trouver un compromis entre les garanties théoriques apportées par le critère d'apprentissage, et la difficulté algorithmique de son optimisation.

Quelques côtés obscurs de l'IA

Depuis la fin des années 2010 cependant, plusieurs conséquences négatives de l'IA sont identifiées, de natures différentes.

Une première limitation tient à la nature des garanties offertes par un modèle appris. Il s'agit de garanties statistiques, établissant une borne supérieure quant à la probabilité des erreurs commises, sous l'hypothèse que les données futures sont issues de la même distribution que les données d'entraînement.

5 Playing Atari with Deep Reinforcement Learning, V. Mnih et al., 2013

" Au cœur de la révolution du *Deep Learning*, une autre révolution apparaît en 2014. Cette révolution s'attaque à la question de savoir si la distribution du mécanisme d'échantillonnage et celle du monde réel sont égales. "

"Les modèles appris sont fondés sur des données et dans de nombreux domaines (commerce, banque, justice), les données peuvent refléter les préjugés humains et les biais de décision. Les modèles, reflétant les données, intègrent et perpétuent ces préjugés. "

Une telle garantie présente plusieurs faiblesses :

- la première faiblesse est que dans des domaines critiques (e.g. les véhicules autonomes, l'aviation), il faut pouvoir garantir des probabilités d'erreur ou de panne très faibles (10^{-7}), ce qui demande des bases d'apprentissage considérables. Or, dans beaucoup de problèmes réels et notamment dans les secteurs industriels, il n'y a pas abondance de données...

- la deuxième faiblesse est que les distributions ne sont pas fixes ; même les immenses distributions des images sur le Web peuvent présenter des évolutions perceptibles sur 10 ans ;

- la troisième faiblesse, la plus grave, est que les modèles appris (les IAs) sont appris sur des espaces de très grande dimension ; ainsi la dimension d'une image est son nombre de pixels, qui se compte en milliers ou en millions. Or, dans de tels espaces, il s'avère faisable de modifier une image de manière imperceptible pour la vision humaine, mais de manière à modifier sa classification par l'IA. On parle ici d'exemples « adversariaux » [6].

Le danger serait ainsi, dans le contexte d'un véhicule autonome, qu'un panneau routier « Stop » ne soit pas interprété comme un stop (mais par exemple comme une limitation de vitesse), en raison de quelques taches de boue ou d'altérations malveillantes, avec les conséquences qu'on imagine sur la sûreté des véhicules. Le problème des exemples « adversariaux » est donc attaqué en cherchant à rendre les réseaux plus robustes. Mais ces réseaux plus robustes peuvent encore être leurrés par des attaques plus puissantes, et ainsi de suite. Fondamentalement, on voudrait pouvoir certifier le comportement d'un réseau neuronal, comme on sait certifier un automate de métro ou un ATM. Ce problème est présentement un problème ouvert.

Une seconde limite est *de nature éthique*. Les modèles appris sont fondés sur des données et dans de nombreux domaines (commerce, banque, justice), les données peuvent refléter les préjugés humains et les biais de décision. Les modèles, reflétant les données, intègrent et perpétuent ces préjugés. Ainsi, dans le cas du système COMPAS *Correctional Offender Management Profiling for Alternative Sanctions* [7], les décisions de mise en liberté conditionnelle apparaissent entachées de biais racistes. Ces biais sont d'autant plus inacceptables que les modèles sont souvent opaques : ils ne peuvent pas expliquer les décisions prises d'une manière qui permettrait de remettre ces décisions en question. Plus généralement, à partir du moment où un modèle intervient dans des décisions importantes pour les gens (allant par exemple de la mise en liberté conditionnelle à l'octroi d'un prêt bancaire, de ParcourSup à l'aide au recrutement), sa transparence et son impartialité (*fairness*) sont cruciales. Le problème est extrêmement épineux parce qu'il couple des questions d'ordre différent, éthiques et algorithmiques. Algorithmiquement, le fait d'expliquer la décision d'un réseau profond constitue encore un problème ouvert (et la définition d'une explication bien fondée est subjective). Ethiquement, la notion d'impartialité n'est pas définie de manière unique ; dans le contexte du recrutement par exemple, les préférences des hommes et des femmes sont différentes. Doit-on négliger les préférences des utilisateurs ou utilisatrices ? Certainement non. Doit-on les respecter ? Mais ces préférences peuvent elles-mêmes résulter d'un conditionnement ou d'une pression sociale ...

Une troisième limite est de nature fondamentale, et touche à l'objectif même de l'apprentissage. Une partie de la fascination exercée par l'apprentissage et l'IA tient au programme d'Auguste Comte (1798-1857) : savoir pour prévoir afin de pouvoir. Supposons en effet que nous soyons capables de prédire si un malade va se rétablir ; il est tentant de penser que cette connaissance peut être utilisée pour le soigner différemment, et augmenter ainsi sa probabilité de rétablissement... Cependant, la prédiction et l'identification d'interventions appropriées sont des opérations différentes. Il est facile de prédire qu'il pleut, si on voit des parapluies dans la rue. Mais cette connaissance ne permet pas de faire pleuvoir (sortir avec un parapluie ?). Formellement, les prédictions peuvent être fondées sur des corrélations (la pluie et les parapluies) ; mais les interventions doivent être fondées sur des causalités (les parapluies ne causent pas la pluie). En résumé, dans de nombreux domaines (éducation, climat, santé, commerce, pour en citer quelques-uns), l'espoir est que les modèles d'IA conduisent à définir de bonnes politiques d'action, permettant de savoir que faire pour que les élèves apprennent bien, que le climat et la biodiversité se remettent au ●●●

6 Explaining and Harnessing Adversarial Examples, I. Goodfellow, J. Shlens & C. Szegedy, 2014

7 Artificial Intelligence Predictive Policing: Efficient, or Unfair?, Akhara 2021.

mieux, que la santé s'améliore, qu'un plan de marketing obtienne les effets voulus, etc. Mais ces politiques d'action consistent à intervenir sur le monde, et donc à modifier la distribution des données que nous observons ici et maintenant.

Selon la théorie de J. Pearl [8], l'intelligence est structurée de manière hiérarchique. Au premier niveau, l'observation des régularités permet de prédire ce qui va se passer a priori. Au second niveau, la question est de prédire les effets des interventions (que se passe-t-il si je lance une pierre, si j'absorbe un médicament, etc.). Le troisième niveau est celui des raisonnements contrefactuels : que se serait-il passé si je n'avais pas pris ce médicament ; si Jean Jaurès n'avait pas été assassiné ?

Il est clair que chaque niveau de cette hiérarchie est beaucoup plus complexe que le précédent. Le premier niveau est en cours de résolution (sous réserve de données et de moyens de calcul suffisants, nous y reviendrons). Les deuxième et troisième niveaux sont des problèmes ouverts.

Où en sommes-nous ?

Un dernier rebondissement majeur de l'IA est celui des grands modèles de langue (LLM), comprenant chatGPT et ses alter ego (LLAMA, Bloom, PALM,...). Il ne s'agit pas d'un saut technologique : les ingrédients de chatGPT existaient depuis quelques années. L'innovation sans précédent est que la plateforme chatGT a été ouverte à tous par OpenAI le 30 novembre 2022, et que les gens en explorent depuis les usages possibles avec une créativité incroyable, de nature à remettre en cause les fondamentaux de l'organisation actuelle de l'information, du droit d'auteur à l'enseignement, des centres d'appels à la programmation.

Les essais précédents des acteurs majeurs de l'IA (Microsoft, Google, ...) de mettre un agent conversationnel à la disposition du grand public avaient tourné au désastre, exposant la fragilité de l'IA à la malveillance volontaire ou involontaire. Concrètement, une IA ne disposant pas de garde-fous éthiques, il est aisé de l'induire à faire preuve de racisme, de sexisme, etc. Cette faiblesse a été partiellement corrigée, en entraînant chatGPT à savoir ce qui ne doit pas être dit. Mais la correction nécessite de garder des armées de gardiens humains dans la boucle, et dépend toujours de la vigilance des gardiens... Le remède complet consisterait à disposer d'une définition calculable de l'éthique : ce qui semble impossible, si on admet la diversité des cultures humaines.

Les LLM fournissent une réponse à la plupart des questions, sous la forme désirée (longueur, niveau de langage, langue...) et s'appuyant sur l'ensemble de l'information existante ; il s'agit d'une démocratisation sans précédent de l'information disponible. Ces réponses sont presque toujours plausibles ; mais elles ne sont pas nécessairement vraies. Pour situer la différence et le danger, Google fournit des moyens de répondre aux questions posées, sous forme de documents sources et sans masquer les contradictions possibles de ces sources. Par opposition, chatGPT fournit une réponse, présentée comme « la » réponse. La seule option possible, pour explorer la fiabilité de cette réponse, consiste à reposer la question sous une forme différente ; en d'autres termes, le fait de développer le sens critique de chacun devient une nécessité vitale.

Il faut aussi revenir sur la question des moyens : le coût, tant au niveau des infrastructures de calcul, que de l'énergie nécessaire pour apprendre et pour exploiter les modèles, augmente régulièrement. Les précisions sont difficiles à obtenir ; mais l'énergie nécessaire pour faire tourner chatGPT serait de 1 GWh par jour (septembre 2023). Le fait que chacun des acteurs majeurs de l'IA s'engage dans la course des LLM, et qu'il existerait plus de 70 LLM à plus d'un milliard de paramètres, est une menace non négligeable dans un monde fini et appelle une régulation transnationale. ■

8 Causality: Models, Reasoning, and Inference, J. Pearl, 2000.

"Un dernier rebondissement majeur de l'IA est celui des grands modèles de langue (LLM), comprenant chatGPT et ses alter ego (LLAMA, Bloom, PALM,...). Il ne s'agit pas d'un saut technologique : les ingrédients de chatGPT existaient depuis quelques années."

L'auteure

Michèle Sebag est directrice de recherche au CNRS, au Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) à l'Université Paris-Saclay.
De formation mathématique (ENS), elle s'est formée à l'informatique dans l'industrie, avant de passer sa thèse et de rentrer au CNRS. Elle est responsable de l'équipe Apprentissage et Optimisation au LISN, et co-responsable avec Marc Schoenauer de l'équipe INRIA TAU (*Tackling the Underspecified*). Ses thèmes de recherche comprennent la modélisation causale, l'apprentissage par renforcement et les applications de l'apprentissage aux sciences humaines et sociales (pour les embauches, pour la santé). Elle a été élue European AI Fellow, et membre de l'Académie des Technologies.