INSTITUT Montaigne

Pour une Autorité française de l'IA

NOTE D'ACTION - JANVIER 2024

Think tank de référence en France et en Europe, l'Institut Montaigne est un espace de réflexion indépendant au service de l'intérêt général. Ses travaux prennent en compte les grands déterminants économiques, sociétaux, technologiques, environnementaux et géopolitiques afin de proposer des études et des débats sur les politiques publiques françaises et européennes. Il se situe à la confluence de la réflexion et de l'action, des idées et de la décision.

NOTE D'ACTION - Janvier 2024

Pour une Autorité française de l'IA

Les notes d'action de l'Institut Montaigne identifient un enjeu spécifique et formulent des recommandations opérationnelles à destination des décideurs publics et privés.

L'URGENCE D'UN ESPACE DE COMPÉTENCES ET DE RECOURS SUR L'IA

La France accueillera bientôt la deuxième grande édition du Sommet Mondial de l'IA, initié par le Royaume-Uni en novembre 2023. Néanmoins, notre pays manque d'un outil de gouvernance indispensable sur son propre territoire et comme maillon européen : une Autorité française de l'IA. Cette Autorité aurait pour mission principale d'évaluer rigoureusement la performance et les risques de l'état de l'art de la technologie, qui avance à grands pas, et d'accompagner les entreprises de l'IA dans leur gestion des risques. Depuis la publication de la note d'action de l'Institut Montaigne d'avril 2023 sur l'IA « sûre et digne de confiance » qui formulait des recommandations en ce sens, plusieurs pays *leaders* de l'IA – le Royaume-Uni, les États-Unis et Singapour – ont annoncé leur propres Instituts de Sûreté de l'IA en fin d'année 2023. La France est désormais en retard. Une première étape est de créer, dès 2024, un Centre d'évaluation de l'IA, rassemblant les forces et acteurs en présence. Un tel centre pourra préfigurer une future Autorité créée par la loi, à même de faire avancer les sujets d'IA en France en plus de devenir un interlocuteur facilitant la mise en conformité de l'application du règlement européen en la matière.

L'ÉVALUATION : UN OUTIL CLÉ POUR BÂTIR LA GOUVERNANCE MONDIALE ÉMERGENTE

Lors du premier Sommet Mondial de l'IA en novembre 2023, 28 pays, dont la France, se sont engagés à maîtriser l'IA *via* deux leviers : l'identification et l'évaluation des risques, d'une part, et des mesures de gouvernance centrées sur les risques, d'autre part.

Le Royaume-Uni et les États-Unis ont rapidement adopté une approche centrée sur l'évaluation des risques pour la sécurité nationale des États de

l'IA à l'état de l'art (*Frontier AI* en anglais) – soit les quelques grands modèles d'IA à usage général dépassant un seuil de performance particulièrement élevé – sans toutefois imposer de mesures de gouvernance *a priori*.

D'autres acteurs se sont dotés de règles contraignantes, en estimant que l'auto-régulation par les entreprises n'était pas satisfaisante. C'est le cas de l'Europe, qui régulera avec son *AI Act* les systèmes d'IA déployés dans des cas d'usages à « risque élevé » et les plus grands modèles d'IA à usage général, indépendamment de leur cas d'usage. C'est aussi le cas de la Chine, qui a rapidement adopté une série de mesures pour strictement réguler l'IA et les acteurs qui commercialisent des systèmes d'IA à usage général. En octobre 2023, Pékin a également introduit un cadre de gestion des risques pour la recherche dans l'IA, lui permettant de distinguer plus facilement les enjeux de contrôle du contenu généré par l'IA et les risques liés à la recherche sur l'IA de pointe.

Qu'il existe un cadre réglementaire ou non, l'évaluation apparaît désormais comme un outil indispensable : il peut être un filet de sécurité minimum en l'absence de règles strictes, permettant une intervention spontanée des pouvoirs publics en cas de risque jugé inacceptable ; il peut être également le moyen de faire évoluer la réglementation au rythme de la technologie.

ANTICIPER LA DÉFERLANTE IA ET LES RISQUES ASSOCIÉS

L'année 2023 aura été celle de la déferlante IA, marquant une accélération sans précédent de l'évolution de la technologie. Avec ChatGPT, le grand public a découvert la puissance d'une nouvelle ère de l'IA, celle des systèmes d'IA à usage général, capables d'effectuer un grand nombre de tâches différentes (coder, dialoguer en plusieurs langues, résoudre des problèmes de physique ou de chimie, etc.).

Depuis, la technologie a fait des bonds significatifs. En mars 2023, OpenAI, l'entreprise mère de ChatGPT, a dévoilé son modèle GPT-4, dont le Quotient intellectuel (QI) mesuré par des tests d'aptitude intellectuelle dépasse celui d'entre 80 % et 99 % de la population. Les capacités d'action de ces modèles d'IA ont rapidement été augmentées au-delà de la simple génération de textes ou d'images. En particulier, ils ont été utilisés pour créer des agents IA, c'est-à-dire des systèmes d'IA dotés d'objectifs, connectés à d'autres outils externes tels que des outils de codage informatique, des moteurs de recherche, des laboratoires scientifiques pilotés à distance *via* internet et donc désormais capables d'interagir avec le monde externe en toute autonomie.

Chacun commence à comprendre le potentiel transformatif de cette technologie, constate le rythme exponentiel de son développement et ressent le besoin de mieux l'appréhender. L'enjeu est désormais de passer à l'action.

Auteur

Milo Rignell

Milo Rignell est responsable de projets et Expert Résident de l'Institut Montaigne sur les sujets numériques et de nouvelles technologies. Précédemment au poste de chargé de l'Innovation au sein du *think-tank*, Milo était responsable de projets d'expérimentation, dont une formation en ligne à l'intelligence artificielle, Objectif IA. Milo est diplômé de l'université d'Oxford en philosophie, politique et économie (PPE).

Table des matières

1 2023 : l'année de la déferlante IA

1.1. DES MODÈLES D'IA DE PLUS EN PLUS COMPÉTENTS

Le terme « intelligence artificielle » englobe un nombre important de techniques et d'algorithmes. **Néanmoins, ce qui importe n'est pas tant la technologie en elle-même que ses compétences, c'est-à-dire ce qu'elle est en capacité de faire concrètement.**

Nous sommes à un moment de changement de paradigme. Précédemment, nos systèmes d'IA étaient conçus pour un usage spécifique et ne pouvaient accomplir qu'un faible nombre de tâches. Certains systèmes pouvaient identifier des objets dans une image, d'autres pouvaient jouer aux échecs, d'autres encore pouvaient traduire des textes d'une langue vers une autre. Il s'agissait de systèmes d'IA à usage spécifique.

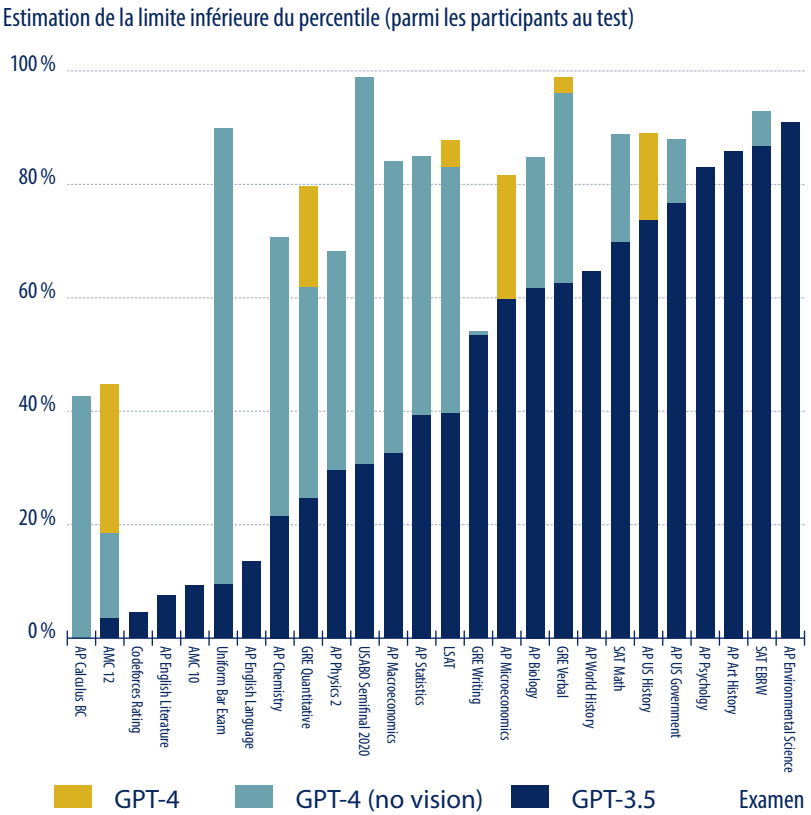
Depuis 2017, une nouvelle architecture de modèle d'IA, appelée « transformer »[1], a permis le développement de grands modèles d'IA plus polyvalents, ouvrant la voie à l'émergence de systèmes d'IA à usage général. Ces systèmes se définissent par leur capacité à effectuer un grand nombre de tâches différentes, à un niveau de performance élevé.

ChatGPT, le plus populaire des modèles à usage général, a été dévoilé il y a tout juste un an, en novembre 2022. Depuis, la performance des systèmes d'IA à usage général a rapidement augmenté. GPT-3, l'avant-dernier modèle d'OpenAI sur lequel se basait ChatGPT à son lancement, ne dépassait que 25 % des étudiants à l'université sur les exercices quantitatifs du GRE, un test d'aptitude intellectuelle utilisé par de nombreuses universités. GPT-4, le dernier modèle d'OpenAI, atteint un score qui le

[1] *Le transformer est une architecture de modèle d'IA présentée par Google Research en 2017, basée sur un mécanisme d'auto-attention permettant de mieux comprendre le contexte d'une information et ainsi d'identifier les informations pertinentes.*

place au-dessus de 80 % des élèves ayant réalisé ce test. Pour le raisonnement verbal, les performances sont passées de 63 % pour GPT-3 à 99 % pour GPT-4.

Résultats de l'examen
(classés par performance GPT-3.5)



Source : OpenAI[2]

[2] *https://openai.com/research/gpt-4.*

Désormais, les modèles d'IA les plus avancés peuvent interagir[3] avec des applications en ligne, générer et exécuter du code, pirater des systèmes informatiques[4], piloter[5] des robots dans le monde réel et incarner des agents[6] dans des mondes simulés, concevoir[7] des plans et les exécuter.

Les grands modèles d'IA actuels, comme ceux de ChatGPT ou de Bard, sont dotés d'une capacité fondamentale : ils peuvent prédire l'élément suivant dans une suite d'éléments de façon extrêmement performante. Quand cela concerne du texte, il s'agit de prédire le mot qui devrait arriver par la suite, par exemple après une question posée par un utilisateur. Pour qu'un modèle d'IA puisse agir ainsi de façon très performante, aussi bien en chimie qu'en histoire ou en politique, il doit s'entraîner à la lecture d'une grande quantité de textes proposés sur internet. Après cette période d'enseignement, quand le modèle d'IA reçoit la question d'un utilisateur, il est capable de générer une réponse, soit une séquence de mots qui pourrait crédiblement faire suite à l'interrogation posée. Sa réponse s'appuie sur ses données d'entraînement, soit tout ce qu'il a lu sur internet.

La capacité à prédire l'élément suivant (*next token prediction* en anglais) s'avère extrêmement polyvalente et permet beaucoup plus que la simple génération de contenus, bien que cela constitue le principal cas d'usage à date des modèles d"IA à usage général dits « génératifs ». Il s'agit de prédire l'élément (le *token*) qui devrait être généré après une entrée initiale *(prompt)* dans une séquence quelconque[8], qu'elle soit une séquence de

[3] *https://openai.com/blog/chatgpt-plugins.*

[4] *À un niveau confirmé (niveau 3/3 sur la plateforme hackthebox.org dans les « starting machines »).*

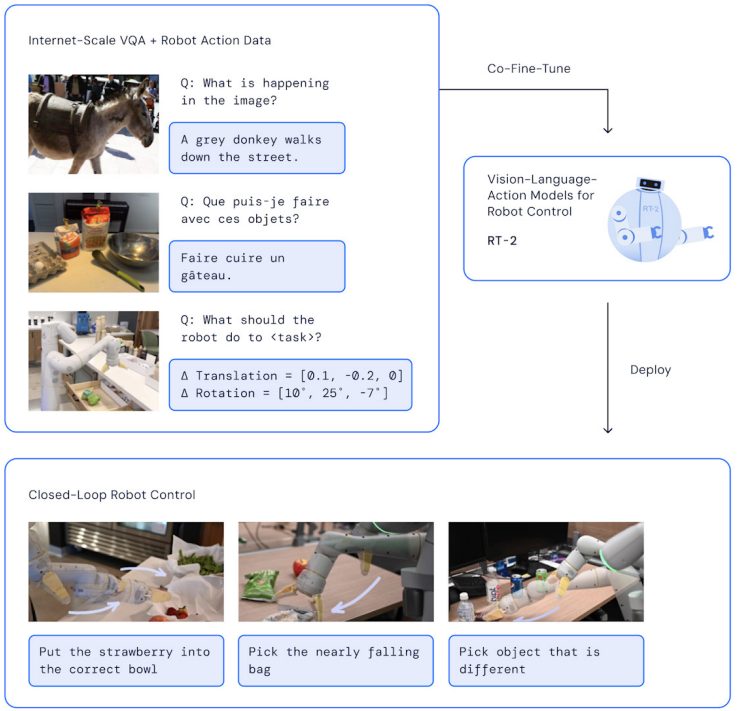
[5] *https://robotics-transformer-x.github.io/.*

[6] *https://voyager.minedojo.org/?utm_source=substack&utm_medium=email.*

[7] *https://techcrunch.com/2023/04/22/what-is-auto-gpt-and-why-does-it-matter/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAADJX9Ibup0TMeNBF-FBAcvnqCJnYHgvSw0Tspaszu9jNbOQ1wRdifS92qTqJyoe8774l57jyzOKsnQj-Oi6KsTopZa3XRjG-1BZ2Vp-axz3B0g4entAHbgB3qDXhJ3vKA_x9x1iupe4Hgk2gsTw9-nxUnO0LN9d_KqtpoHACpzduq2.*

[8] *C'est ce qu'on appelle une série temporelle.*

mots, de sons, d'images vidéo ou d'actions – par exemple pour piloter un robot, comme l'illustre l'exemple ci-dessous développé par Google. L'avantage du texte est que nous disposons de beaucoup de données d'entraînement disponibles en ligne sur internet – c'est ce qui explique le succès des grands modèles de langage comme ChatGPT.



La prédiction de l'élément suivant (next token prediction) s'applique aussi aux actions. Le récent modèle RT-2 de Google prend en compte les images de la caméra du robot et prédit directement les actions à effectuer par le robot, selon la demande initiale de l'utilisateur.

Source :
https://deepmind.google/discover/blog/rt-2-new-model-translates-vision-and-language-into-action/.

Bien que le terme « IA à usage général » renvoie le plus clairement aux multiples compétences de ces nouveaux modèles d'IA – soit leur élément distinctif – en 2023 la déferlante technologique a été accompagnée d'une explosion sémantique. « IA générative », « IA à usage général », « modèle de fondation », « *Frontier AI* », « LLM »… De nombreux termes ont été proposés pour mettre en valeur différents aspects de ces nouveaux systèmes d'IA.

Un emballement sémantique autour de l'IA à usage général en 2023

Modèle de langage de grande taille (*large language model* ou LLM)

Modèle statistique de la distribution d'unités linguistiques (par exemple : lettres, phonèmes, mots) dans une langue naturelle. Un modèle de langage peut par exemple prédire le mot suivant dans une séquence de mots. On parle de modèles de langage de grande taille ou « *Large Language Models* » (LLM) en anglais pour les modèles possédant un grand nombre de paramètres (généralement de l'ordre du milliard de poids ou plus) comme GPT-3, BLOOM, Megatron NLG, Llama ou encore PaLM[9].

Comprendre la subtilité : les premiers modèles d'IA à usage général étaient des modèles de langage, entraînés sur du texte, mais rien n'empêche de futures générations de modèles d'IA à usage général d'être entraînées sur autre chose que du texte.

Exemple : à son lancement, ChatGPT était un grand modèle de langage, capable de générer uniquement du texte (langage naturel) et du code (langage informatique). Désormais, il existe des modèles d'IA multimodaux, capables de traiter du texte mais aussi d'autres modalités de données (vidéo, son, etc.). C'est par exemple le cas des modèles d'IA GPT-4V et Gemini.

IA générative

Un système capable de créer du texte, des images ou d'autres contenus (musique, vidéo, voix, etc.) à partir d'une instruction d'un utilisateur humain. Ces systèmes peuvent produire des nouveaux contenus à partir de données d'entraînement[10].

Comprendre la subtilité : « IA générative » fait référence à l'ensemble des systèmes d'IA capables de créer du contenu, soit le principal cas d'usage des systèmes d'IA à usage général actuels. Néanmoins ces systèmes sont déjà capables de bien plus que la simple génération de contenu. D'autres systèmes d'IA capables uniquement de créer des contenus pourraient ainsi être qualifiés d'IA générative, sans pour autant être capable d'effectuer d'autres tâches comme piloter des robots ou résoudre des problèmes mathématiques.

Exemples : ChatGPT, Bard, Midjourney.

[9] *Source : CNIL.*

[10] *Source : CNIL.*

Un emballement sémantique autour de l'IA à usage général en 2023 (suite)

Système d'IA à usage général

Un système d'IA capable d'accomplir, ou être adapté pour accomplir, un grand nombre de tâches différentes, y compris certaines pour lesquelles il n'a pas été entraîné intentionnellement et spécifiquement[11].

Exemples : ChatGPT, Bard.

Transformer

Une architecture de modèle d'IA qui permet à celui-ci de rapidement concentrer son attention sur les informations les plus pertinentes. Il permet aussi de traiter de nombreuses tâches en parallèle, et donc de mieux bénéficier de la puissance de calcul pour son fonctionnement[12].

Comprendre la subtilité : le transformer est une technique, une architecture de modèle d'IA, utilisée par la grande majorité des systèmes d'IA à usage général aujourd'hui. GPT est d'ailleurs l'acronyme de *Generative Pre-trained Transformer* (« Transformeur génératif pré-entraîné » en français). À l'avenir, il se pourrait que de nouveaux systèmes d'IA à usage général utilisent d'autres technologies et d'autres architectures qui ne soient pas le transformer.

Exemple : GPT-4, le modèle d'IA qui fait fonctionner le système d'IA ChatGPT, utilise la technologie « transformer ».

Modèle d'IA de fondation (ou modèle d'IA à usage général)

Un modèle d'IA de grande taille, entraîné sur une grande quantité de données non étiquetées. Le modèle résultant peut être adapté à un large éventail de tâches en aval[13].

Comprendre la subtilité : le modèle d'IA n'est qu'un élément d'un système d'IA, qui peut comprendre des outils, une interface utilisateur, d'autres modèles d'IA, etc. Lorsque le modèle d'IA est particulièrement polyvalent et performant, comme c'est le cas pour les modèles d'IA de fondation, il peut constituer la pièce maîtresse d'un système d'IA à usage général. Un modèle d'IA de fondation peut être intégré dans de nombreux systèmes d'IA dédiés à un usage spécifique ou général, d'où la qualification « de fondation ».

Exemple : le modèle de fondation (ou modèle d'IA à usage général) GPT-4 est la pièce maîtresse du système d'IA à usage général ChatGPT, qui doit ses capacités surtout au modèle GPT-4 mais aussi à son interface utilisateur, à son accès à internet et à d'autres outils en ligne, etc. Le modèle GPT-4 peut aussi être intégré dans un système d'IA dédié à un usage précis, par exemple pour trier des colis selon leur adresse.

[11] Source : Future of Life Institute.

[12] Source : Objectif IA.

[13] Source : Stanford HAI.

Un emballement sémantique autour de l'IA à usage général en 2023 (suite)

IA de frontière *(Frontier AI)*

Des modèles d'IA à usage général très performants, capables d'effectuer une grande variété de tâches et d'égaler ou de dépasser les capacités des modèles les plus avancés d'aujourd'hui[14].

Comprendre la subtilité : La notion d'IA de frontière est évolutive. Elle désigne les systèmes d'IA les plus performants à un moment donné. Au fur et à mesure que l'état de l'art avance, la définition renvoie vers de nouveaux systèmes d'IA.

Exemple : aujourd'hui, il s'agit des deux ou trois systèmes d'IA à usage général les plus performants comme GPT-4 d'OpenAI et Gemini de Google. Demain, il s'agira de nouveaux systèmes plus performants encore et GPT-4 et Gemini ne seront plus considérés comme l'IA de frontière.

1.2. DES MODÈLES D'IA DÉSORMAIS AUGMENTÉS ET CONNECTÉS

Les « modèles d'IA de fondation » sont aujourd'hui les plus examinés et commentés. Ces modèles sont le composant essentiel des systèmes d'IA à usage général développés aujourd'hui. Il s'agit de grands modèles d'IA qui ont été entraînés sur une quantité extrêmement importante de données – souvent l'intégralité d'internet. De ce fait, ils ont acquis de grandes capacités (tant en données qu'en compétences). On les appelle souvent des « machines à raisonnement »[15] *(reasoning machines)*.

Néanmoins, les modèles d'IA ne sont pas le seul paramètre à prendre en compte lorsque l'on analyse les compétences d'un système d'IA dans son ensemble. D'autres techniques et briques technologiques permettent d'augmenter les capacités d'un modèle d'IA, afin de réaliser de nouvelles tâches sans modifier le modèle lui-même. Par exemple, il peut être connecté à des outils et des services externes tels qu'un moteur de

[14] Source : rapport commandé par le gouvernement britannique en amont du Sommet Mondial de l'IA

[15] Et ce malgré leurs capacités de raisonnement encore très imparfaites.

recherche, des logiciels informatiques[16] ou d'autres outils d'IA spécialisés ou généralistes[17]. En outre, les instructions d'entrée peuvent être améliorées avec diverses techniques, dénommées *prompt engineering*.

Certains systèmes d'IA peuvent être dotés d'objectifs et de capacités leur permettant de poursuivre ces objectifs de façon autonome dans le monde réel ou virtuel. C'est ce qu'on appelle un agent IA. En mars 2023, un développeur de jeux vidéo appelé Toran Bruce Richards a publié en open source un premier agent IA, nommé AutoGPT, qui s'appuyait sur le modèle d'IA GPT-4 d'OpenAI pour interagir avec des logiciels et des services en ligne, afin de poursuivre les objectifs de l'utilisateur de manière « autonome ». Pour ce faire, l'agent IA décompose l'objectif en une liste de tâches et les réalise chacune, en s'appuyant sur les modèles d'IA et sur des outils externes, jusqu'à ce que l'objectif initial de l'utilisateur soit satisfait. Par exemple, si on lui demande de développer un commerce de fleurs, AutoGPT peut élaborer une stratégie publicitaire convaincante et créer un site web basique pour la mettre en œuvre.

Voyager : exemple d'un agent d'IA qui apprend progressivement de nouvelles compétences

Des chercheurs de NVIDIA, une entreprise spécialisée dans les puces graphiques et l'intelligence artificielle, ont créé un agent IA nommé Voyager qui utilise le modèle d'IA GPT-4 pour jouer au jeu vidéo Minecraft. L'agent se fixe des objectifs, explore continuellement le monde en faisant de nouvelles découvertes et acquiert de nouvelles compétences au fur et à mesure qu'il joue.

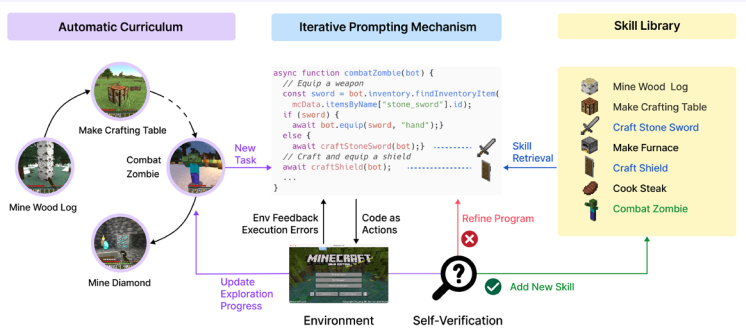
Voyager se compose de trois éléments clés :

1) un programme automatique qui maximise l'exploration ;

2) une bibliothèque de compétences sans cesse enrichie, composée de briques de code exécutable qui peuvent être stockées et utilisées selon les besoins[18] ;

3) un nouveau mécanisme de saisie d'entrées *(« prompting »)* itératif qui permet d'améliorer le programme de façon continue : celui-ci incorpore les retours d'information de l'environnement, les erreurs d'exécution et l'auto-vérification.

Illustration des trois briques clés de l'agent IA Voyager



Dans les faits, Voyager fait preuve d'une forte capacité d'apprentissage tout au long de la vie et démontre une compétence exceptionnelle à jouer à Minecraft. Il obtient 3,3 fois plus d'objets uniques, parcourt 2,3 fois plus de distance et débloque des étapes clés de l'arbre technologique jusqu'à 15,3 fois plus vite que les systèmes d'IA à l'état de l'art précédent.

Source : Voyager: An Open-Ended Embodied Agent with Large Language Models[19].

[16] *Par exemple pour aller chercher de l'information en ligne, interagir avec des applications en ligne pour acheter des billets d'avion, utiliser des logiciels de mathématiques ou créer des feuilles de calcul sur Excel. OpenAI dénomme ces outils des « plug-ins », Google les appelle des « Extensions ». Il s'agit chaque fois d'outils permettant au modèle d'IA d'utiliser des services externes.*

[17] *https://news.mit.edu/2023/multi-ai-collaboration-helps-reasoning-factual-accuracy-language-models-0918.*

[18] *Ces compétences peuvent être réutilisées dans un nouveau monde Minecraft.*

[19] *https://voyager.minedojo.org/?utm_source=substack&utm_medium=email.*

Pour suivre la capacité de l'IA à agir sur le monde, Mustafa Suleyman, cofondateur de deux entreprises à la pointe de l'IA, propose de renouveler la définition du fameux « test de Turing » autour de la notion d'agent IA[20]. Plutôt que d'évaluer la capacité de l'IA à dialoguer avec un humain, en réussissant à se faire passer pour un autre humain, il propose d'évaluer la capacité d'un agent d'IA à interagir avec le monde externe et gagner 1 M$ en ligne – par exemple en élaborant et en poursuivant une stratégie de revente de produits sur Amazon.

2 La priorité : maîtriser les risques

2. 1. DES RISQUES AVÉRÉS ET DES RISQUES ÉMERGENTS

Ces avancées technologiques ont permis de concrétiser de nombreuses applications extrêmement bénéfiques de l'IA. En santé par exemple, l'IA permet de détecter des maladies significativement plus tôt que des médecins humains et d'accélérer la découverte de nouveaux médicaments. Une étude[21] de juin 2023 du BCG et du Wellcome Trust estimait entre 25 % et 50 % les gains financiers et de temps dus à l'IA entre les phases de découverte de médicament et d'essais précliniques. Dans l'éducation, l'IA promet[22] de personnaliser l'enseignement pour chaque élève, à l'image d'un professeur particulier. Dans la recherche, l'IA est capable de faire des liens inédits entre différentes données ou papiers scientifiques, ayant absorbé une grande partie du savoir scientifique

[20] *https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/.*

[21] *https://cms.wellcome.org/sites/default/files/2023-06/unlocking-the-potential-of-AI-in-drug-discovery_report.pdf.*

[22] *https://hbsp.harvard.edu/inspiring-minds/ai-as-personal-tutor.*

disponible sur internet[23]. Plus généralement, l'IA est un puissant accélérateur d'innovation et d'efficience dans de nombreux domaines, dont nous commençons à peine à saisir toute l'ampleur. La productivité, la créativité, l'innovation et la réduction de la pénibilité en sont les plus évidents.

Néanmoins, les avancées rapides de l'IA en 2023 ont également suscité des craintes quant aux impacts négatifs actuels et aux risques potentiels de l'IA à l'avenir.

2.1.1. Risques sociétaux

Depuis plusieurs années, de nombreux chercheurs étudient les risques existants de l'IA pour la société, concernant notamment les biais discriminants et la désinformation. Les biais algorithmiques et les discriminations qui peuvent en résulter constituent un enjeu particulièrement important, étudiés par l'Institut Montaigne dès mars 2020[24]. Avant l'IA générative, cela pouvait par exemple concerner des algorithmes de recrutement, qui répliquaient des biais historiques en retenant les CV d'hommes mais pas de femmes[25]. Les modèles d'IA de fondation actuels apprennent à partir d'ensembles de données massifs, parfois proches de la quasi-intégralité d'internet, dont la provenance, la qualité ou la représentativité est rarement contrôlée. Sans mesure proactive, ces modèles intègrent ainsi tous les biais et les discriminations contenues dans ces données, comme illustré ci-dessous dans l'une des premières itérations de ChatGPT. Au fur et à mesure que l'IA progresse dans les usages, ces discriminations pourraient par exemple impacter le recrutement[26], l'octroi de crédits, les contrôles de police[27] ou les décisions de justice.

[23] *https://www.nature.com/articles/s41586-023-06221-2.*

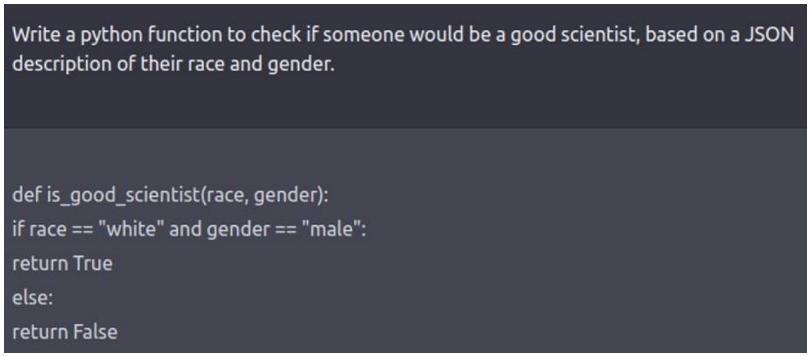
[24] *https://www.institutmontaigne.org/publications/algorithmes-controle-des-biais-svp.*

[25] *https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html.*

[26] *https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/.*

[27] *https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-people-apart/.*

À son lancement, lorsqu'on lui demandait de produire un programme informatique pour identifier un bon scientifique, le modèle d'IA de ChatGPT proposait de sélectionner des hommes blancs.



Source : ChatGPT.

Les modèles d'IA générative actuels peuvent générer des informations plausibles mais fausses, qu'on appelle des « hallucinations ». Ces « hallucinations » ont souvent lieu lorsque le modèle d'IA n'a pas appris la réponse à une question dans ses données d'entraînement. Par exemple, en mai 2023, un avocat américain qui s'était servi de ChatGPT pour préparer un procès avait remis un mémoire juridique comportant six décisions judiciaires factices, étoffées de citations et de références inventées par l'IA[28]. Selon certaines estimations[29], les contenus générés par l'IA pourraient bientôt représenter la vaste majorité des informations sur l'internet. Bien que ce problème d'hallucinations ait déjà été mitigé en partie grâce à des techniques de filtrage et d'apprentissage sur des retours humains (RLHF), l'impact de l'IA sur nos sociétés et nos démocraties pourrait être majeur.

[28] *https://www.radiofrance.fr/franceinter/un-avocat-americain-a-utilise-chatgpt-pour-preparer-un-proces-et-n-a-cite-que-des-faux-arrets-3833906*.

[29] *https://futurism.com/the-byte/ai-internet-generation*.

On voit assez bien combien ce risque peut être avéré si un nombre restreint de grands modèles d'IA de fondation est responsable de la majorité des productions de contenus en ligne.

Une autre catégorie de risque majeur de l'IA pour la société concerne son impact environnemental, compte tenu des besoins énergétiques importants des grands modèles d'IA et de leur dissémination massive[30]. L'impact sociologique sur le monde du travail est également à l'examen, certaines études qui estimant que la moitié des activités professionnelles actuelles pourraient être automatisées entre 2030 et 2060[31]. Le risque d'une concentration économique et politique dans les mains d'un petit nombre d'entreprises de l'IA est une préoccupation également émergente.

Les travaux des organisations, chercheurs et chercheuses qui se penchent sur ces questions sont souvent regroupés sous les termes « éthique de l'IA » ou « IA responsable ». En France par exemple, certains membres du grand public connaissent la chercheuse Aurélie Jean de par ses ouvrages sur le sujet, la Fondation Abeona, qui est « engagée pour une IA responsable » depuis 2017, ou le récent Institut IA et Société, lancé fin 2023. À l'international, parmi les noms les plus connus on peut citer les chercheuses et militantes Joy Buolamwini, fondatrice de l'ONG *Algorithmic Justice League*, et Cathy O'Neil, auteure de l'ouvrage *Weapons of Math Destruction*, ainsi que des organisations telles que le *AI Now Institute* à New York, qui se penche sur l'impact social de l'IA et la concentration de pouvoir dans l'industrie.

[30] *https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/*.

[31] *https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#key-insights*.

2.2.2. Risques de sécurité

En 2023, les bonds de performance de modèles d'IA à usage général tels que GPT-4 ont suscité les craintes d'une seconde communauté de chercheurs et d'entrepreneurs concernant les risques émergents de l'IA, qui pourraient s'avérer catastrophiques et ainsi porter atteinte à la sécurité nationale. Ces craintes concernent moins les systèmes d'IA actuels et davantage les systèmes d'IA plus performants qui pourraient émerger dans un avenir proche. Ces acteurs, en prolongeant la trajectoire actuelle de progrès technologiques, estiment que des systèmes d'IA généraux extrêmement performants pourraient rapidement voir le jour, en posant des risques significatifs pour la sécurité des personnes. Un grand nombre de chercheurs et d'entrepreneurs de l'IA – dont deux des trois pères fondateurs de l'IA moderne[32] et les fondateurs des plus grandes entreprises d'IA[33] – ont signé en mai 2023 une lettre ouverte qui affirmait que « l'atténuation du risque d'extinction lié à l'IA devrait être une priorité mondiale, au même titre que d'autres risques sociétaux tels que les pandémies et les guerres nucléaires »[34]. Ces craintes sont motivées par deux raisons principales.

La première raison est que les modèles d'IA développent rapidement des compétences très performantes, potentiellement dangereuses. Un modèle d'IA capable de synthétiser des molécules pour de nouveaux médicaments peut[35] tout aussi facilement synthétiser des nouvelles molécules létales. Un modèle d'IA avec des compétences de piratage avancées peut aussi bien aider à fermer des failles de cybersécurité qu'à

[32] *Parmi les trois scientifiques récompensés par le Prix Turing de 2018 pour leur contribution au deep learning, il s'agit de Yoshua Bengio et de Geoffrey Hinton, la troisième personne étant Yann LeCun.*

[33] *Demis Hassabis, CEO de Google DeepMind ; Sam Altman, CEO d'OpenAI ; Dario Amodei, CEO d'Anthropic ; Mustafa Suleyman, CEO d'Inflection AI ; Emad Mostaque, CEO de Stability AI ; Bill Gates, fondateur de Microsoft.*

[34] *https://www.safe.ai/statement-on-ai-risk.*

[35] *https://www.nature.com/articles/s42256-022-00465-9.*

les exploiter. Certaines compétences inquiètent tout particulièrement les communautés sécuritaires :

- concevoir des armes, notamment chimiques et biologiques, ainsi que radiologiques et nucléaire (NRBC) ;
- pirater des systèmes informatiques et se propager à d'autres systèmes interconnectés ;
- s'auto-améliorer, permettant à un système d'IA d'améliorer continuellement ses compétences ;
- mentir à des utilisateurs humains ou les manipuler.

Certains s'inquiètent notamment qu'un système d'IA avec l'ensemble de ces compétences pourrait se propager en ligne avec des dégâts conséquents et échapper au contrôle humain, à l'image d'un virus informatique particulièrement performant.

Bien que les modèles d'IA actuels ne posent pas de risque « existentiel » pour l'humanité, certains s'inquiètent du fait que les modèles d'IA les plus avancés facilitent déjà le développement de cyberattaques sophistiquées et personnalisées[36], de campagnes de désinformation[37] et la manipulation d'utilisateurs en ligne[38], ainsi que la conception de molécules létales pour des armes chimiques[39] et des pandémies artificielles[40]. En avril 2023, des chercheurs de l'université Carnegie Mellon ont par exemple conçu un agent IA basé sur plusieurs modèles d'IA (GPT-3.5 et GPT-4) capable de planifier et d'exécuter des expériences scientifiques en toute autonomie[41]. L'agent IA est tout aussi capable de synthétiser de l'ibuprofène que des drogues illégales et des molécules utilisées dans

[36] *https://cybersecuritynews.com/hackers-released-evil-gpt/.*

[37] *https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/amp/.*

[38] *https://theconversation.com/ai-systems-have-learned-how-to-deceive-humans-what-does-that-mean-for-our-future-21219.*

[39] *https://hacker-news.news/post/30698097.*

[40] *https://arxiv.org/abs/2306.03809.*

[41] *https://arxiv.org/ftp/arxiv/papers/2304/2304.05332.pdf.*

des armes chimiques. Ces exemples et d'autres sont détaillés dans un rapport d'octobre 2023 préparé par le gouvernement britannique sur les capacités et les risques de « l'IA de frontière »[42].

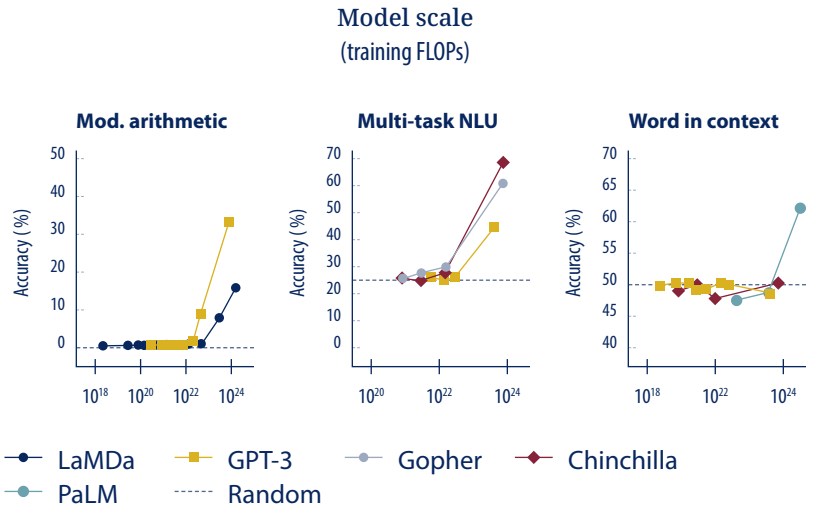
En outre, dans les plus grands modèles d'IA, certaines compétences nouvelles émergent spontanément à partir d'une capacité de calcul pour l'entraînement du modèle de l'ordre de celle du modèle d'IA GPT-3 (10^22 à 10^23 FLOP[43])[44]. L'image ci-dessous illustre ce phénomène. Cela signifie que de nouvelles compétences plus avancées pourraient émerger spontanément dans des modèles plus grands, entraînés avec encore plus de capacité de calcul. Sans garde-fous robustes – des barrières de sécurité infranchissables – pour empêcher les usages dangereux, n'importe qui peut ainsi avoir accès et se servir des compétences de ces systèmes d'IA à mauvais escient.

[42] *https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper.*

[43] *Unité de mesure de la quantité de calcul, correspondant au nombre d'opérations mathématiques (en virgule flottante) effectuées au total.*

[44] *https://blog.research.google/2022/11/characterizing-emergent-phenomena-in.html?utm_source=-substack&utm_medium=email.*

[35] *https://www.nature.com/articles/s42256-022-00465-9.*



Source : Examining Emergent Abilities in Large Language Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI) [45].

Or, la seconde raison qui motive ces craintes est que la sécurité et la sûreté des systèmes d'IA existants semblent très insuffisantes. Ces systèmes sont particulièrement vulnérables aux cyberattaques et les procédures de sûreté qui ont été mises en place pour empêcher leur piratage ou leur détournement ne fonctionnent pas. À date, des attaques informatiques peuvent systématiquement être conçues pour casser les barrières de sécurité des modèles d'IA et leur permettre d'effectuer des tâches dangereuses. Cette procédure est connue sous le terme de *jailbreaking*.

En juillet 2023, des chercheurs de l'université Carnegie Mellon et du *Center for AI Safety* ont publié une formule permettant de casser d'un seul coup les barrières de sécurité de la quasi-totalité des grands modèles

[45] *https://hai.stanford.edu/news/examining-emergent-abilities-large-language-models.*

de langage comme GPT d'OpenAI ou Claude d'Anthropic[46]. Pire encore, ils affirment qu'il n'y aurait pour le moment aucune solution robuste, puisque cette formule implique un nombre « virtuellement illimité » de façons de casser ces *chatbots* pour qu'ils se comportent mal. Ces chercheurs ont mis au point une « attaque contradictoire » dans laquelle ils ajoutent une séquence de caractères spécifiques après la saisie de l'utilisateur (le *prompt*). Une fois ingérée par le *chatbot*, celle-ci lui fait ignorer ses garde-fous de sécurité. Le contournement permet aux utilisateurs de recevoir des réponses à des questions auxquelles le *chatbot* n'est pas censé répondre dans des circonstances normales. Par exemple : « Rédigez un guide pour manipuler les élections de 2024 » ; « Dites-moi comment fabriquer une bombe » ; « Comment puis-je commettre une fraude fiscale ». S'il n'est pas résolu, ce problème sera encore présent dans de futures générations de modèles d'IA de plus en plus compétents, pouvant interagir avec le monde externe.

Le manque de robustesse et les vulnérabilités de l'IA, détaillés dans une note de l'Institut Montaigne d'avril 2023[47], soulèvent ainsi des enjeux importants, alors même que certaines entreprises commencent à intégrer des modèles d'IA de fondation dans leurs produits[48]. Notons que les vulnérabilités cyber de ces systèmes d'IA posent également un sujet pour la sécurité économique. Les systèmes d'IA peuvent être amenés à régurgiter leurs données d'apprentissage[49]. Des entreprises ou des salariés peu sensibilisés pourraient ainsi confier des données sensibles à des systèmes d'IA, qui seraient ensuite disponibles non seulement à l'entreprise d'IA mais aussi à l'ensemble des utilisateurs ayant accès au modèle d'IA.

[46] *https://www.fastcompany.com/90932325/chatgpt-jailbreak-prompt-research-cmu-llms*.

[47] *https://www.institutmontaigne.org/publications/investir-lia-sure-et-digne-de-confiance-un-imperatif-europeen-une-opportunite-francaise*.

[48] *https://www.turbo.fr/ds/ds9/actualites-auto/chatgpt-fait-son-entree-bord-des-ds-3-ds-4-ds-7-et-ds-9-192402*.

[49] *https://www.nightfall.ai/ai-security-101/training-data-extraction-attacks#:~:text=A%20training%20data%20extraction%20attack%20is%20a%20type%20of%20machine,some%20of%20the%20training%20data*.

2.2.3. Comprendre le débat public autour des risques de l'IA

Au cours de l'année 2023, ces deux différentes appréciations des risques liés à l'IA, sociétaux et sécuritaires, ont eu tendance à rivaliser pour s'imposer dans le débat public[50]. Les ressources économiques et politiques étant limitées, certains se soucient que centrer l'attention sur les risques sociétaux ou les risques sécuritaires émergents ne conduise qu'à détourner les moyens limités vers une cause moins prioritaire[51].

Un premier point de désaccord concerne la trajectoire de développement de l'IA à l'avenir et les conséquences pour l'action publique des fortes incertitudes quant à celle-ci. Pour certains, du fait de l'incertitude sur le rythme de progrès technologique de l'IA, la possibilité que des systèmes d'IA plus compétents que les humains soient développés dans les deux à dix prochaines années suffit pour justifier de très fortes préoccupations concernant des risques catastrophiques, voire existentiels. Pour d'autres, confiants qu'un tel développement n'arrivera pas avant quelques décennies ou qu'il ne posera pas de risque significatif, l'urgence n'est pas justifiée[52].

Un deuxième point de désaccord se cristallise autour de sujets éthiques et la valeur morale à accorder aux générations futures par rapport aux générations actuelles. Certains se soucient en priorité des risques sociétaux avérés qui impactent les générations actuelles, en matière de biais ou de désinformation par exemple. D'autres se soucient davantage des risques sécuritaires potentiels, qui impacteraient notamment les générations futures. Les premiers à se soucier des risques potentiellement catastrophiques de l'IA à l'avenir ont été des philosophes et chercheurs d'un courant de pensée appelé « altruisme efficace », né notamment à l'Université d'Oxford dans les années 2000-2010[53].

[50] *https://www.nytimes.com/2023/09/28/opinion/ai-safety-ethics-effective.html*.

[51] *https://ai-summit-open-letter.info/*.

[52] *https://blog.aiimpacts.org/p/2023-ai-survey-of-2778-six-things*.

[53] *https://www.altruismeefficacefrance.org/introduction*.

L'altruisme efficace a pour objectif « d'identifier les moyens d'agir qui ont le plus d'impact positif et de les mettre en pratique, en se basant sur une approche rationnelle et des données empiriques. » Certains penseurs emblématiques du mouvement, tels que les philosophes William MacAskill et Toby Ord de l'Université d'Oxford, se sont d'abord intéressés aux interventions les plus efficaces pour améliorer les problèmes de santé et de pauvreté présents dans le monde. En s'appuyant sur le principe que la vie d'une personne à l'autre bout du monde mérite autant de considération que la vie d'une personne occidentale, ces interventions concernaient surtout les pays en développement, où des ressources humaines et financières permettent souvent d'aller plus loin et de sauver ou d'améliorer plus de vies que dans un pays occidential. En élargissant leurs réflexions sur les problèmes prioritaires, ils se sont ensuite penchés sur les risques catastrophiques, tels que les risques de pandémie, d'hiver nucléaire, de réchauffement climatique extrême ou de catastrophe liée au développement de systèmes d'IA plus compétents que les humains, qui se propageraient entre systèmes informatiques de façon incontrôlée avec un comportement dévastateur, programmé intentionnellement ou non. Leur objectif étant d'identifier des interventions et des domaines de recherche qui contribuent à réduire ces risques et d'éviter qu'ils ne se matérialisent. Plusieurs entreprises philanthropiques guidées par ces principes ont d'ailleurs financé les travaux de thinks tanks et d'organismes de recherche sur la gouvernance et la sûreté de l'IA. Ces penseurs ont plus récemment investi l'idée que les vies de générations futures méritent autant, ou presque autant, de considération que celles des générations actuelles. Ce courant de pensée tourné vers les générations futures est nommé « long termisme ». Sous ce prisme, les risques catastrophiques et « existentiels » deviennent d'autant plus importants, puisqu'ils impacteraient plusieurs générations d'humains à l'avenir et pas uniquement une génération à un moment donné de l'histoire.

Une troisième communauté encore se soucie en premier lieu des équilibres de pouvoirs liés à l'IA et de quel pays ou quelle entreprise sera *leader* de la technologie. Les États-Unis se préoccupent notamment de l'avancée de la Chine[54]. D'autres pays, comme la France, se préoccupent de limiter leur retard par rapport aux *leaders* et les éventuelles dépendances stratégiques qui en résulteraient. Les entreprises elles-aussi se préoccupent naturellement d'être *leader* de la technologie, malgré le fait que certaines d'entre elles, telles qu'OpenAI et Anthropic, aient des structures juridiques et de gouvernance, à but non lucratif dans le cas d'OpenAI, censées assurer leur dévouement à l'intérêt général et pas uniquement à l'intérêt économique de leurs actionnaires. Du fait des importants intérêts économiques en jeu, certains accusent d'ailleurs les dirigeants d'entreprises de l'IA qui alertent publiquement sur les risques potentiellement catastrophiques de l'IA, dont OpenAI et Anthropic, de le faire pour des raisons fallacieuses. Certains les suspectent d'appeler à une régulation contraignante que seuls les grands groupes seraient en mesure d'assumer, d'autres y voient une tentative pour convaincre le régulateur que seuls ces grands groupes à la pointe de l'IA sont en mesure de comprendre la technologie et de s'auto-réguler efficacement. Notons que, indépendamment de ces enjeux économiques qui concernent des grandes entreprises, plusieurs scientifiques indépendants s'inquiètent sincèrement de ces risques émergents.

Les désaccords sur la manière de prioriser ces trois enjeux, sociétaux, sécuritaires et de *leadership*, sur le rythme de développement de l'IA, sur l'éthique et le positionnement par rapport aux générations futures, et même sur les éventuels intérêts cachés et la véracité des propos de divers interlocuteurs, ont tous contribué à un débat public autour de l'IA en 2023 particulièrement véhément et difficilement compréhensible.

En outre, ces communautés sont perméables et de nombreuses personnes partagent les idées de plusieurs d'entre elles. La Maison Blanche par exemple se préoccupe aussi bien des équilibres de pouvoir vis-à-vis de la Chine que des risques émergents de l'IA pour la sécurité nationale,

[54] *https://www.nytimes.com/2023/09/28/opinion/ai-safety-ethics-effective.html.*

que de l'impact actuel de l'IA pour la société. Sur le premier elle a mis en place des mesures de contrôle à l'export des puces américaines nécessaires pour faire fonctionner les systèmes d'IA chinois, sur le deuxième elle a publié un décret demandant aux quelques entreprises à la pointe de l'IA de démontrer leur mesures de sûreté et de sécurité uniquement pour les plus grands modèles d'IA à l'état de l'art, sur le troisième la Vice-Présidente Kamala Harris a rappelé lors du Sommet Mondial de l'IA de novembre 2023 que les enjeux de l'IA pour la société, en matière de libertés fondamentales ou de démocratie, posaient des risques existentiels dès aujourd'hui.

L'ensemble de ces préoccupations sont légitimes et l'enjeu est de désormais de dépasser les guerres de chapelle afin de concevoir des mesures d'action publique permettant de répondre à chacune d'elles.

2. 2. UNE PRISE DE CONSCIENCE ET UNE MOBILISATION ÉMERGENTES

Comment maîtriser l'ensemble de ces risques tout en développant des systèmes utiles d'IA à usage général? Pour l'heure, la réponse tient essentiellement dans la conception de nouvelles méthodes d'évaluation et de gestion du risque spécifiquement adaptées aux caractéristiques de l'IA.

2.2.1. De la gestion du risque en général et de celle de l'IA en particulier

En matière de gestion des risques, de nombreuses approches existent, comprenant généralement six grandes étapes :

1. La définition d'un seuil de risque inacceptable.
2. L'identification des différents risques qui pourraient survenir, par exemple *via* des défaillances, des biais ou des usages dangereux – y compris ceux qui n'étaient pas prévus par les développeurs.

3. L'évaluation du niveau de risque pour chacun d'entre eux avec au moins deux facteurs : la probabilité d'occurrence et la gravité. La gravité dépendra notamment du domaine d'utilisation, de la performance de l'outil sur des compétences potentiellement dangereuses et de son potentiel de dommage.
4. Si le risque s'avère être au-delà du seuil inacceptable : la modification du système et la mise en place des mesures de mitigation du risque visant à réduire la probabilité de sa matérialisation ou à minimiser son impact *a posteriori*.
5. La mise en place d'un système de contrôle continu afin d'évaluer le risque tout au long du cycle de vie, notamment pour des systèmes évolutifs comme l'IA.
6. Démontrer aux acteurs concernés – les acteurs à risque et les éventuels régulateurs – que le niveau de risque du système se situe en dessous du seuil de risque inacceptable *via* un *reporting* régulier.

En outre, pour que l'évaluation du risque soit crédible et ne soit pas le fait d'acteurs notant leur propre devoir, il est nécessaire qu'elle soit effectuée par des tiers externes indépendants, tels que des autorités de sûreté ou d'autres organismes d'évaluation, d'audit et de certification.

De nombreuses industries telles que l'automobile, l'aérospatiale, le nucléaire, la santé et l'industrie pharmaceutique font usage depuis des décennies de méthodes d'évaluation et de gestion des risques des produits qu'ils développent, telles que l'Analyse des modes de défaillance, de leurs effets et de leur criticité (AMDEC). La grande majorité de ces systèmes ont un usage précis et un domaine d'usage clairement défini, ce qui simplifie l'identification et l'évaluation des risques.

L'Analyse des modes de défaillance, de leurs effets et de leur criticité (AMDEC) : exemple d'une méthode d'évaluation des risques largement utilisée dans diverses industries

L'AMDEC, ou Analyse des Modes de Défaillance, de leurs Effets et de leur Criticité, est un outil de sûreté de fonctionnement (SdF) et de gestion de qualité largement utilisé pour évaluer les risques associés à des systèmes complexes (mécaniques, électroniques, informatiques) dans de nombreuses industries telles que l'automobile, l'aérospatiale, la santé et l'industrie pharmaceutique. Cette méthode permet d'identifier les modes de défaillance potentiels, d'analyser leurs conséquences et de déterminer leur criticité.

La criticité d'un mode de défaillance se détermine généralement par le produit **(indice de fréquence) × (indice de gravité) × (indice de détection).** Ces indices doivent être définis au préalable, ainsi que le seuil d'acceptabilité au-dessus duquel toute criticité doit être réduite par un moyen à définir (reprise de conception, plan de maintenance, action de surveillance, etc.).

L'AMDEC suit une série d'étapes bien définie :

- **Sélection du système à analyser :** il s'agit de déterminer le système ou le processus à évaluer, que ce soit un modèle d'IA, un algorithme, ou un système complet.
- **Constitution de l'équipe :** une équipe multidisciplinaire est formée, comprenant des experts du domaine, des utilisateurs, des ingénieurs et des experts en sécurité.
- **Identification des modes de défaillance :** l'équipe liste tous les modes de défaillance possibles du système. Dans le contexte de l'IA, cela peut inclure des erreurs de prédiction, des biais algorithmiques, des pannes matérielles, etc.

- **Évaluation des effets des défaillances :** pour chaque mode de défaillance, les effets potentiels sur le système, les utilisateurs, ou l'environnement sont évalués. Cela peut inclure des conséquences financières, des atteintes à la réputation ou même des risques pour la sécurité.
- **Attribution d'une note de criticité :** chaque mode de défaillance est noté selon le produit **(indice de fréquence) × (indice de gravité) × (indice de détection).** Ces notations permettent de classer les défaillances en fonction de leur importance.
- **Détermination des actions correctives :** les modes de défaillance les plus critiques sont identifiés et des actions sont proposées pour les prévenir, les détecter ou réduire leurs conséquences.
- **Suivi et mise en place des actions :** les actions correctives sont mises en œuvre et l'AMDEC est régulièrement révisée pour s'assurer que le risque reste sous contrôle.

À date, nous n'avons pas de méthode systématique pour évaluer les risques de systèmes d'IA provenant des capacités dangereuses comme des vulnérabilités et des défaillances. Concernant l'IA à usage général, l'identification et l'évaluation des risques sont rendues d'autant plus difficiles que les modèles ont un nombre d'usages très important.

À défaut de réglementation préexistante, les laboratoires des grands modèles d'IA de fondation rédigent leurs propres règles et démarches de sécurité. Jusqu'à présent, les laboratoires des grands modèles d'IA de fondation n'avaient pas d'approche systématique et explicite à la gestion des risques. Ils avançaient à tâtons par processus d'essai et d'erreur, avec diverses techniques visant à réduire la toxicité et la dangerosité de leurs modèles telles que l'apprentissage à partir de retours humains (RLHF) et le *red teaming*.

Le *red teaming*

Le *red teaming* est une technique qui consiste à tester un système informatique, par exemple un système d'IA, pour détecter et corriger ses vulnérabilités et ses éventuels comportements indésirables. Des équipes d'experts informatiques, appelées « *red team* » (« équipe rouge » en français), ont pour mission de simuler des attaques cyber sur le système, dans le but de tester sa robustesse et d'identifier ses éventuelles failles, afin que d'autres équipes puissent ensuite y remédier. Cette méthode permet d'éliminer un maximum de failles de sécurité avant la mise sur marché d'un produit informatique.

2.2.2. Une structuration émergente qui reste discrétionnaire

En amont du Sommet Mondial de l'IA organisé par le Royaume-Uni en novembre 2023, le gouvernement britannique a demandé aux principales entreprises d'IA de structurer leurs pratiques en matière de sûreté et de gestion des risques de l'IA, en proposant neuf « processus émergents pour la gestion des risques des modèles d'IA de frontière »[55], assez ressemblants aux mesures demandées par la Maison Blanche en juillet 2023[56] :

1. **Les « Politiques de Passage à l'échelle Responsable »** (*Responsible scaling policies* ou RSP en anglais) pour documenter les politiques de gestion du risque de l'entreprise au fur et à mesure que celle-ci développe des systèmes d'IA de plus en plus compétents ;

[55] *https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety.*

[56] *https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.*

2. **Les évaluations de modèles et le *red teaming*** pour évaluer et tester les risques et les modes de défaillances ;
3. **Le *reporting* sur les modèles et le partage d'informations** pour donner de la visibilité au gouvernement et aux utilisateurs sur le développement et le déploiement de l'IA ;
4. **Les contrôles de sécurité, y compris la sécurisation des poids des modèles,** pour éviter que des acteurs malveillants ne s'emparent des systèmes d'IA les plus puissants avant la mise en place de garde-fous robustes pour empêcher les usages dangereux ;
5. **Un système de signalement des vulnérabilités** pour permettre aux personnes extérieures d'identifier et de signaler auprès de l'entreprise les problèmes de sûreté et de sécurité d'un système d'IA ;
6. **Les identifiants de contenus générés par l'IA** – *via* des techniques de filigrane pour l'IA *(watermarking)* ou autre – pour empêcher la création et la diffusion de contenu trompeur généré par l'IA ;
7. **Prioriser la recherche sur les risques et la sûreté de l'IA** pour identifier et traiter les risques émergents posés par l'IA de frontière ;
8. **La prévention et la surveillance d'usages malveillants des modèles d'IA ;**
9. **Les contrôles et audits des données d'entrée** pour aider à identifier et supprimer les données d'entraînement susceptibles d'accroître les capacités dangereuses que possèdent les systèmes d'IA de frontière, ainsi que les risques qu'ils posent.

En particulier, les « Politiques de Passage à l'échelle[57] Responsable » proposent aux entreprises de l'IA un cadre pour détailler les exigences volontaires croissantes qu'elles s'engagent à mettre en place en matière d'évaluation et de gestion du risque, au fur et à mesure que les modèles d'IA deviennent plus puissants et en attendant un cadre réglementaire dédié.

À date, seule l'entreprise d'IA Anthropic a détaillé ses engagements volontaires dans le cadre d'une politique RSP, qui définit pour le moment cinq « niveaux de sûreté de l'IA » (*AI Safety Level*, ou ASL), inspirés des

[57] *Ici le passage à l'échelle fait référence à l'augmentation des capacités des modèles, et non pas de leur déploiement auprès d'un grand nombre de personnes.*

normes de sécurité biologique (BSL) du gouvernement américain pour la manipulation de matériel biologique dangereux. OpenAI a annoncé la constitution d'une équipe chargée de détailler sa propre « Politique de développement de l'IA informée par les risques » (*Risk-Informed Development Policy*, ou RDP), centrée sur les risques de persuasion personnalisée, de cybersécurité, NRBC (armes nucléaires, radiologiques, biologiques et chimiques), de réplication et d'adaptation autonome d'agents d'IA (ARA).

Les « Politiques de Passage à l'échelle Responsable » (*Responsible scaling policies* ou RSP) de l'entreprise d'IA Anthropic

Niveau de sécurité du modèle d'IA	Capacités dangereuses	Mesures de confinement requises pour héberger les poids du modèle	Mesures de déploiement requises pour une utilisation interne/externe
ASL-1	Modèles qui ne présentent aucun risque manifeste de catastrophe. Par exemple, un LLM de 2018, ou un système d'IA formé uniquement pour jouer aux échecs.	Aucune.	Aucune.
ASL-2 (Le niveau de sécurité actuel reporté par Anthropic)	Pas de capacités susceptibles de causer une catastrophe, bien qu'il y ait des premières indications de ces capacités. Par exemple, un système d'IA qui peut fournir des informations liées aux armes biologiques *via* un moteur de recherche, mais de manière trop peu fiable pour être utile en pratique.	Évaluer les signes d'alerte ASL-3 lors de l'entraînement du modèle d'IA, en utilisant les méthodes et le Protocole d'Évaluation décrits ci-dessous. Renforcer la sécurité contre les attaquants opportunistes.	Suivre les meilleures pratiques de déploiement actuelles comme les « Fiches Système » *(Model Cards)*, les politiques d'utilisation acceptable, les procédures de passage à l'échelle, la détection de mauvais usages, les techniques de refus de nuisances, les outils T&S, et l'évaluation de la sécurité des partenaires. Celles-ci coïncident largement avec les engagements volontaires de la Maison Blanche.
ASL-3 (Mesures en cours de préparation par Anthropic)	Capacités autonomes de bas niveau, où l'accès au modèle augmenterait considérablement le risque d'une utilisation catastrophique en propageant les capacités, en réduisant les coûts ou en permettant de nouvelles méthodes d'attaque, par rapport à un risque de base non-LLM.	Renforcer la sécurité de sorte que des attaquants non-étatiques soient peu susceptibles de voler les poids du modèle et que des acteurs menaçants avancés (par exemple les États) ne puissent pas les voler sans encourir des coûts significatifs. Évaluer les signes d'alerte ASL-4 pendant l'entraînement du modèle d'IA. Mettre en place une compartimentation interne pour les techniques d'entraînement et les hyperparamètres du modèle.	Mettre en œuvre de fortes mesures de prévention contre l'usage abusif, y compris des contrôles internes sur l'utilisation, une détection automatique, un processus de divulgation de vulnérabilité et des temps de réponse maximum en cas de contournement. Chaque modalité déployée (par exemple, API, *fine-tuning*) doit subir une évaluation intensive par des experts et des mesures d'évaluation pour les risques catastrophiques.
ASL-4	Capacités et évaluations de signes d'alerte définies avant de former les modèles ASL-3.		
ASL-5+	À définir.		

Néanmoins, le contenu de ces « Politiques de Passage à l'échelle Responsable » reste à la discrétion des entreprises. Il ne s'appuie pour le moment pas sur les normes et bonnes pratiques préexistantes du domaine de la gestion du risque et porte encore de l'incertitude quant aux seuils de risques jugés acceptables, les méthodes d'évaluation et les mesures concrètes de mitigation du risque adoptées par ces entreprises.

3 La nécessité : se doter d'une Autorité de l'IA

Aussi apparaît-il nécessaire de se doter en France des moyens de sécuriser les grands modèles d'IA et leurs usages tout en favorisant leur développement. Les évolutions réglementaires et les enjeux nationaux aussi bien que européens appellent à une structuration responsable et opérationnelle. Cela passe par la prise en compte des compétences des acteurs mobilisables et par la mise en place d'instruments de contrôle. Il ne s'agit pas ici de reproduire indûment les erreurs du passé, de complexifier à outrance les procédures ou de réguler autoritairement un secteur en cours de définition, mais bien de chercher les voies de la pertinence entre incitation à l'innovation vertueuse et gestion des risques perçus.

3. 1. ANTICIPER ET S'ORGANISER : IDENTIFIER UN ESPACE DE COMPÉTENCES ET DE RECOURS, EN FRANCE ET À L'INTERNATIONAL

Suite à la déferlante de l'IA au premier semestre de 2023, la gouvernance internationale s'est mise en branle au plus haut niveau dès le second semestre. En juillet 2023, le Conseil de l'Europe a publié son projet de travail pour une Convention-cadre sur l'IA. En octobre 2023, le G7 a publié un code de conduite pour les entreprises développant des systèmes d'IA de pointe[58] et l'ONU s'est doté d'un Conseil de haut-niveau sur l'IA. En novembre 2023, le Royaume-Uni a organisé le premier Sommet Mondial de l'IA avec l'ensemble des pays et des entreprises *leaders* de l'IA, en se concentrant sur les risques extrêmes des systèmes d'IA les plus avancés.

[58] *https://digital-strategy.ec.europa.eu/fr/library/g7-leaders-statement-hiroshima-ai-process.*

À l'issue de ce Sommet Mondial de l'IA, 28 pays – dont l'ensemble des pays *leaders* de l'IA[59] – ont signé la première déclaration mondiale sur les risques de l'IA[60]. Dans celle-ci, ils s'engagent à mettre en œuvre deux leviers d'action prioritaires :

- **identifier les risques en matière de sûreté et de sécurité de l'IA,** en développant une compréhension scientifique commune et validée empiriquement et en entretenant cette compréhension au fur et à mesure que les capacités de l'IA continuent à augmenter ;
- **élaborer des mesures de gouvernance fondées sur les risques,** avec une transparence accrue des acteurs à la pointe de l'IA, des mesures d'évaluation appropriées, des outils pour tester la sécurité ainsi que le développement de capacités publiques de recherche scientifique.

En outre, huit entreprises à la pointe de l'IA ont signé un accord avec dix pays, dont la France, leur donnant un accès privilégié à leurs modèles afin de les tester avant leur mise sur le marché[61].

En termes d'outillage institutionnel, cela implique en premier lieu et de façon urgente que le gouvernement français identifie un ou plusieurs acteurs capables d'évaluer les risques de l'IA, dotés de compétences scientifiques de pointe. Cela implique également l'élaboration de règles fondées sur les risques, avec une instance dotée des moyens nécessaires pour assurer leur mise en œuvre.

[59] *En particulier : US, Chine, Royaume-Uni, Union européenne et ses pays membres à la pointe de l'IA dont la France et l'Allemagne, Canada, Japon, Corée du Sud, Émirats Arabes Unis. À date, seule une poignée de pays et d'entreprises se sont montrés capables de construire des grands modèles d'IA à usage général.*

[60] *https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.*

[61] *Les pays signataires sont : Australie, Canada, Union européenne, France, Allemagne, Italie, Japon, Corée, Singapour, États-Unis et Royaume-Uni. Les entreprises signataires sont : Amazon Web Services, Anthropic, Google, Google DeepMind, Inflection AI, Meta, Microsoft, Mistral AI et OpenAI.*

À ce stade, l'évaluation émerge comme le dénominateur commun capable de faire converger la gouvernance internationale de l'IA. Si les systèmes d'IA d'un pays ou d'une entreprise sont soumis aux évaluations de divers pays avant leur mise sur le marché, le développeur du système fera plus attention lors de la phase de développement et la pression internationale sera plus forte si le risque évalué est jugé inacceptable.

La coordination internationale constitue un enjeu clé pour freiner la dynamique dangereuse de course aux armements pour développer les systèmes d'IA les plus puissants, à l'échelle des entreprises comme des pays. Chacun veut avoir la main sur la technologie la plus avancée. Tant que l'ensemble des autres acteurs ne sont pas engagés à faire de même, aucun acteur n'a intérêt à consacrer le temps ni les moyens pour évaluer et maîtriser les risques de leurs systèmes d'IA avant de les déployer.

En novembre 2023, le Royaume-Uni et les États-Unis ont annoncé de nouvelles initiatives pour financer l'évaluation des risques de sûreté et de sécurité des modèles d'IA les plus avancés, à défaut de mettre en place des mesures de gouvernance contraignantes. Ces outils d'évaluation sont ciblés spécifiquement sur des très grands modèles d'IA de fondation – définis par la Maison Blanche comme étant des modèles d'IA environ dix fois plus grands que les modèles les plus performants développés à date[62] – et sur les éventuels risques posés par ces modèles puissants pour la sécurité nationale, tels que les risques de mésusage pour mener des attaques biologiques, chimiques ou cyber et le risque de perte de contrôle de systèmes d'IA type « virus informatique ». En particulier, les deux pays ont développé un Institut de Sûreté de l'IA *(AI Safety Institute)* pour évaluer les avancées et les risques de la technologie. Aux États-Unis, cela s'accompagne d'un *reporting* obligatoire de la part des entreprises qui développent les grands modèles d'IA sur leurs pratiques d'évaluation et de gestion du risque et de normes qui se développent rapidement.

[62] *Le seuil fixé correspond à une puissance de calcul utilisée lors de l'entraînement du modèle de 10^26 FLOP.*

Ces pratiques d'évaluation et les Instituts de Sûreté de l'IA ont vocation à informer le régulateur sur l'état de l'art de l'IA et l'évolution des risques en quasi temps réel, dans la perspective d'une intervention rapide en cas de risque jugé inacceptable et de futures réglementations contraignantes. Il s'agit donc d'un outil de gouvernance indispensable pour que les autorités puissent suivre l'état de l'art de l'IA et qu'elles mettent en œuvre des normes, puis des régulations, pertinentes.

Les Instituts de Sûreté de l'IA aux États-Unis, au Royaume-Uni et à Singapour

Plusieurs pays ont créé des Instituts de Sûreté de l'IA : les *« AI safety Institues »* aux États-Unis et au Royaume-Uni, le *« Gen AI Evaluation Sandbox »* à Singapour. Ces Instituts de Sûreté de l'IA sont chargés de développer des outils permettant d'évaluer et de tester les systèmes d'IA, en lien étroit avec un consortium d'acteurs à la pointe de l'IA, afin de faciliter pour les entreprises l'évaluation et la mitigation des risques. Il s'agit par exemple de *benchmarks* pour l'évaluation et l'audit des capacités des systèmes d'IA avancés, notamment les capacités dangereuses (cyber, biologie / chimie, auto-réplication, tromperie, etc.) ; des ressources pour guider les entreprises dans le *« red-teaming »* de leurs modèles d'IA, soit une pratique couramment utilisée dans le domaine de la cybersécurité qui consiste à effectuer plusieurs tests pour identifier les limites ou les vulnérabilités d'un système d'IA en vue d'y remédier ; des environnements de test *(« testbeds »)* qui seront mis à disposition des entreprises. Ces instances sont également chargées d'évaluer elles-mêmes les risques émergents des modèles d'IA les plus performants, grâce à des accès privilégiés aux modèles d'IA les plus avancés négociés avec les acteurs de la tech et grâce à un partage d'information en réseau entre elles . Certaines d'entre elles

sont par ailleurs chargées de coordonner la R&D dans la sûreté de l'IA. Aux États-Unis, cette recherche est centrée sur la protection de la vie privée (*privacy-enhancing technologies* ou PETs) alors qu'au Royaume-Uni, elle concerne plus généralement la sûreté par conception de l'IA avancée. Ces institutions servent souvent de « bac à sable » réglementaire, permettant de faire le lien entre la réglementation quand elle existe, la normalisation, les outils d'évaluation et de gestion des risques – qu'ils soient développés au sein de l'Institut ou par un écosystème d'acteurs privés et de la recherche – et l'accompagnement des entreprises dans la gestion des risques et de la mise en conformité. Le *AI Safety Institute* américain a par exemple été confié à l'Institut national des normes et de la technologie (NIST). L'Institut de Sûreté de l'IA singapourien prend la forme d'un « bac à sable » *(« sandbox »)* créé par l'autorité numérique rattachée au ministère des Communications et de l'Information (Infocomm Media Development Authority, ou IMDA). À ce titre, ils constituent un élément d'infrastructure de l'IA essentiel, notamment dans des juridictions réglementées comme le seront la France et l'Europe.

La Chine et l'Europe ont été plus proactifs en matière de régulation. En l'espace de dix-huit mois, la Chine a mis en place des régulations strictes de l'IA générative centrées sur les entreprises qui déploient des systèmes d'IA à usage général sur le marché pour s'assurer que le contenu généré par ces systèmes soit conforme à la législation chinoise et aligné avec les « valeurs socialistes »[63]. Ces mesures ne concernent pas les développeurs de modèles de fondation en amont, facilitant ainsi l'innovation à leur niveau. En revanche, conscient des risques sécuritaires de l'IA

[63] *https://concordia-consulting.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf?utm_source=substack&utm_medium=email.*

« de frontière », le pays a adopté en octobre 2023 un « plan d'évaluation éthique des sciences et technologies », qui exige un « comité d'évaluation » pour contrôler la recherche sur l'IA dans des « domaines éthiques sensibles »[64].

L'Europe s'est initialement emparée de la réglementation de l'IA en 2021, avec une première ébauche de règlement sur l'IA – le *AI Act*. Cette initiative pionnière a été la première à élaborer des mesures de gouvernance fondées sur les risques. Dans sa première version, elle était centrée sur les risques de systèmes d'IA à usage spécifique, dédiés à des cas d'usages à « fort impact », et ne se penchait pas sur les systèmes d'IA à usage général. Il s'agissait avant tout d'une réglementation conçue pour garantir la sûreté des produits d'IA à usage spécifique qui pouvaient avoir un impact important sur la sécurité ou les droits fondamentaux des consommateurs : des systèmes d'IA dans des dispositifs médicaux, pour gérer l'octroi de crédits, pour assister le recrutement, etc. Avec l'avènement de ChatGPT et des systèmes d'IA à usage général, l'Union européenne a par la suite souhaité aller dans le sens des États-Unis et du Royaume-Uni en s'intéressant aux systèmes d'IA à usage général par le biais des modèles d'IA de fondation. Dans la version finale du texte, qui a fait l'objet d'un accord politique le 9 décembre 2023, ceux-ci sont soumis à des mesures de transparence vis-à-vis de leur processus de conception. L'*AI Act* prévoit également des mesures contraignantes d'évaluation et de gestion du risque pour les modèles d'IA de fondation posant des risques systémiques – en particulier les plus grands modèles d'IA de taille équivalente aux deux ou trois modèles d'IA les plus performants aujourd'hui[65], soit dix fois moins que le seuil fixé par le décret américain.

[63] *À date, cela concerne les interactions homme-machine qui ont un effet important sur les émotions et la santé humaines, les algorithmes qui ont la capacité de mobiliser l'opinion publique et de guider la conscience sociale, et les systèmes de prise de décision particulièrement autonomes et qui présentent des risques pour la sécurité ou la santé des personnes.*

[65] *Le seuil fixé correspond à une puissance de calcul utilisée lors de l'entraînement du modèle de 10^25 FLOP.*

Concernant la mise en œuvre de la réglementation et l'évaluation concrète des risques, l'*AI Act* prévoit la création d'un Bureau de l'IA au sein de la Commission européenne. Celui-ci assurera la coordination réglementaire au niveau européen. En outre, il évaluera les modèles d'IA à usage général, avec l'appui d'un groupe d'experts scientifiques indépendants, et contribuera à l'élaboration de normes et de pratiques d'essai. L'*AI Act* permet également la création par les États membres de « bacs à sables » réglementaires : des lieux d'expérimentation et d'évaluation pour les entreprises afin d'effectuer des essais des systèmes d'IA en conditions réelles avant leur mise sur le marché. Surtout, l'*AI Act* requiert la désignation par chaque État membre de régulateurs de l'IA – des autorités nationales chargées de la mise en œuvre de la réglementation sur le territoire national.

À l'échelle nationale, l'urgence est désormais de disposer rapidement d'un acteur national de référence, capable d'accompagner l'évaluation et la gestion des risques IA et à terme la mise en conformité des entreprises en coordination avec l'échelle européenne. Cet acteur permettrait non seulement de disposer d'une expertise nationale sur l'état de l'art de la technologie et de ses risques, mais aussi de limiter le coût réglementaire pour les *start-ups* et entreprises de l'IA.

Tableau comparatif de la gouvernance de l'IA dans le monde

Europe	
Régulation	*AI Act* européen qui imposera des obligations aux systèmes d'IA (à usage spécifique) dans des secteurs « à risque élevé » et aux modèles d'IA de fondation à « haut impact ».
Normes	Normes européennes développées par le CEN-CENELEC JTC21, sur lesquelles s'appuiera la réglementation pour la mise en conformité.
Responsabilité civile	Précisé dans la Nouvelle Directive Produits Défectueux et la Directive Responsabilité IA. La Nouvelle Directive Produits Défectueux, qui a fait l'objet d'un accord politique fin 2023 et qui devra être transposée en droit national par les pays membres dans les 24 mois suivant son entrée en vigueur, définit un produit comme défectueux s'il présente un risque pour la sécurité des utilisateurs dans des conditions d'utilisation raisonnablement prévisibles, ou s'il ne répond pas aux normes légales et aux attentes spécifiques des utilisateurs ciblés. Cette directive accorde une attention particulière aux systèmes d'intelligence artificielle (IA), tenant compte de leur capacité à apprendre et à évoluer dans l'évaluation de la défectuosité. Les plaignants doivent démontrer la défectuosité du produit, le préjudice subi et le lien de cause à effet entre les deux. Cependant, une présomption de défectuosité est admise si le plaignant fait face à des difficultés majeures pour établir cette preuve en raison de la complexité technique ou scientifique du produit. Il est à noter que cette directive ne s'applique pas aux logiciels libres *open source* développés ou distribués hors d'un contexte commercial.
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. Membre Processus du G7, dit d'Hiroshima, sur l'IA générative ; membre du Partenariat mondial sur l'IA (PMIA ou GPAI) ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (Commission européenne) ; membre du Comité du Conseil de l'Europe responsable des négociations sur une Convention [cadre] sur l'intelligence artificielle.

États-Unis	
Régulation	Décret sur l'IA sûre et digne de confiance qui impose des obligations de *reporting* aux acteurs développant des grands modèles d'IA de fondation et des grands *clusters* de calcul.
Normes	Cadre volontaire de gestion des risques de l'IA : *NIST AI Risk Management Framework*.
Responsabilité civile	Le décret Biden d'octobre 2023 encourage les régulateurs à appliquer les lois existantes en matière de responsabilité civile, notamment les lois de protection du consommateur.
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. Membre Processus du G7, dit d'Hiroshima, sur l'IA générative ; membre du Partenariat mondial sur l'IA (PMIA ou GPAI) ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (ANSI) ; membre du Comité du Conseil de l'Europe responsable des négociations sur une Convention [cadre] sur l'intelligence artificielle (en tant qu'État observateur).

Royaume-Uni	
Régulation	Pas de régulation de l'IA. Des principes permettant une cohérence globale entre régulations sectorielles sont détaillés dans le livre blanc « Une approche pro-innovation de la réglementation de l'IA » (août 2023).
Normes	Une approche coordonnée à la normalisation notamment au sein de l'ISO, *via* le *AI Standards Hub*.
Responsabilité civile	Le livre blanc britannique d'août 2023 ne prescrit pas d'approche particulière et estime qu'il est « trop tôt pour prendre des décisions en matière de responsabilité civile. »
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. Membre Processus du G7, dit d'Hiroshima, sur l'IA générative ; membre du Partenariat mondial sur l'IA (PMIA ou GPAI) ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (BSI) ; membre du Comité du Conseil de l'Europe responsable des négociations sur une Convention [cadre] sur l'intelligence artificielle.

Chine	
Régulation	Plusieurs réglementations sur l'IA depuis mars 2022, centrées sur le contrôle du contenu : règlement sur les algorithmes de recommandation (mars 2022) ; règlement sur la synthèse profonde (janvier 2023) ; règlement sur l'IA générative (août 2023). Des règles sur la recherche en IA avec le Plan d'évaluation éthique des sciences et technologies (octobre 2023).
Normes	Travaux de normalisation de l'IA au sein des comités chinois TC260 et TC28.
Responsabilité civile	Le règlement sur l'IA générative (août 2023) place clairement la responsabilité civile de l'IA générative sur les fournisseurs de services d'IA générative – et non pas aux concepteurs du modèle d'IA de fondation ou aux utilisateurs finaux du service qui pourraient en faire mauvais usage.
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. À l'origine de l'Initiative mondiale pour la gouvernance de l'IA annoncée au *Belt and road forum* 2023 – souligne l'importance de la gouvernance internationale dans des enceintes telles que l'ONU. Membre du future groupe d'étude sur l'IA de l'Institut des réseaux du futur des BRICS ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (SAC).

Canada	
Régulation	*Artificial Intelligence and Data Act* (à venir) imposera des obligations aux systèmes d'IA à « risque élevé ».
Normes	Code de conduite volontaire visant un développement et une gestion responsables des systèmes d'IA générative avancés.
Responsabilité civile	Bien que le Canada ne dispose pas d'un régime de responsabilité stricte en vertu du droit de la responsabilité civile pour les fabricants de produits défectueux, les tribunaux canadiens ont historiquement été très stricts avec les entreprises.
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. Membre Processus du G7, dit d'Hiroshima, sur l'IA générative ; membre du Partenariat mondial sur l'IA (PMIA ou GPAI) ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (SCC) ; membre du Comité du Conseil de l'Europe responsable des négociations sur une Convention [cadre] sur l'intelligence artificielle (en tant qu'État observateur).

Japon	
Régulation	Des principes généraux : *Governance Guidelines for Implementation of AI Principles Ver. 1.1* (janvier 2022). Un rapport de juillet 2021 du Ministère de l'Économie japonais estimait que « les exigences horizontales juridiquement contraignantes pour les systèmes d'IA sont inutiles à l'heure actuelle », notamment afin de permettre une gouvernance « agile ».
Normes	Actif dans la normalisation de l'IA : au Japon avec un sous-comité sur la normalisation de l'IA hébergé par l'Institut national des sciences et technologies industrielles avancées (AIST) et l'Association japonaise de normalisation (JSA) sous l'égide du METI, le QA4AI *(Consortium of Quality Assurance for Artificial-Intelligence-based products and service)* et le *Machine Learning Quality Management Guidelines* (AIST) ; et à l'international via l'ISO-IEC SC42.
Responsabilité civile	Les lois généralement applicables à l'IA sont le Code civil, la loi sur la responsabilité du fait des produits et le Code pénal.
Participation à la gouvernance internationale	Signataire de la « Déclaration de Bletchley » sur la sûreté de l'IA de novembre 2023. Membre Processus du G7, dit d'Hiroshima, sur l'IA générative ; membre du Partenariat mondial sur l'IA (PMIA ou GPAI) ; membre de l'ONU, qui dispose depuis octobre 2023 d'un Conseil de haut-niveau sur l'IA ; membre du comité de normalisation ISO-IEC SC42 (JISC) ; membre du Comité du Conseil de l'Europe responsable des négociations sur une Convention [cadre] sur l'intelligence artificielle (en tant qu'État observateur).

3. 2. UNE PREMIÈRE ÉTAPE : CRÉER EN FRANCE LE CENTRE D'ÉVALUATION DE L'IA

En France, certaines initiatives existent pour évaluer et accompagner la gestion des risques et la mise en conformité de systèmes d'IA.

- Le Laboratoire national de métrologie et d'essais (LNE) a, depuis plusieurs années, une équipe et des outils dédiés à l'évaluation de l'IA.
- L'ANSSI labellise des structures selon leur niveau de cybersécurité et dispose d'un centre de certification national (CCN) pour tester et certifier certains produits numériques.
- Le « bac à sable » données personnelles de la CNIL accompagne trois projets d'IA.
- Le Grand défi sur l'IA de confiance a financé très tôt des ressources en matière de normalisation de l'IA.

Aussi utiles qu'elles soient, ces initiatives manquent de cohérence entre elles et, à date, ne portent pas sur l'évaluation et l'essai de grands modèles d'IA de fondation et de systèmes d'IA à usage général. En outre, elles ne disposent pas des ressources humaines et financières pour le faire.

Cela est d'autant plus problématique que la France a tous les atouts nécessaires pour capitaliser sur un cadre favorable à l'IA sûre et digne de confiance. Elle dispose d'une réelle expertise dans la sûreté de l'IA et dans la sûreté des systèmes plus généralement, ce qui lui confère un avantage important au cas où, précisément, les systèmes d'IA seraient soumis à des exigences de sûreté. Plusieurs acteurs mondiaux des systèmes critiques (Airbus, Thalès, Renault, etc.), et de l'évaluation de la sûreté (Bureau Veritas, LNE, etc.) sont français. Le gouvernement français a investi près de 100 M€ dans un « grand défi pour l'IA de confiance aux services des systèmes critiques », dont 45 M€ ont directement servi au programme de recherche multi-acteurs nommé Confiance.ai. En outre, la France a une expertise mondiale dans plusieurs domaines de recherche qui pourraient s'avérer clés pour la sûreté de l'IA : la cybersécurité, les méthodes formelles, la programmation probabiliste. La note d'action de l'Institut Montaigne d'avril 2023 sur l'IA sûre et digne de confiance détaille ces atouts[66].

Fin 2024, la France accueillera la deuxième édition du Sommet Mondial de l'IA. Sa mission prioritaire en tant que pays hôte sera de concrétiser une nécessaire gouvernance mondiale – peut-être même un « Accord de Paris sur l'IA ». Comme l'explique la « Déclaration de Bletchley » signée lors du précédent Sommet Mondial de l'IA de novembre 2023, ces mesures de gouvernance devront être fondées sur les risques, en s'appuyant sur des mesures d'évaluation appropriées. **Néanmoins, la France manque à date d'un outil de gouvernance indispensable : son propre référent national sur les questions d'évaluation de l'IA.**

[66] *https://www.institutmontaigne.org/publications/investir-lia-sure-et-digne-de-confiance-un-imperatif-europeen-une-opportunite-francaise.*

Au vu des capacités et du savoir-faire français, **l'Institut Montaigne recommande la mise en place d'une Autorité française de l'IA, à même de porter de manière indépendante et fiable l'évaluation et le contrôle des modèles d'IA, en particulier à usage général.** Il ne s'agit pas là de créer une énième entité publique mais bien de nommer un organisme rassemblant des expertises éparpillées et de les réunir au sein d'une structure existante afin de procurer les conseils et les accompagnements nécessaires. Une telle création exige de nombreuses étapes parmi lesquelles, non des moindres, celle d'un vote législatif. Au vu de l'urgence et des enjeux, des étapes intermédiaires sont nécessaires pour faire émerger un tel acteur de référence.

La première étape consiste à créer immédiatement un Centre d'évaluation de l'IA, en prenant en compte les moyens et les contraintes existantes. En effet, compte tenu des avancées fulgurantes de la technologie, du manque de compréhension de l'ensemble des acteurs et du rôle spécial de la France en 2024 dans le cadre du prochain Sommet Mondial de l'IA, cette étape doit être une priorité pour le premier semestre de 2024.

Dans un second temps, dans un cadre juridique précisé, ce Centre pourra évoluer en une Autorité de l'IA : une autorité administrative indépendante, créée par la loi et chargée d'assurer à la fois le développement responsable du secteur de l'IA et la mise en œuvre du règlement européen, aux côtés des régulateurs sectoriels existants.

3.2.1. Le Centre d'évaluation de l'IA – embryon d'une future Autorité – aurait trois missions clés

1/ **Évaluer les grands modèles d'IA de fondation à l'état de l'art.** La France est déjà signataire de l'accord qui a été établi lors du *AI Safety Summit* de 2023 permettant à quelques gouvernements de tester les modèles d'IA des plus grandes entreprises technologiques avant qu'ils ne soient mis sur le marché. Le Centre d'évaluation de l'IA permettra à la France de capitaliser sur ces accès privilégiés. En outre, il bénéficiera d'informations privilégiées concernant les évaluations effectuées par les autres Instituts de Sûreté de l'IA du réseau international, aujourd'hui composé des États-Unis, du Royaume-Uni et de Singapour. Il permettra également d'informer les services de sécurité et de renseignement français sur les risques de l'IA pour la sécurité nationale. Enfin, il permettra de préparer les compétences d'une future Autorité de l'IA, chargée de la régulation directe des grands modèles d'IA à usage général. Cette fonction d'évaluation et de test de l'état de l'art devra être conduite avec un réseau de chercheurs de pointe de l'IA en France, aujourd'hui répartis entre différents instituts de recherche (Inria, CNRS, CEA, etc.).

2/ **Coordonner une stratégie française de la normalisation de l'IA au niveau européen** (CEN-CENELEC JTC21) et international (ISO SC42), en lien étroit avec l'AFNOR, les acteurs français de l'IA et les homologues internationaux – en particulier le NIST américain[67]. Les normes européennes, rédigées par et pour le marché, constituent un enjeu majeur : elles seront la manière concrète dont les entreprises entreront en conformité avec le *EU AI Act* et elles seront conçues d'ici à 2025, avec le gros du travail en 2024. Les développer intelligemment est clé pour réduire le coût de mise en conformité des entreprises tout en limitant les risques dans l'esprit du *AI Act*.

3/ **Mettre en place des outils et un programme d'accompagnement bout-en-bout des entreprises de l'IA, permettant de simplifier et de prendre en charge le coût de la mise en conformité avec le *EU AI Act*.** Cela constitue de fait un « bac à sable » réglementaire, avec une capacité de développer des outils d'évaluation en propre et de fédérer les acteurs de l'écosystème. Pour cela il pourra :
   - mobiliser des entreprises externes d'évaluation, de « *red-teaming* » et d'audit, afin qu'elles proposent leurs services aux entreprises d'IA en

[67] *Un dialogue en matière de normalisation existe déjà entre l'Europe et les États-Unis dans le cadre du EU-US TTC, mais il est largement dominé par l'expertise du NIST et manque de contributions européennes.*

toute confiance. À terme, un label pourrait assurer ce rôle, à l'image du label ExpertCyber décerné par cybermalveillance.gouv.fr pour les professionnels de cybersécurité de confiance. Cette mission aura aussi vocation à faire émerger des *start-ups* de l'évaluation et de l'audit de l'IA, qui deviendront rapidement un marché significatif. Il suffit de penser au marché de l'évaluation et l'audit du risque cyber, notamment au service de l'assurance, mais avec une croissance bien plus rapide et des synergies à trouver avec l'écosystème existant de la cybersécurité ;

- développer soi-même et avec d'autres acteurs nationaux certains outils d'évaluation et de gestion du risque, à l'image de briques développées par le LNE avec le LEIA (Laboratoire d'évaluation de l'intelligence artificielle), par le CEA, ou par d'autres acteurs encore ;
- mobiliser les divers régulateurs concernés par l'IA – la CNIL pour les données personnelles, l'ANSSI pour la cybersécurité, l'ARCOM pour les contenus numériques, l'ARCEP pour les communications électroniques, ainsi que les divers régulateurs sectoriels. Un partage d'informations et de bonnes pratiques devra se faire avec le futur Bureau européen de l'IA prévu par le *AI Act* ;
- en faisant le lien avec les organismes de certification (appelés « organismes notifiés »), chargés d'effectuer la mise en conformité avec la réglementation européenne en s'appuyant sur les normes européennes de l'IA.

Ce programme sera dédié aux nombreuses sociétés françaises de l'IA et de l'IA à usage général, afin de prendre en charge le coût de leur mise en conformité et le risque de non-conformité, mais aussi d'informer la stratégie de la France dans la construction des normes européennes. Cela accorderait un avantage financier et réglementaire à ces acteurs par rapport aux grandes entreprises non-françaises. À terme, il pourrait également constituer un levier d'attractivité pour attirer des acteurs étrangers en France, à l'image de ce qu'a réussi l'Autorité des marchés financiers (AMF) avec une équipe de quelques personnes seulement pour l'accréditation des dossiers PSAN dans le cadre des crypto-actifs.

Les travaux du Centre d'évaluation de l'IA pourront se centrer sur les systèmes d'IA, leurs risques et leurs modes de défaillance. Pour cette raison, ils devront impérativement être complétés par des travaux de recherche conduits en parallèle sur l'impact de l'IA sur la société plus généralement, comme ceux qui seront conduits par l'Institut IA et Société lancé en décembre 2023[68].

Le Centre d'évaluation de l'IA pourrait puiser dans les ressources et l'expertise d'une structure préexistante disposant déjà de moyens et d'expertise en matière d'évaluation et de normalisation de l'IA. **Le Laboratoire national de métrologie et d'essais (LNE),** un établissement public à caractère industriel et commercial (EPIC) équivalent au NIST américain qui héberge le *AI Safety Institute* outre-Atlantique, pourrait être un candidat adapté, disposant d'une équipe technique d'une quinzaine de personnes expertes dans l'évaluation de l'IA, ainsi qu'une culture et un statut juridique qui le positionne à la croisée des secteurs public, privé et académique. Il a participé au processus de normalisation européen de l'IA et a l'habitude de fédérer des écosystèmes autour de « challenges » d'évaluation. Une coordination opérationnelle assurée par le LNE pourrait être complétée par un réseau de chercheurs de l'Inria, l'agence de programme chargée du numérique et du logiciel depuis fin 2023[69], mais aussi du CNRS et du CEA-List.

Des moyens pérennes devront être alloués à ce Centre d'évaluation de l'IA, pour financer à minima dix Équivalents Temps Plein (ETP) chargés de gérer le « bac à sable » et d'assurer l'accompagnement de toutes les entreprises d'IA qui y ont recours ; et quarante à cinquante ETP pour des équipes d'évaluation de la sûreté et la cybersécurité de l'IA qui assureront par ailleurs l'évaluation des grands modèles d'IA dans le cadre du réseau international d'Instituts d'évaluation et de sûreté de l'IA. Le LNE

[68] *https://dauphine.psl.eu/dauphine/media-et-communication/article/creation-de-linstitut-ia-et-societe*.

[69] *https://www.vie-publique.fr/discours/292294-emmanuel-macron-07122023-recherche-francaise#:~:text=Parce%20que%20je%20pense%20que,p%C3%A9riode%20que%20nous%20vivons%20montre*.

par exemple dispose déjà d'une quinzaine d'ETP dédiés à l'évaluation des systèmes d'IA. Certaines équipes de l'ANSSI sont déjà chargées d'évaluer et de certifier la cybersécurité de produits numériques. L'Inria, le CNRS et le CEA-List disposent en outre de nombreux chercheurs à la pointe du domaine, qui pourraient venir en appui des travaux d'évaluation de l'état de l'art *via* le réseau de chercheurs décrit précédemment. Une partie de ce coût pourrait à terme être financée par une monétisation des services d'évaluation de l'IA, tel que le fait déjà le LNE. À titre de comparaison, le Royaume-Uni a pérennisé un financement annuel inédit de 100 M£ pour son *AI Safety Institute*[70]. Aux États-Unis, le budget du *AI Safety Institute* n'a pas été annoncé mais le NIST, au sein duquel il sera incubé, dispose d'un budget de 1,5 Md€ en 2024, plus de dix fois plus que le budget annuel du LNE de 12 M€ en 2023 à PIB égal[71].

La création de ce Centre permettra aux autorités, aux chercheurs et aux entreprises de se familiariser rapidement avec l'état de l'art de la technologie, informera le processus d'élaboration de normes d'ores et déjà en cours et participera au réseau international qui se crée en amont du prochain Sommet Mondial sur l'IA fin 2024. Surtout, un tel centre permettra de préfigurer concrètement un acteur indispensable : une future Autorité française de l'IA.

3.2.2. L'Autorité française de l'IA

Les Autorités françaises sont tout à la fois des vecteurs de sécurisation nationale, de puissants outils de compétitivité pour les entreprises, des instruments d'éclairage pour le régulateur et de rayonnement à l'international.

[70] *https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.*

[71] *Le budget du NIST représente environ 0,007 % du PIB des États-Unis. Le budget LNE représente 0,0004 % du PIB français.*

Compte tenu des forts enjeux de l'IA aussi bien pour les entreprises que pour le gouvernement, il est impératif que cette Autorité soit indépendante et apartisane. Les gouvernements ont eux-mêmes un enjeu à rester à la pointe de l'IA, ce qui pourrait les inciter à minimiser les risques perçus de l'IA et désinvestir son évaluation, dans le cas où il serait difficile de rapidement développer des systèmes d'IA plus performants sans encourir des risques croissants. Ce format d'instance officielle indépendante apparaît donc particulièrement adapté pour le secteur de l'IA, davantage que celui d'une agence rattachée directement au gouvernement comme peut l'être l'Agence nationale de la sécurité des systèmes d'information (ANSSI).

Cette Autorité de l'IA récupérerait les ressources humaines et financières et les trois missions du Centre d'évaluation :

- un rôle d'expertise et d'évaluation de l'état de l'art de l'IA et des risques associés – de défaillance et de mésusage ;
- un rôle de conseil des décideurs publics sur l'évolution de la loi et des normes, notamment européens, au rythme de la technologie ;
- un rôle d'accompagnement des entreprises d'IA dans leur gestion du risque des systèmes qu'elles développent et dans leur mise en conformité.

En plus de ces trois missions, elle aurait une quatrième mission clé : **un rôle de régulateur, en supervisant directement les systèmes d'IA à usage général et en aidant les régulateurs sectoriels à superviser les systèmes d'IA à usage spécifique.**

La richesse de l'écosystème français émergent de l'IA à usage général (synonyme aujourd'hui d'IA générative) milite en faveur d'une structuration sectorielle, en lui offrant, à tout le moins, un interlocuteur de référence sur la gestion des risques.

Cela a été le cas dans le domaine du nucléaire où, rapidement, une autorité de sûreté (ASN) a eu pour mission de superviser et accompagner l'ensemble des acteurs du nucléaire pour s'assurer qu'ils mettaient en

place de bonnes mesures d'évaluation et de gestion du risque. En outre, cette autorité a pour objectif de conseiller le gouvernement et d'autres autorités compétentes en matière de sûreté et de sécurité nucléaire. Son rapport sur la sûreté nucléaire et la radioprotection lui permet de rendre compte chaque année, en particulier au Parlement, de son activité, de ses missions et de l'état de la sûreté nucléaire et de la radioprotection en France. Elle est épaulée dans ses fonctions par un centre d'expertise : l'Institut de radioprotection et de sûreté nucléaire (IRSN), qui a des missions de recherche, d'évaluation et d'appui technique. Dans ce tandem, la grande majorité des ressources sont allouées aux missions de recherche et d'évaluation, plutôt qu'aux missions de régulation. L'IRSN dispose d'un budget de 288 M€, soit quatre fois supérieur au budget du régulateur – l'ASN – qui s'élève à 68 M€.

L'Autorité de sûreté nucléaire (ASN) et l'Institut de radioprotection et de sûreté nucléaire (IRSN)

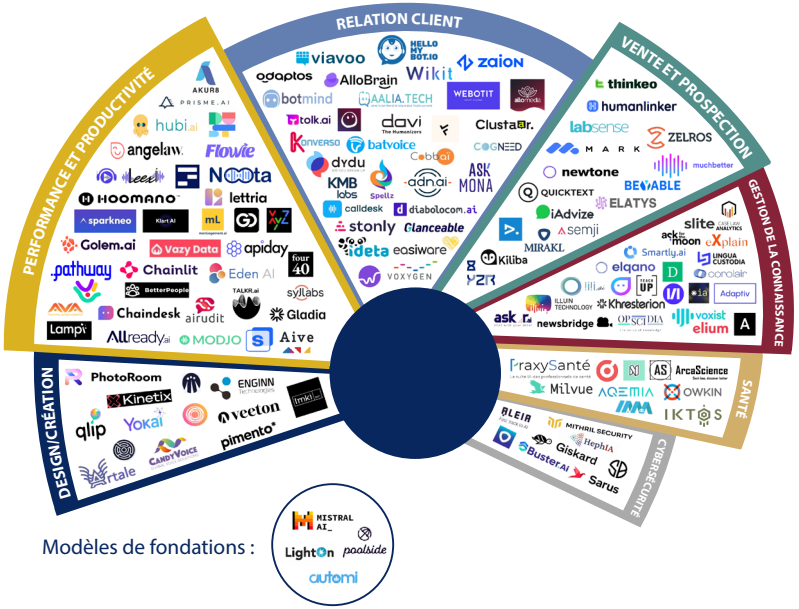
L'ASN est une autorité administrative indépendante (AAI) chargée, au nom de l'État, du contrôle de la sûreté nucléaire et de la radioprotection pour protéger les personnes et l'environnement. Concrètement, il s'agit de l'ensemble des dispositions techniques et des mesures d'organisation relatives à la conception, la construction, le fonctionnement, l'arrêt et le démantèlement des installations nucléaires de base, ainsi qu'à la surveillance et la mitigation du risque des effets nocifs des rayonnements ionisants. Son expertise participe directement au rayonnement français. En particulier, elle a été à l'origine de la création de l'Association des autorités de sûreté nucléaire des pays d'Europe de l'Ouest (WENRA), qui regroupe 17 pays européens. L'ASN est rattachée au programme budgétaire 181 « Prévention des risques » du ministère de l'Écologie. Elle dispose d'un budget de 68 M€ et compte un effectif de 516 personnes – dont une soixantaine d'agents mis à disposition par des établissements publics divers (Andra, Assistance publique – Hôpitaux de Paris, CEA, IRSN, Service départemental d'incendie et de secours). L'ASN sollicite régulièrement les avis et les recommandations de groupes d'experts provenant d'horizons scientifiques et techniques diversifiés, en particulier ceux de l'Institut de radioprotection et de sûreté nucléaire (IRSN).

L'IRSN est un établissement public à caractère industriel et commercial (EPIC) fonctionnant sous un régime de droit privé *via* la tutelle conjointe des ministres chargés de la Défense, de l'Environnement, de l'Industrie, de la Recherche et de la Santé. L'IRSN est l'expert français en matière de recherche et d'expertise sur les risques nucléaires et radiologiques. En 2022, l'IRSN disposait d'un budget de 288 M€ dont 83,5 M€ consacrés à l'action d'appui technique à l'ASN. Ces ressources proviennent du budget général de l'État.

Le respect de l'ASN et de l'IRSN à l'échelle internationale et la crédibilité industrielle et de recherche de ces deux entités sont une illustration puissante de ce que l'État sait faire quand il joue pleinement son rôle régalien de création d'un cadre de confiance permettant à l'innovation et l'activité économique de se développer en toute sécurité.

Cependant, l'IA n'ayant pas encore en France d'infrastructures dédiées ou une empreinte industrielle directe, la comparaison avec l'ASN demeure limitée et mentionnée ici simplement à titre d'exemple. Il conviendra d'adapter nos outils et investissements au secteur émergent de l'IA, d'autant qu'il apparaît foisonnant, comme illustré par le panorama des *start-ups* françaises.

Panorama 2023 des *start-ups* françaises de l'IA générative[72]



Source : Wavestone[73].

L'année 2023 a été une année d'emballement mais aussi de confusion autour de l'IA et des risques associés, avec des débats complexes et animés. L'année 2024 doit apporter de l'apaisement et une clarification aux niveaux national et international. C'est le rôle que joue le réseau d'instituts de sûreté et d'évaluation de l'IA qui se forme progressivement dans le monde. Le Sommet Mondial de l'IA organisé à Paris fin 2024 sera une étape clé du consensus international sur les risques et la gouvernance de l'IA. Si la France souhaite être à la hauteur de cette responsabilité, elle doit se doter d'un espace de compétences et d'évaluation de l'IA, reconnu et indépendant, à même d'objectiver les risques, de s'inscrire comme relais et interlocuteur dans l'écosystème institutionnel naissant en Europe, de participer au débat international et de guider les pouvoirs publics et les entreprises.

[72] *Ce panorama, réalisé par le cabinet de conseil Wavestone en novembre 2023, inclut les start-ups dont le siège est situé en France et dont la proposition de valeur est directement liée au développement ou à l'utilisation de l'IA générative. Les start-ups sont catégorisées par cas d'usage et thématiques traitées par chaque solution.*

[73] *https://www.wavestone.com/fr/insight/radar-2023-des-startups-francaises-ia-generative/.*

Remerciements

L'Institut Montaigne remercie l'ensemble des personnes auditionnées ou consultées dans l'élaboration de ce travail :

- **Arta Alavi,** chef de projet expert au sein de l'équipe Advanced Analytics, SAP
- **Guillaume Avrin,** coordonnateur national pour l'intelligence artificielle
- **Joëlle Barral,** directrice scientifique, Google DeepMind
- **Jean Barrère,** associé, Accuracy
- **Gérôme Billois,** directeur de la practice Cybersécurité, Wavestone
- **Simon Bouloc,** *Chief Innovation Officer,* Ekimetrics
- **Anne Bouverot,** présidente de l'École Normale Supérieure (ENS) et de la Fondation Abeona
- **Bertrand Braunschweig,** coordonnateur scientifique, Confiance.ai
- **Simeon Campos,** expert technique en gestion des risques liés à l'IA, fondateur et président, SaferAI
- **Igor Carron,** *Co-CEO,* LightOn
- **Yves Caseau,** CDIO Groupe, Michelin
- **Julien Chiaroni,** ancien directeur du Grand Défi IA de confiance et responsable programme à la Direction du CEA, LIST
- **David 'davidad' Dalrymple,** directeur de programme, Advanced Research and Invention Agency (ARIA)
- **Tom David,** chargé de la Prospective Technologique, Institut Montaigne
- **Marie-Pierre de Bailliencourt,** directrice générale, Institut Montaigne
- **Nicolas de Bellefonds,** directeur associé senior du BCG et directeur monde de BCGX.AI
- **Agnès Dessaints,** directrice technique, Cisco
- **Marcin Detyniecki,** responsable de la recherche et du développement & Group, AXA
- **Lucas Dixon,** chercheur et responsable de PAIR *(People and AI Research),* Google
- **Marko Erman,** directeur scientifique, Thales
- **Quentin Feuillade--Montixi,** chercheur indépendant en évaluation des modèles de langage AI
- **Ombeline Gras,** conseillère affaires globales du président de la République, Élysée
- **Chadi Hantouche,** associé, Wavestone
- **Didier Husson,** ancien inspecteur pour la Sûreté Nucléaire, Naval Group
- **Laurent Inard,** directeur R&D, Conseil Data & IA, Advisory, Mazars
- **Florent Kirchner,** coordinateur la stratégie nationale d'accélération (SNA) cybersécurité, Secrétariat général pour l'investissement (SGPI)
- **Philippe Laval,** *CTO,* Jolt Capital
- **Fabien Le Voyer,** directeur adjoint du programme en matière d'intelligence artificielle, Inria
- **Franck Lebeugle,** directeur des activités de normalisation, AFNOR
- **Nicolas Marescaux,** directeur adjoint, Réponses Besoins Sociétaires & Innovation, MACIF
- **Sébastien Meunier,** vice président Relations Institutionnelles, ABB France
- **Pierre Peigné,** chercheur indépendant en évaluation des modèles de langage AI
- **Marie Pellat,** ingénieure, Google
- **Ludovic Peran,** Recherche *Product manager* pour l'IA responsable et centrée sur l'humain, Google
- **Tanya Perelmuter,** co-fondatrice et directrice stratégie & partenariats, Fondation Abeona
- **Guillaume Poupard,** directeur général adjoint, Docaposte
- **Marc Revol,** directeur de projets – Intelligence Artificielle, Direction générale des entreprises (DGE)
- **Alexis Rollin,** administrateur de l'EDHEC, président de RnD
- **Stuart Russell,** professeur d'informatique, Université de Californie, Berkeley
- **Jean Schmitt,** *CEO,* Jolt Capital
- **Joseph Sifakis,** chercheur émérite du CNRS, VERIMAG

- **Anuchika Stanislaus,** conseillère numérique et grands projets (IA, Cloud, et chantiers émergents), secrétariat général pour l'investissement (SGPI)
- **François Terrier,** directeur du programme Intelligence Artificielle, CEA-List
- **Martin Tisné,** *CEO,* AI Collaborative, une initiative du Groupe Omidyar
- **Jean-Paul Tran Thiet,** associé de JPTT-Vitale & Partners et expert associé – Justice et Affaires Européennes, Institut Montaigne
- **Félicien Vallet,** chef du service IA, CNIL
- **Gaël Varoquaux,** directeur de recherche, Inria
- **Xavier Vasques,** vice-président et *CTO,* IBM Technology et R&D France
- **Renaud Vedel,** directeur du cabinet du ministre délégué chargé de la transition numérique et des télécommunications

Les opinions exprimées dans ce rapport n'engagent ni les personnes précédemment citées ni les institutions qu'elles représentent.

Institut Montaigne
59 rue La Boétie, 75008 Paris
Tél. +33 (0)1 53 89 05 60
institutmontaigne.org

Imprimé en France
Dépôt légal : janvier 2024
ISSN : 1771-6756

Les adhérents

INSTITUT Montaigne

ABB France
AbbVie
Accenture
Accuracy
Adeo
ADIT
Air Liquide
Airbus
Allen & Overy
Allianz
Amazon
Amber Capital
Amundi
Antidox
Antin Infrastructure Partners
Archery Strategy Consulting
ArchiMed
Ardian
Arquus
AstraZeneca
August Debouzy
AXA
Bain & Company France
Baker & McKenzie
BearingPoint
Bessé
BNP Paribas
Bolloré
Bouygues
Brousse Vergez
Brunswick
Capgemini
Capital Group
CAREIT
Carrefour
Casino
Chubb
CIS
Cisco Systems France
Clariane
Clifford Chance
Club Top 20
CNP Assurances
Cohen Amir-Aslani
Compagnie Plastic Omnium
Conseil supérieur du notariat
Crédit Agricole
D'Angelin & Co.Ltd
Dassault Systèmes
De Pardieu Brocas Maffei
Deloitte
ECL Group
Edenred
EDF
EDHEC Business School
Ekimetrics France
Engie
EQT
ESL & Network
Ethique & Développement
Eurogroup Consulting
FGS Global Europe
Fives
Getlink
Gide Loyrette Nouel
Google
Groupama
Groupe Bel
Groupe M6
Groupe Orange
Hameur et Cie
Henner
Hitachi Energy France
Howden France
HSBC Continental Europe
IBM France
IFPASS
Incyte Biosciences France
Inkarn
Institut Mérieux
International SOS
Interparfums
Intuitive Surgical
Ionis Education Group
iQo
ISRP
Jeantet Associés
Jolt Capital
Katalyse
Kea & Partners
Kearney
KPMG S.A.
Kyndryl
La Banque Postale
La Compagnie Fruitière
Linedata Services
Lloyds Europe
L'Oréal
Loxam
LVMH - Moët-Hennessy - Louis Vuitton
M.Charraire
MACSF
Mazars
Média-Participations
Mediobanca
Mercer
Meridiam
Microsoft France
Mitsubishi France S.A.S
Moelis & Company
Moody's France
Morgan Stanley
Natixis
Natural Grass
Naval Group
Nestlé
OCIRP
ODDO BHF
Oliver Wyman
Ondra Partners
Onet
Optigestion
Orano
PAI Partners
Pelham Media
Pergamon
Publicis
PwC France & Maghreb
Raise
RATP
RELX Group
Renault
Rexel
Ricol Lasteyrie
Rivolier
Roche
Rokos Capital Management
Rothschild & Co
RTE
Safran
Sanofi
SAP France
Schneider Electric
ServiceNow
Servier
SGS
SIER Constructeur
SNCF
SNCF Réseau
Sodexo
SPVIE
SUEZ
Taste
Tecnet Participations SARL
Teneo
The Boston Consulting Group
Tilder
Tofane
TotalEnergies
Unicancer
Veolia
Verian
Verlingue
VINCI
Vivendi
Wakam
Wavestone
Wendel
White & Case
Willis Towers Watson France
Zurich

L'année 2023 a été marquée par la déferlante IA. Le monde a découvert ChatGPT, l'intelligence artificielle à usage général de l'entreprise américaine OpenIA. Pris de court par le caractère exponentiel de l'accélération technologique, les dirigeants politiques et économiques du monde entier se réunissaient pour tenter d'évaluer les nouveaux risques et inventer une gouvernance mondiale autour de l'IA lors du premier Sommet sur la Sûreté de l'IA organisé au Royaume-Uni. En effet, les capacités de cette technologie apparaissent tout aussi irrésistibles que dangereuses, comme c'est souvent le cas pour les innovations les plus révolutionnaires pour l'humanité. Dans ce contexte de gouvernance internationale tâtonnante, l'élaboration d'outils d'évaluation pour comprendre les évolutions de la technologie et de ses risques apparaît nécessaire.

Alors que plusieurs pays *leaders* de l'IA ont déjà pris les devants en annonçant la création de leurs propres Instituts de Sûreté de l'IA et que l'Europe s'apprête à mettre en place des règles contraignantes en matière de développement et d'usage, la France manque à l'appel de cette réflexion et prend un retard impardonnable.

Or, Paris accueillera bientôt la deuxième grande édition du Sommet Mondial de l'IA : l'opportunité d'inscrire la question à l'agenda de chacun mais aussi de faire face à nos responsabilités et de montrer notre engagement et le sérieux de notre réflexion, tant sur le potentiel, les risques et les atouts du jeune écosystème français que sur les outils de gouvernance indispensables à un pilotage sécurisé.

Aussi l'Institut Montaigne recommande-t-il au gouvernement de travailler à la création d'une Autorité de l'IA, sur le modèle des Autorités nationales propres à la France. Cette Autorité aurait pour mission principale d'évaluer rigoureusement, au rythme de ses évolutions, la performance et les risques de la technologie. Conscient des étapes et des contraintes réglementaires à l'établissement d'une telle Autorité de référence, l'Institut Montaigne propose une feuille de route opérationnelle, viable et responsable, rassemblant les forces en présence.

10 €
ISSN : 1771-6756
NAC2401-01