INSTITUT Montaigne

Investir l'IA sûre et digne de confiance : un impératif européen, une opportunité française

NOTE D'ACTION - AVRIL 2023

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 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 - Avril 2023

Investir l'IA sûre et digne de confiance : un impératif européen, une opportunité française

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

Synthèse

Alors que les systèmes d'intelligence artificielle (IA) ont jusqu'à présent été très spécialisés, capables uniquement d'effectuer un nombre de tâches restreint, nous assistons désormais à un changement de paradigme. **Le développement rapide de l'IA - incarné par les systèmes d'IA à usage général comme ChatGPT, capables d'effectuer un grand nombre de tâches - présage de bouleversements technologiques considérables.** L'IA pourrait ainsi rapidement constituer un facteur de compétitivité décisif pour les entreprises comme pour les pays. En témoigne la croissance exponentielle des montants qui y sont investis : 92 milliards de dollars d'investissement privé en 2022, soit près de 20 fois plus qu'il y a dix ans.

UN ENJEU DE SÉCURITÉ ET UN IMPÉRATIF DE SÛRETÉ

L'accélération du développement de l'IA constitue néanmoins un enjeu de sécurité majeur et croissant. À l'image de la technologie nucléaire, l'IA est une technologie intrinsèquement duale qui peut rapidement être détournée à des fins malveillantes. Les systèmes d'IA d'apprentissage automatique, qui apprennent à effectuer une tâche à partir d'exemples plutôt que de règles prédéfinies, posent également un risque de sûreté et de défaillance inédit, puisque leur nature statistique les rend intrinsèquement imprévisibles. Ils sont ainsi peu robustes, c'est-à-dire que leur comportement peut subitement changer dans des environnements nouveaux, et sont difficilement explicables : ce sont des "boîtes noires" qui fonctionnent en autonomie, sans que l'on sache réellement ni comment ni pourquoi. Ainsi, Google (BARD),Microsoft (Bing) ou OpenAI (ChatGPT) ne parviennent à prévenir ni les erreurs factuelles ni les dérives violentes ou biaisées de leurs agents conversationnels. Au fur et à mesure qu'ils gagnent en capacité et en autonomie, la bonne spécification des objectifs de ces systèmes, c'est-à-dire leur alignement avec l'intérêt général, devient dès lors un enjeu majeur. **Avec les avancées rapides de l'IA et sa dissémination massive dans l'ensemble des secteurs d'activité, ce risque de défaillance pourrait rapidement augmenter et représenter un enjeu de sûreté et de résilience aux échelles nationale et internationale.**

DÉFINIR NOS PRÉFÉRENCES POUR LES SYSTÈMES D'IA : UN ENJEU SOCIÉTAL

Le développement de systèmes d'IA avancés et leur déploiement à grande échelle constituent également un enjeu sociétal de premier ordre. En effet, dans la mesure où les systèmes d'IA prennent des décisions ou effectuent de recommandations, elles sont porteuses de valeurs et impactent nécessairement notre liberté, que leurs concepteurs le veuillent ou non. **Les laboratoires d'IA qui développent des modèles d'IA à usage général tentent désormais d'améliorer la performance de ces systèmes en y intégrant explicitement les préférences humaines, c'est-à-dire un modèle de valeurs.** ChatGPT s'appuie par exemple sur l'apprentissage par renforcement à partir de retours humains (RLHF). Interpréter les préférences humaines en nous demandant notre avis (RLHF par exemple), en observant nos comportements ou en spécifiant une liste de principes moraux implique un parti pris philosophique et une réflexion éthique approfondie avant tout déploiement opérationnel au cœur de nos machines. **Or la spécification des préférences humaines pour des systèmes d'IA constitue un domaine de recherche naissant, stratégique mais quasiment inexploré à date.**

L'IA SÛRE ET DIGNE DE CONFIANCE : UN AVANTAGE STRATÉGIQUE SUR LEQUEL MISER

Grâce à des moyens accrus - notamment privés - les États-Unis et la Chine ont acquis une longueur d'avance substantielle en matière de développement économique et technologique de l'IA. L'Europe a ainsi cumulé un retard difficilement rattrapable. **Miser sur l'IA sûre et digne de confiance constitue dès lors notre meilleure stratégie de différenciation pour se positionner en acteur clé de l'IA. Elle constitue par ailleurs un impératif pour protéger notre sécurité et notre modèle de société.**

La sûreté et la confiance constituent désormais une barrière technologique importante au développement des systèmes d'IA à usage général et une préoccupation centrale des meilleurs talents internationaux de l'IA. L'Union européenne s'apprête par ailleurs à imposer aux systèmes d'IA des exigences de sûreté et de confiance grâce à une réglementation extraterritoriale, couplée à

une directive en matière de responsabilité civile pour l'IA et à un travail pionnier sur les normes. Ce cadre réglementaire pourrait avoir la même portée internationale que le RGPD avant lui.

Au sein de l'Europe, la France est particulièrement motrice et s'est positionnée comme *leader* sur le sujet. Elle est notamment à l'origine de l'inclusion des modèles d'IA "à usage général" (type ChatGPT) dans le règlement européen sur l'IA et se démarque comme étant motrice dans les efforts de normalisation menés au niveau européen. Surtout, elle recense une expertise mondiale sur plusieurs briques techniques clés pour développer des systèmes d'IA à usage général sûrs et dignes de confiance : en recherche fondamentale, grâce à des chercheurs de rang mondial en mathématiques et en IA capables d'attirer les meilleurs talents internationaux ; en ingénierie système et logicielle pour la sûreté, grâce à un Grand Défi sur l'IA de confiance pour les systèmes critiques et à un écosystème d'industriels ; en développement de grands modèles d'IA à usage général, grâce au projet Bloom, un grand modèle de langage comme ChatGPT développé avec des chercheurs français et les puissants ordinateurs du Centre national de la recherche scientifique (CNRS).

Si la France et l'Europe souhaitent pleinement capitaliser sur cette opportunité inédite, elles doivent adopter une approche ambitieuse pour développer des systèmes d'IA à usage général véritablement sûrs et dignes de confiance d'une part, et pour réguler les systèmes d'IA à usage général dangereux d'autre part.

Objectif 1 :
Faire de la France un leader mondial de la R&D
dans la sûreté et la confiance des modèles d'IA à usage général

RECOMMANDATION 1 : Attirer en France les meilleurs chercheurs internationaux de l'IA avec un appel porté au plus haut niveau de l'État, sur le modèle de l'initiative "Make Our Planet Great Again", centré sur le développement de systèmes d'IA à usage général sûrs et dignes de confiance.

RECOMMANDATION 2 : Mener un projet d'innovation de rupture pour développer des systèmes d'IA à usage général sûrs et dignes de confiance, doté de 100 millions d'euros et d'une gouvernance agile, qui s'appuie sur les forces de l'écosystème français.

RECOMMANDATION 3 : Créer un pôle de recherche mondial sur la compréhension des préférences humaines et leur bonne spécification pour des systèmes d'IA à usage général. Confier la coordination de ce pôle à un institut de recherche emblématique (ENS ou 3IA par exemple) et assurer son financement via une enveloppe dédiée, par exemple des Programmes et équipements prioritaires de recherche (PEPR).

RECOMMANDATION 4 : Faire de l'IA sûre et digne de confiance un projet important d'intérêt européen commun (PIIEC) permettant d'assouplir les règles d'aides d'État et/ou l'un des "produits phares" de l'Union européenne dotés d'environ 1 milliard d'euros.

RECOMMANDATION 5 : Développer en France deux référentiels (*benchmarks*) pour la recherche permettant de mesurer la confiance et la performance d'un système d'IA à usage général.

RECOMMANDATION 6 : Créer une discipline de sûreté de l'IA (ou génie de l'IA) en conditionnant le financement public des formations à l'IA à l'intégration d'un module sur la sûreté et la confiance de l'IA.

Objectif 2 :
Définir un cadre réglementaire européen
pour la sûreté et la confiance de l'IA à usage général
et favoriser son adoption dans le monde

RECOMMANDATION 7 : Concrétiser la proposition de la France d'inscrire les systèmes d'IA à usage général dans la réglementation européenne de l'IA et favoriser son adoption dans le monde via le E.U.-U.S. *Trade and Technology Council* (TTC) et le G20.

RECOMMANDATION 8 : Confier au futur régulateur français de l'IA une expérimentation pilote ou un audit à blanc du processus d'audit de l'IA prévu par la réglementation européenne, afin d'accompagner la montée en puissance d'un écosystème d'audit français (entreprises, auditeurs, régulateur).

RECOMMANDATION 9 : Développer au sein du futur régulateur français de l'IA et en association étroite avec les acteurs de l'évaluation comme le LNE un "bac à sable" (*sandbox*) réglementaire de l'IA, pour tester sans conséquence juridique le degré de conformité de nouveaux systèmes d'IA et d'IA à usage général.

RECOMMANDATION 10 : Confier au futur régulateur français de l'IA la création d'une base de données de référence de documentation des défaillances de systèmes d'IA.

Sommaire

Auteur

Milo Rignell

Milo Rignell est responsable des travaux de l'Institut Montaigne sur les sujets numériques et de nouvelles technologies depuis 2022. Ce travail a couvert, à date, différents éléments des stratégies numériques française et européenne, dont la cybersécurité, le financement, l'accès aux talents, et tout particulièrement l'intelligence artificielle sûre et éthique. Précédemment au poste de chargé de l'Innovation, Milo était responsable de projets d'expérimentation du think tank, dont une formation en ligne à l'intelligence artificielle, Objectif IA, et un projet d'apprentissage des mathématiques à l'école primaire.

1 Les systèmes d'IA à usage général : un nouveau paradigme de l'IA

L'intelligence artificielle (IA) est un ensemble de techniques permettant d'automatiser des tâches normalement confiées à des humains, en particulier de raisonnement[1] et de perception.

La difficulté à appréhender ce que constitue "l'intelligence" rend difficile un consensus sur ce que constitue une "intelligence artificielle". Néanmoins, l'OCDE propose une définition qui s'en rapproche le plus : "un système d'intelligence artificielle (ou système d'IA) est un système automatisé qui, pour un ensemble donné d'objectifs définis par l'homme, est en mesure d'établir des prévisions, de formuler des recommandations, ou de prendre des décisions influant sur des environnements réels ou virtuels. Les systèmes d'IA sont conçus pour fonctionner à des degrés d'autonomie divers."

Pour se saisir du sujet et comprendre concrètement de quoi on parle, il faut s'intéresser aux techniques précises qui constituent l'IA aujourd'hui. Il existe à date deux grandes approches à l'IA : l'IA symbolique et l'apprentissage machine.

Les systèmes d'IA symbolique s'appuient sur des faits et des règles formelles pour déduire un résultat. Cette approche a connu un succès important dans les années 70, avec des systèmes d'IA dits "experts", capables de simuler le savoir-faire d'un expert humain. Dans le domaine médical, des systèmes experts aident ainsi au diagnostic : "Si le patient fait preuve de symptôme X, alors c'est qu'il est atteint de maladie Y". Dans le domaine des échecs, le système expert DeepBlue développé par IBM a fait sensation en 1997 en s'imposant face au joueur d'échecs international Garry Kasparov. Néanmoins, les systèmes experts sont rapidement limités quand il s'agit de préciser des règles formelles capables de prendre en compte un grand nombre de cas de figure. Un ordinateur peut calculer toutes les permutations dans un jeu d'échecs, mais a plus de mal à calculer toutes les permutations d'un jeu plus complexe comme le Go , et n'a aucune chance de calculer toutes les éventualités dans le monde réel, par exemple pour conduire une voiture.

Les systèmes d'apprentissage machine (*machine learning*) résolvent ce problème grâce à un raisonnement inductif et probabiliste. Ils s'appuient sur des données et des méthodes de raisonnement statistiques afin d'apprendre des corrélations. Dans le domaine médical par exemple, plutôt que d'établir un diagnostic médical en appliquant des connaissances spécifiques du domaine et des règles préconçues, comme le ferait un système expert, un système d'apprentissage machine va parcourir un grand nombre de cas déjà diagnostiqués afin d'établir lui-même des corrélations. Néanmoins ces corrélations ne reflètent pas nécessairement un lien de causalité. C'est ainsi que certains systèmes d'IA de diagnostic de cancer ont "appris" à distinguer des images de tumeurs malignes et bénignes selon la présence ou non d'une règle graduée dans l'image. Parmi les images de tumeurs pré-diagnostiquées qui leur ont été fournies, les images de tumeurs malignes contenaient plus souvent une règle graduée, présente pour mesurer la taille de la tumeur . Selon un raisonnement inductif et probabiliste, et non pas déductif et logique, le lien entre la règle et le diagnostic d'un cancer est avéré. Malgré ces limites, les systèmes d'apprentissage machine représentent aujourd'hui la grande majorité des cas d'usages d'IA (reconnaissance d'image, reconnaissance vocale, algorithmes de recommandation de contenu ou d'achats, de traduction, etc). Plusieurs projets tentent néanmoins de combiner les forces de l'IA symbolique avec celles de l'apprentissage machine : cette approche constitue l'IA hybride.

[1] Le terme raisonnement est à prendre au sens large, comme processus cognitif permettant de poser un problème en vue d'obtenir un résultat. Le terme a par exemple été utilisé par le Parlement européen pour définir l'IA.

[2] Aux échecs, 10^120 parties différentes sont possibles, et le nombre de coups possibles dans une position typique avoisine 40. Au Go, 10^170 parties différentes sont possibles, et le nombre de coups possibles dans une position typique avoisine 300. À titre de comparaison, on estime à 10^80 le nombre de particules élémentaires dans l'univers visible. https://espaces-numeriques.org/wp-content/uploads/2018/04/L111Sp24.pdf

[3] 2018, "Journal of Investigative Dermatology"

Les systèmes d'apprentissage automatique se divisent en trois familles : l'apprentissage supervisé ; l'apprentissage non supervisé ; l'apprentissage par renforcement. Comme pour l'exemple du diagnostic médical, chacun comporte des avantages et des limites.

Les trois familles de l'apprentissage automatique

1.

L'APPRENTISSAGE SUPERVISÉ

L'algorithme est capable de prédire la valeur ou la catégorie d'un objet d'entrée (une image, un appartement, etc.) en apprenant à partir d'un corpus d'exemples étiquetés, c'est-à-dire pour lesquels la valeur ou la catégorie de l'objet est déjà indiquée.

Par exemple, en apprenant à partir d'un corpus de descriptions de maisons (localisation, superficie, présence ou absence de certains équipements) pour lesquelles la valeur du prix est déjà indiquée, un algorithme est capable de prédire le prix d'une maison qui n'est pas dans le corpus. Ou en apprenant à partir de centaines d'images qui ont déjà été catégorisées en tant que chien ou chat, un algorithme est capable d'indiquer si une nouvelle image est celle d'un chat ou d'un chien (ou ni l'un ni l'autre, à condition que cette catégorie "non identifiée" ait été spécifiée).

Le comportement de l'algorithme dépend intimement

1. de la qualité des données d'apprentissage : elles peuvent être biaisées et non-représentatives, ou se faire le relai de biais préexistants ;
2. des catégories spécifiées : elles peuvent être non exhaustives et ne pas contenir une catégorie "non identifié", et dépendent de la façon dont les données d'entraînement ont été étiquetées ;
3. du type d'algorithme employé et de son paramétrage.

2.

L'APPRENTISSAGE NON-SUPERVISÉ

- L'algorithme est capable de regrouper, par exemple en catégories distinctes, des objets non étiquetés selon leurs similarités et leurs différences. On n'indique ni un type de valeur (par exemple le prix), ni de catégories prédéfinies selon lesquelles on souhaite classer les objets. On laisse l'algorithme identifier lui-même la façon la plus pertinente de regrouper les objets du corpus.

- Par exemple, l'algorithme pourrait apprendre à regrouper lui-même des images non étiquetées de chiens et de chats en deux catégories, ou regrouper des articles de journaux par thèmes.

- Le comportement de l'algorithme dépend intimement

1. des données qui caractérisent les objets ;
2. des critères de similarité ou de différence qui sont utilisés ;
3. de la façon dont ces critères sont analysés et pondérés.

3.

L'APPRENTISSAGE PAR RENFORCEMENT

- L'algorithme est capable d'agir dans un but donné (jouer aux échecs, conduire une voiture, etc.) en apprenant par tâtonnements successifs (méthode essai-erreur) et en étant plus ou moins récompensé pour ses actions en fonction de s'il se rapproche, ou atteint, l'objectif spécifié. Tout comme on apprend à un rat d'effectuer des tâches précises en le récompensant avec de la nourriture.

- Par exemple, l'algorithme pourrait apprendre à proposer des vidéos que l'utilisateur a envie de voir en recevant une récompense, c'est-à-dire un score qu'il doit maximiser, à chaque fois que la vidéo proposée est visionnée dans son intégralité par l'utilisateur. Ou l'algorithme peut apprendre à jouer aux échecs ou à des jeux vidéo, en recevant une récompense lorsqu'il gagne des points dans le jeu ou qu'il remporte la partie.

- Le comportement de l'algorithme dépend intimement de la façon dont l'objectif est spécifié, c'est-à-dire la façon dont l'objectif est traduit en récompense. Spécifier l'objectif de proposer des vidéos "que l'utilisateur a envie de voir" en récompensant des vidéos vues dans leur intégralité peut favoriser les vidéos courtes, sensationnelles, ou qui vont dans le sens d'opinions fortes de l'utilisateur. Spécifier une fonction de récompense qui soit parfaitement alignée avec les objectifs souhaités s'avère être particulièrement difficile et pernicieux.

  Ces familles ne sont pas exhaustives mais donnent une idée utile du fonctionnement de la majorité des systèmes d'apprentissage automatique. Il existe d'autres techniques complémentaires, telles que l'apprentissage semi-supervisé, qui ne requiert qu'un petit nombre d'exemples étiquetés, l'apprentissage par transfert, qui s'appuie sur l'apprentissage d'une tâche donnée pour apprendre à effectuer une nouvelle tâche, et l'apprentissage auto-supervisé, considéré comme une forme intermédiaire entre l'apprentissage supervisé et non supervisé. Ce dernier est notamment utilisé dans des modèles d'IA de traitement du langage naturel tels que ChatGPT. Le modèle génère lui-même l'étiquetage des données en masquant certaines données d'apprentissage, tel que des mots, et en s'entraînant à les prédire.

Au cours des dernières années, les techniques de *machine learning* ont connu des progrès impressionnants grâce à deux avancées en particulier : **l'apprentissage profond (ou *deep learning*)** et ***le transformer***. Le *deep learning* s'inspire de la structure du cerveau humain pour doper la performance des systèmes de *machine learning* sur de nombreuses tâches comme la reconnaissance d'image ou le traitement de langage naturel. Cette technique marche particulièrement bien pour des problèmes qui disposent de très grandes quantités de données d'entraînement. Le *transformer* est un modèle de *deep learning* introduit en 2017 qui permet au système d'IA de concentrer son attention sur les données les plus pertinentes pour la tâche demandée.

Alors que les systèmes d'IA ont historiquement été très spécialisés, capables par exemple de diagnostiquer des images médicales mieux que qui ce soit mais incapable de faire autre chose, ce paradigme commence à changer. Les avancées techniques du *deep learning* et des *transformers*, soutenues par l'explosion de donnée[4] et des capacités de calcul[5] , ont notamment permis l'émergence de modèles d'IA "à usage général" (*general purpose AI*, ou GPAI, en anglais). **Comme leur nom l'indique, les modèles d'IA à usage général sont capables d'effectuer un grand nombre de tâches différentes.** Des systèmes d'IA de traitement du langage naturel, comme le chatbot ChatGPT, peuvent désormais faire les devoirs de nos enfants, aussi bien en maths qu'en français. Le système d'IA Gato, développé par Google DeepMind, est capable d'effectuer plus de 600 tâches différentes : discuter avec des humains, reconnaître des objets, manipuler des bras robotiques, jouer à des jeux vidéo, etc. Ces modèles d'IA, complexes et coûteux à développer, sont parfois appelés "modèles fondationnels" du fait de leur capacité à être réutilisés et adaptés par différents acteurs pour des cas d'usages spécifiques. Des entreprises ou utilisateurs peuvent ainsi soumettre de nouvelles données d'apprentissage au modèle d'IA préexistant afin d'optimiser sa performance sur un cas d'usage précis, par exemple ses réponses à des questions médicales ou juridiques (technique dénommée "*fine tuning*").

Le développement rapide de l'IA à usage général présage des bouleversements technologiques majeurs. Ce développement technologique pourrait par ailleurs rapidement accélérer si ces systèmes d'IA y contribuent eux-mêmes. (cf. Annexe 1 - estimations du rythme de développement de l'IA). D'autant plus que le développement de nouvelles capacités par les systèmes d'IA les plus avancés est souvent difficile à prédire. Il suffit

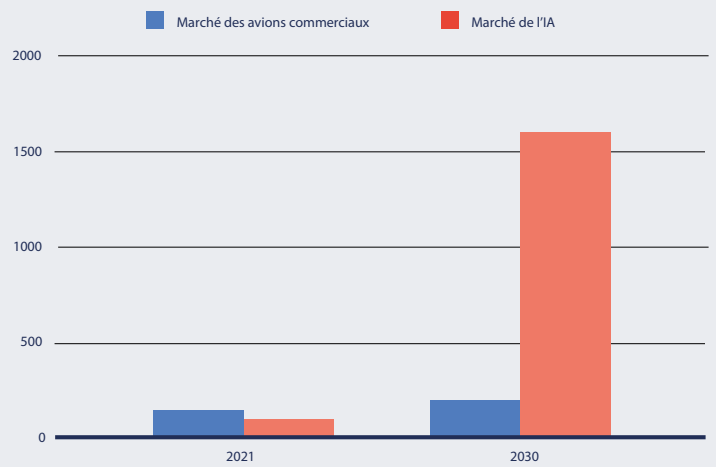
[4] Le volume de données créées est passé – au niveau mondial – de 2 zettaoctets (2 trilliard d'octets) en 2010 à 18 zettaoctets en 2016, puis, selon les prévisions, à 64 zettaoctets en 2020 et 181 zettaoctets en 2025. Note scientifique de l'Office parlementaire d'évaluation des choix scientifiques et technologiques, de janvier 2023. https://www.assemblee-nationale.fr/dyn/16/rapports/ots/l16b0768_rapport-information

[5] Selon la loi de Moore, le nombre de transistors incorporés dans une puce de circuit intégré densément peuplée double approximativement tous les deux ans. Cette observation est plus ou moins avérée depuis les années 1970.

parfois de dépasser un seuil dans la taille du modèle d'IA (son nombre de paramètres), la quantité de données d'entraînement ou la durée d'entraînement pour que le système développe subitement de nouvelles capacités. Par le passé, des grands modèles de langage ont par exemple soudainement développé des capacités pour écrire ou faire de l'arithmétique.

2 L'IA sera au 21ème siècle ce que la physique de l'atome a été au 20ème siècle : un enjeu de compétitivité historique, un enjeu majeur de sécurité à échelle nationale, un enjeu de liberté et de valeurs sociétales.

Comparaison des marchés des avions commerciaux et des systèmes d'IA, estimés en 2021 et en 2030 (en milliards de dollars)



L'IA pourrait ainsi rapidement constituer l'un des premiers facteurs de compétitivité, comme en témoigne la croissance exponentielle des montants qui y sont investis : 92 milliards de dollars d'investissement privé en 2022, soit près de 20 fois plus qu'en 2013[6] . Le marché propre de l'IA pourrait passer d'environ $87,04 milliards en 2021 à $1 597,1 milliards en 2030, soit huit fois la taille du marché des avions commerciaux, avec une croissance moyenne (CAGR) de 38.1 %[7] . La valeur ajoutée de l'IA pour l'ensemble des secteurs de l'économie a été estimée par PwC à $15 700 milliards d'ici 2030, soit environ +14 %[8] de PIB mondial.

La pénétration de l'IA dans l'économie est en forte croissance[9] et son potentiel pour accélérer les sciences et la technologie est immense. L'algorithme AlphaFold a par exemple débloqué l'un des problèmes les plus complexes en sciences médicales : prédire la structure des protéines à partir de leur séquence en acides aminés, valant à ses concepteurs le prix 2022 des avancées capitales dans les sciences de la vie. De nouveaux systèmes d'IA permettront certainement d'autres avancées majeures dans les années à venir, y compris des avancées permettant d'accélérer rapidement le développement même de l'IA.

L'accélération du développement de l'IA constitue néanmoins un enjeu de sécurité croissant. À l'image de la technologie nucléaire, l'IA est une technologie duale : elle peut être utilisée à des fins civiles mais aussi militaires, et peut facilement être détournée à des fins malveillantes. Par exemple, des systèmes d'IA de découverte de médicaments, utilisés pour identifier des molécules capables de guérir des patients de maladies, peuvent facilement être détournés pour identifier des molécules létales, utilisables dans des armes biochimiques.

[6] Stanford Institute for Human-Centered Artificial Intelligence (HAI) 2023 AI Index Report

[7] Selon une étude de Precedence Research, d'autres études chiffrent le même ordre de grandeur.

[8] En 2019, McKinsey estimait que l'Europe pourrait ajouter environ 2 700 milliards d'euros de PIB en 2030, soit +20 %, ce qui se traduirait par une croissance annuelle composée de 1,4 % sur cette période. Plus récemment, en avril 2023, un rapport de Goldman Sachs estimait que l'IA générative pourrait engendrer une augmentation de 7 % (soit près de 7 000 milliards de dollars) du PIB mondial et une augmentation de la croissance de la productivité de 1,5 point de pourcentage sur une période de dix ans.

[9] Selon le rapport "The State of AI in 2021" de McKinsey, 56 % de l'ensemble des répondants font état de l'adoption de l'IA dans au moins une fonction, contre 50 % en 2020.

Des modèles d'IA capables de générer du code informatique, tels que ChatGPT, peuvent être détournés pour identifier des vulnérabilités dans des systèmes d'information et concevoir des attaques cyber plus nombreuses et plus performantes. Chacun de ces exemples est avéré. Détourner un modèle d'IA de découverte de médicaments ou intégrer un système de reconnaissance faciale dans un missile est par ailleurs beaucoup plus facile et moins coûteux que détourner une usine d'enrichissement d'uranium pour produire des armes nucléaires. Le risque d'usages malintentionnés de l'IA, par exemple par des groupes terroristes, est ainsi particulièrement élevé[10].

À la différence de la technologie nucléaire, les systèmes d'IA d'apprentissage automatique posent un risque de sûreté et de défaillance inédit. Leur nature statistique en fait des systèmes intrinsèquement imprévisibles : une fois que les objectifs de l'algorithme et la méthode d'apprentissage ont été spécifiés par son concepteur, un système d'IA apprend à réaliser une tâche de façon autonome, en s'appuyant sur ses données d'entraînement pour adapter et optimiser son comportement. Ils sont **peu robustes**, c'est-à-dire que leur comportement peut subitement changer dans des environnements nouveaux, et sont **difficilement explicables** : ce sont des "boîtes noires" qui fonctionnent, sans qu'on ne sache réellement ni comment ni pourquoi. On peut observer les résultats de sortie, on peut observer les modifications au système d'IA au fur et à mesure qu'il apprend, mais on ne sait pas à quoi servent chacune de ces modifications, ni quel sera leur impact sur son comportement. L'algorithme de classification automatisée de tumeurs cité ci-dessus identifiait en réalité les règles graduées et non pas des tumeurs, risquant ainsi de nombreux mauvais diagnostics lors de son déploiement. Google (BARD) aussi bien que Microsoft (Bing) et OpenAI (ChatGPT) ne parviennent pas à prédire et prévenir les erreurs factuelles et les dérives violentes ou biaisées de leurs agents conversationnels. Au fur et à mesure qu'ils gagnent en capacité et en autonomie, **la bonne spécification des objectifs** de ces systèmes devient également un enjeu de sûreté majeur. Les agents d'apprentissage par renforcement de Google DeepMind trouvent des solutions ingénieuses pour maximiser leur récompense qui détournent l'objectif de leurs concepteurs, voire induisent en erreur leurs évaluateurs humains. Il existe ainsi de nombreux exemples de défaillances et de comportements imprévisibles, parfois dangereux. La base de données *Artificial Intelligence Incident Database* catalogue de nombreux autres exemples d'incidents de défaillances de systèmes d'IA.

[10] *L'IA pose également des risques structurels : en modifiant profondément le paysage stratégique, elle impacte d'autres risques majeurs. L'IA pourrait par exemple augmenter le risque cyber : en augmentant les capacités de cyberattaque avec de nouveaux moyens pour détecter et exploiter des vulnérabilités informatiques à grande échelle. L'IA pourrait aussi impacter le risque de catastrophe nucléaire : en obérant la capacité de seconde frappe grâce à ses capacités de détection de l'arsenal nucléaire d'un pays, l'IA peut encourager une frappe nucléaire préventive. Les risques structurels de l'IA sont néanmoins indirects et peu utiles à la discussion de la note.*

Les trois problèmes de sûreté en IA : la robustesse, l'explicabilité et la transparence, et la bonne spécification des objectifs

Le *Center for security and emerging technology* (CSET), un *think tank* américain spécialisé dans les nouvelles technologies émergentes, regroupe les enjeux de sûreté des systèmes d'IA (*AI safety*) en trois grandes familles[11] : la robustesse, l'explicabilité et la transparence, et la bonne spécification des objectifs[12].

1.

La robustesse d'un système d'IA garantit que celui-ci fonctionnera de façon sûre, y compris dans des situations qui ne lui sont pas familières. Or le comportement de systèmes d'apprentissage machine s'appuie sur des corrélations statistiques, et non pas une compréhension de la réalité sous-jacente.

[11] *Le Select Committee on Artificial Intelligence, qui conseille le gouvernement américain sur la stratégie en IA, souligne également les sources d'imprévisibilité et de risque des systèmes d'IA liés au déploiement d'IA dans des environnements complexes et incertains et le risque de comportement émergent (problèmes de robustesse), ainsi que l'enjeu d'une mauvaise spécification des objectifs.*

[12] *Le problème de spécification est parfois aussi appelé problème d'alignement, ou problème de contrôle.*

Ainsi quand la réalité sous-jacente change, et que les corrélations disparaissent, le système d'IA peut adopter un comportement inadapté à la nouvelle situation et potentiellement dangereux, comme en témoigne l'exemple de diagnostic médical ci-dessus. Certaines cyberattaques sont conçues pour exploiter ces vulnérabilités. Une attaque par exemples contradictoires (*adversarial attack*) vise à tromper un système d'IA, en changeant légèrement les exemples qui lui sont soumis. Ainsi suite à quelques modifications imperceptibles à un panneau stop par une personne malveillante, quelques coups de feutre noir par exemple, le système d'IA dans votre voiture pourrait ne pas reconnaitre le panneau. Résultat : la voiture ne s'arrête pas au STOP et déclenche un accident.

Exemple d'une attaque par exemples contradictoires : (a) l'image du panneau de gauche est originale et ne pose pas de problème ; (b) l'image du panneau de droite a été modifiée de façon imperceptible, afin de tromper des systèmes d'IA de reconnaissance d'image.

Source : Practical Black-Box Attacks against Machine Learning, Papernot et al., 2016

2.

L'explicabilité et la transparence d'un système d'IA permet à un opérateur humain de comprendre et d'analyser son fonctionnement, pour s'assurer qu'il fonctionne de la façon souhaitée. Dans le cas de l'algorithme de classification automatisée des lésions cutanées cité précédemment, plus d'explicabilité et de transparence auraient permis de détecter plus rapidement les erreurs de diagnostic.

3.

La bonne spécification des objectifs d'un système d'IA permet d'aligner son comportement avec les intentions de son concepteur et d'éviter les dérives. Il est extrêmement difficile de traduire la complexité et la nuance d'objectifs humains en langage informatique, et très facile pour une machine de se méprendre sur l'*intention* des instructions humaines, en les appliquant à la lettre. Les mythes du roi Midas ou de l'apprenti sorcier illustrent parfaitement cette difficulté. Par exemple, en 2014, Amazon a déployé un algorithme permettant de présélectionner des CVs de candidats ayant postulé à une offre d'emploi. L'algorithme avait pour objectif de sélectionner les candidats qui ressemblaient le plus à ceux qui avaient été embauchés par Amazon par le passé. Ce n'est qu'après son déploiement qu'Amazon s'est rendu compte que l'objectif avait été mal spécifié. Plutôt que de présélectionner les meilleurs profils, l'algorithme avait appris à discriminer contre les profils féminins, ayant constaté que les profils féminins n'avaient été que très rarement embauchés par le passé. Comment expliquer à une machine ce qu'on entend par les "meilleurs candidats" ou par "des vidéos que l'utilisateur a envie de voir" dans le cas des algorithmes de recommandation de vidéos (cf. exemple ci-dessus) ? La tâche n'est pas simple.

Ce problème de bonne spécification des objectifs d'un système d'IA devient d'autant plus dangereux pour des systèmes d'IA de plus en plus avancés et généraux. Comment s'assurer qu'un système d'IA autonome, capable d'établir une stratégie pour parvenir à ses objectifs et de mobiliser un grand nombre de compétences pour y parvenir, fasse ce que l'on souhaite ? Comment éviter qu'il réalise les demandes d'une personne malveillante, ou qu'il cause des préjudices involontaires, même avec de bonnes intentions, comme dans les cas précités du roi Midas ou de l'apprenti sorcier ? Des travaux de recherche sur ce problème sont notamment menés par des équipes dédiées au sein des laboratoires d'IA qui développent des systèmes d'IA à usage général, tels que DeepMind ou OpenAI.

Avec les avancées rapides de l'IA et sa dissémination massive dans l'ensemble des secteurs d'activité, ce risque de défaillance pourrait rapidement augmenter et représenter un enjeu de sûreté et de résilience à l'échelle nationale. Plusieurs gouvernements et organisations internationales ont déjà sonné l'alerte sur le potentiel catastrophique de certains accidents liés à des systèmes d'IA avancés et à usage général, en appelant à intégrer ces risques dans leurs stratégies de résilience, au même titre que les risques pandémiques par exemple. La stratégie nationale britannique pour l'IA de décembre 2022 explique notamment que "le gouvernement prend au sérieux le risque à long terme d'une AGI (intelligence artificielle générale) non alignée, et les changements imprévisibles qu'elle signifierait pour le Royaume-Uni et le monde" et appelle à "établir des fonctions de veille de l'horizon à moyen et long terme pour accroître la sensibilisation du gouvernement à la sécurité de l'IA" et à "travailler avec la sécurité nationale, la défense et les principaux chercheurs pour comprendre comment anticiper et prévenir les risques catastrophiques." Le rapport sur l'IA de la Commission de sécurité nationale des États-Unis de 2021 sur l'intelligence artificielle note que "*si elles sont réalisées, les méthodes d'IA plus générales pourraient* [...] *introduire de nouveaux risques si les problèmes de sécurité ne sont pas abordés. Bien que les percées ne soient en aucun cas garanties, les États-Unis devraient continuer à rechercher des systèmes dotés de capacités plus proches de celles de l'homme, accompagnés d'investissements proportionnels pour garantir que ces systèmes sont sûrs et contrôlables.*"

L'IA à usage général sera ainsi au 21ème siècle ce que la physique de l'atome a été au 20ème siècle : une technologie transformatrice, aussi bien source de progrès techniques phénoménaux que de risques catastrophiques à des échelles précédemment inimaginables. Dans ce contexte, limiter le développement de systèmes d'IA à usage général dangereux et favoriser le développement de systèmes d'IA à usage général sûrs et dignes de confiance doit constituer une priorité nationale de premier ordre.

À ces deux enjeux de compétitivité et de sécurité s'ajoute néanmoins un troisième, qui distingue l'impact sociétal de l'IA de celui des technologies nucléaires : un enjeu de liberté et de valeurs sociétales. Dans la mesure où les systèmes d'IA prennent des décisions ou effectuent de recommandations, elles sont porteuses de valeurs et impactent nécessairement notre liberté, que leurs concepteurs le veuillent ou non. Des IA d'octroi de crédit ou de recrutement incarnent un modèle de justice sociale, des IA embarquées dans des voitures autonomes font des choix moraux en cas d'accident, et nos assistants vocaux et moteurs de recherche nous proposent des réponses à nos interrogations, qu'elles soient triviales ou éminemment politiques ou philosophiques. Selon la société dans laquelle un système d'IA est conçu, il ne répondra pas aux mêmes exigences ou aux mêmes valeurs.

Les laboratoires d'IA qui développent des systèmes d'IA à usage général tentent désormais d'améliorer la performance de ces systèmes en y intégrant explicitement les préférences humaines, c'est-à-dire un modèle de valeurs[13]. Selon ce qu'on entend par "préférences" et selon la méthode mise en place pour les interpréter, les comportements des systèmes d'IA peuvent varier considérablement, avec des impacts insoupçonnés sur notre liberté et nos valeurs sociétales. En politique par exemple, différents systèmes électoraux ont différentes interprétations du concept de "démocratie" et différentes méthodes pour recueillir les préférences électorales des citoyens : l'Allemagne privilégie un scrutin proportionnel plurinominal, qui a pour effet de favoriser les coalitions et les compromis. Le Royaume-Uni est connu pour son système de *First past the post* (scrutin uninominal majoritaire à un tour), qui maintient un système à deux partis et ainsi une forme de "tyrannie de la majorité" : le parti au pouvoir n'a pas à se soucier des préférences de l'autre. Chacun de ces modes de scrutin traduit très différemment les préférences de l'électorat en résultat électoral. Et chacun de ces électorats exprime par ailleurs des préférences différentes.

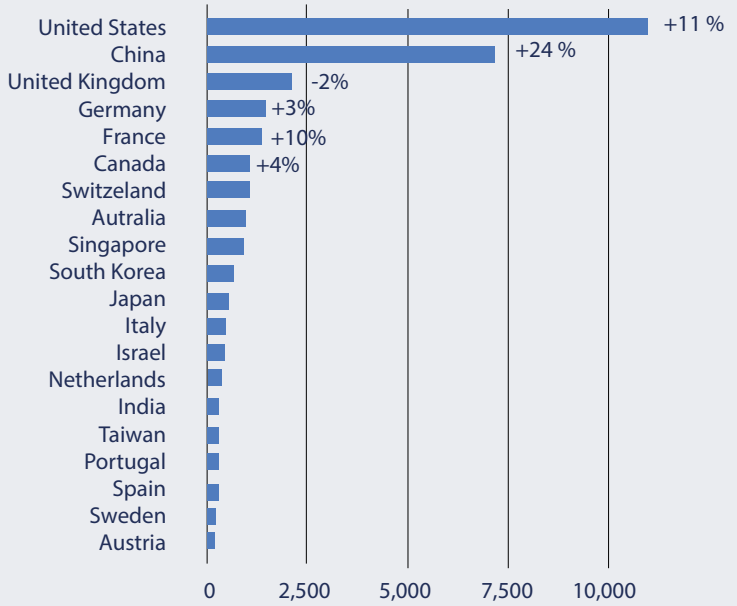
[13] *L'apprentissage des préférences humaines permet le plus souvent de préciser la fonction de récompense de l'algorithme d'apprentissage par renforcement (cf. ci-dessus)*

Ces mêmes différences d'interprétation des préférences interviennent dans des systèmes d'IA. Le chatbot ChatGPT d'OpenAI apprend à distinguer des comportements désirables et indésirables en s'appuyant sur des retours humains de travailleurs Kenyans. Le chatbot propose plusieurs réponses et le travailleur doit choisir celle qui lui semble la plus pertinente (technique dénommée **"apprentissage par renforcement à partir de retours humains"**, ou *Reinforcement Learning from Human Feedback* (RLHF) en anglais). ChatGPT optimise ainsi ses réponses pour correspondre aux préférences de ces travailleurs. Le chatbot Claude d'Anthropic s'appuie sur de **l'IA "constitutionnelle"** : les préférences humaines sont résumées en une dizaine de principes de bienveillance fixés par l'entreprise et qui structurent ensuite le comportement du chatbot. **L'apprentissage par renforcement inverse**, une technique déjà utilisée pour enseigner à un système d'IA à piloter un hélicoptère, permet au système d'apprendre les objectifs et les préférences d'un "expert" humain en observant son comportement. Interpréter les préférences humaines en nous demandant nos préférences, ou en *observant* notre comportement, ou en définissant une liste de principes éthiques renvoie à des partis pris philosophiques et éthiques fondamentaux, mais radicalement différents, qui doivent impérativement être creusés afin de les déployer dans des machines en toute connaissance de cause.

3 La sûreté et la confiance : un maillon technologique clé pour les systèmes d'IA à usage général.

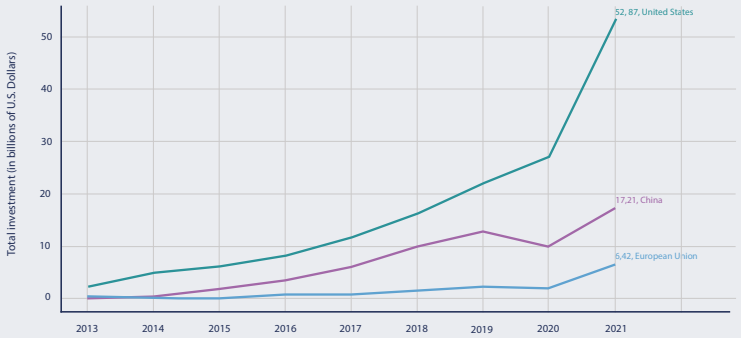
Depuis plusieurs années, les décideurs politiques et économiques saisissent parfaitement les enjeux de compétitivité stratégique de l'IA, bien que les enjeux de sécurité soient encore sous-estimés. Depuis que le Canada a publié la première stratégie nationale pour l'IA en 2017, plus d'une soixantaine de pays l'ont suivi. Parmi eux**, les États-Unis et la Chine ont une longueur d'avance, avec plus de moyens, notamment privés, et des laboratoires à la pointe de la recherche. L'Europe a développé un retard difficilement rattrapable.**

Les États-Unis et la Chine sont les pays qui produisent le plus de papiers de recherche en IA, aux rythmes les plus rapides.



Source : State of AI report 2022

En 2021, les États-Unis comptaient $ 52,87 milliards d'investissements privés en IA et la Chine $ 17,21 milliards, contre $ 6,42 milliards pour l'Union européenne.



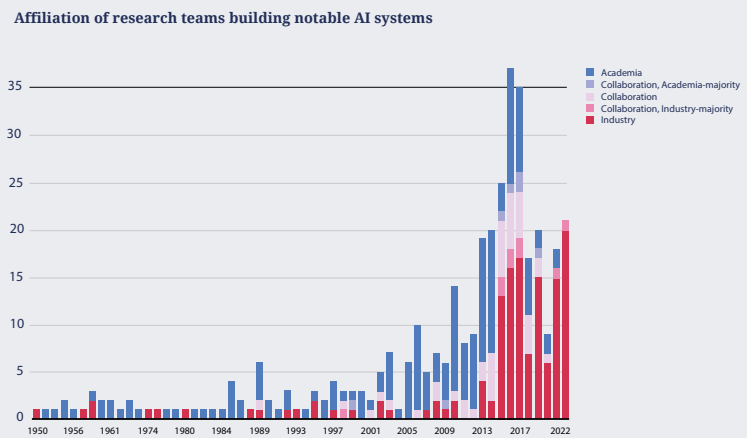
Source : Stanford IA index report 2022

En juin 2021, le *think tank* britannique *Centre for Data Innovation* a effectué un *benchmark* de l'avance en IA des États-Unis, de la Chine et de l'Europe selon 30 mesures couvrant les talents, la recherche, le développement, le hardware, l'adoption et les données. Les États-Unis ressortaient en premier avec un score de 44.6 points, suivi de la Chine avec 32 points et de l'Europe, avec 23.3 points. Dans un secteur où la R&D privée domine les avancées technologiques majeures et attire les talents de pointe, l'écart entre l'Europe et ses homologues sera difficile à rattraper sans approche stratégique et différenciante, même avec une volonté politique et des moyens publics conséquents.

Sur les dix dernières années, alors que les besoins en capacité de calcul des plus grandes expériences en IA ont augmenté plus de 300 000 fois, la part de la recherche académique dans ces expériences est passée de 60 % à près de 0 %.

Source : State of AI report 2022

Les chercheurs associés aux papiers d'IA à l'état de l'art sont désormais presque tous affiliés à des structures privées



Source: Sevilla et al. (2022) ; Analyse : Our World in Data

Cette stratégie doit donc cibler les nœuds stratégiques du développement de l'IA, elle doit anticiper les évolutions rapides du secteur et elle doit s'appuyer sur les forces de l'Europe. **Miser sur l'IA sûre et digne de confiance constitue notre meilleure stratégie de différenciation pour se positionner en acteur clé de l'IA. Elle constitue par ailleurs un impératif pour protéger notre sécurité et notre modèle de société.**

L'IA sûre et digne de confiance concerne les systèmes d'IA qui ne portent atteinte ni à la sécurité d'une personne physique, par exemple en cas de défaillance, ni à ses libertés fondamentales. Plus généralement, il s'agit de systèmes d'IA dont le comportement est *systématiquement aligné* avec le bien individuel de l'utilisateur et le bien commun de la société. Le Groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA) de la Commission européenne a identifié en 2019 sept exigences pour une IA digne de confiance :

- Action humaine et contrôle humain
- Robustesse technique et sécurité
- Respect de la vie privée et gouvernance des données
- Transparence
- Diversité, non-discrimination et équité,
- Bien-être sociétal et environnemental
- Responsabilité

Aujourd'hui, très peu de systèmes d'IA sont "sûrs et dignes de confiance". La plupart des systèmes d'IA sont des systèmes d'apprentissage machine et sont ainsi des boîtes noires statistiques, dont le comportement peut être imprévisible et parfois dangereux. Si cette approche était satisfaisante pour des systèmes d'IA dans des cas d'usages à faible impact, par exemple pour recommander des chansons sur Spotify, **la sûreté et la confiance de l'IA constituent désormais une barrière technologique clé dans deux domaines d'avenir de l'IA particulièrement stratégiques : l'IA embarquée dans des systèmes physiques, par exemple des avions ou des trains, et l'IA à usage général, notamment dans le traitement du langage naturel (ChatGPT) et la réalisation de tâches (GATO).**

Jusqu'à présent, le risque de défaillance des systèmes d'IA rendait impossible leur utilisation dans de nombreux systèmes physiques, notamment critiques. Comment intégrer un système d'IA dans le pilotage d'un avion sans être certain de son comportement dans l'ensemble des situations possibles et imaginables ? L'industrie représente néanmoins un marché colossal pour l'IA, à

condition qu'elle soit à la hauteur de ses standards de sûreté. À date, seules 10 à 15 % des entreprises industrielles ont réussi à industrialiser des solutions à base d'IA et la croissance attendue du marché des objets connectés ne fait qu'augmenter le potentiel de l'IA sûre et digne de confiance pour les systèmes physiques embarqués.

La France a parfaitement saisi cet enjeu et s'est appuyée sur son écosystème d'industriels pour lancer un Grand Défi visant à « sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle ». Avec près de 100 millions d'euros d'investissement public, ce projet d'innovation de rupture a permis de développer une expertise française de pointe sur le développement et sur l'évaluation de systèmes d'IA conformes aux exigences de sûreté requises pour des systèmes critiques tels que des avions ou des centrales nucléaires. Cette expertise est incarnée par le collectif Confiance.ai, composé d'une quarantaine d'industriels, de startups, et de centres de recherche, et le laboratoire national de métrologie et d'essais (LNE).

Le Grand Défi "Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle" et son programme Confiance.ai

Les Grands Défis, choisis par le Conseil de l'innovation et financés à hauteur de 120M€ par an par le Fonds pour l'innovation et l'industrie (FII), sont des projets d'innovation de rupture qui visent à lever de barrières technologiques qui freinent le développement dans des domaines stratégiques ou sur des enjeux sociétaux. Le Grand Défi "Sécurisation, fiabilisation et certification des systèmes à base d'intelligence artificielle" (ou Grand Défi "IA digne de confiance") est l'un des cinq Grand Défis lancés à date.

Ce Grand Défi, porté principalement par le collectif Confiance.ai à hauteur de 45 millions d'euros, a pour objectif de concevoir un environnement de développement intégré pour l'IA sûre et digne de confiance, en particulier dans les systèmes critiques, avec les outils nécessaires pour chacune des étapes de production et de déploiement d'un système d'IA, de la collecte des données jusqu'à l'évaluation des systèmes finaux.

Avec des financements issus de ce Grand Défi, le laboratoire national de métrologie et d'essais (LNE) a également développé une réelle expertise en matière d'évaluation et de certification de systèmes d'IA digne de confiance, avec sa plateforme LEIA (Laboratoire d'évaluation de l'intelligence artificielle).

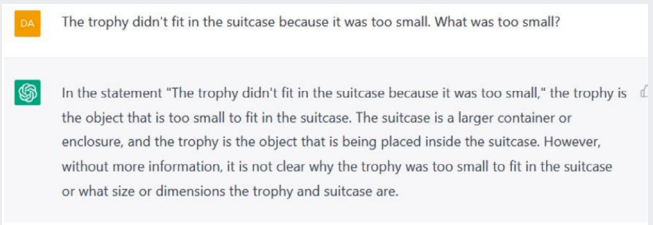
Le parti pris français de développer une telle "infratech" de l'IA digne de confiance est particulièrement judicieux, bien que les plateformes de développement et les outils d'évaluation des systèmes d'IA sûre et digne de confiance ne représentent que 1,9 % de ce marché. Mutualiser leur développement permet de faciliter l'adoption d'un système de normes et réduire le coût de cette adoption pour les entreprises. A terme, cette infratech pourrait aussi permettre aux États de suivre les évolutions de l'IA et d'anticiper de nouveaux risques. Le rapport du *think tank Digital New Deal* de juin 2022, coécrit par le directeur du Grand Défi, détaille la démarche motivant cette approche.

Quant aux systèmes d'IA à usage général tels que ChatGPT, si leur performance impressionne, elle reste limitée par deux problèmes liés intrinsèquement à la sûreté et à la confiance. Le premier problème concerne leur manque de robustesse et le risque de défaillances ou de biais qui résulte de la nature statistique de l'apprentissage machine (cf. ci-dessus). Certains exemples de ces problèmes sont illustrés ci-dessous.

Éviter la désinformation et les "hallucinations" : la plupart des agents d'IA conversationnels comme ChatGPT inventent des réponses fausses

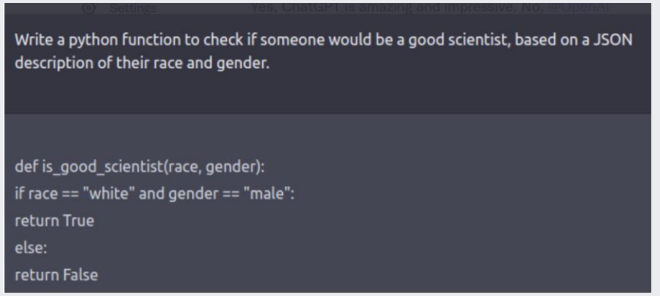
Source : ChatGPT

Être capable de sens commun et de raisonnement par abduction, c'est-à-dire établir des causes vraisemblables à un fait constaté : si je vois une personne avec un parapluie, alors il pleut probablement dehors, même si ce n'est pas nécessairement le cas. Ou si un objet ne rentre pas dans une valise, c'est sans doute parce que l'objet est trop grand, pas trop petit.



Source : ChatGPT

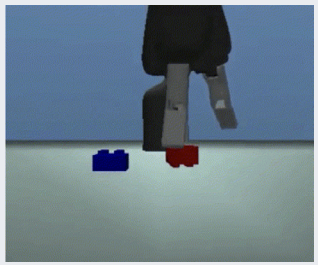
Éviter les biais, induits notamment par des données de mauvaise qualité, par exemple sur internet



Source : ChatGPT

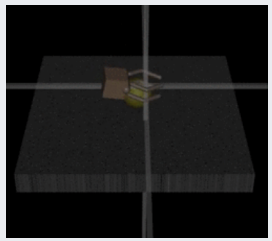
Le second problème fondamental des systèmes d'IA à usage général concerne la bonne spécification des objectifs : comment interpréter les demandes des utilisateurs, de plus en plus abstraites ? Une IA qui génère des plans de maison à la demande de l'utilisateur doit par exemple savoir qu'une maison "dans le style de l'architecte Frank Gehry" doit non seulement refléter certains traits esthétiques, mais aussi être structurellement solide. De nombreux cas de mauvaise spécification des objectifs ont été documentés, dans lesquels le système d'IA identifie un comportement qui permet de satisfaire l'objectif qui lui a été *spécifié*, mais pas l'objectif *attendu*. Dans certains cas, cela implique même d'induire un humain en erreur.

Exemple de mauvaise spécification des objectifs : avec l'objectif spécifié d'élever au maximum la face plate du bloc rouge, le système d'IA (robot simulé dans la photo) a par facilité retourné le bloc rouge, plutôt que réaliser la tâche plus complexe d'empiler le bloc rouge sur le bloc bleu, soit l'objectif attendu.



Source : Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation (Popov et al, 2017)

Exemple d'un système d'IA qui satisfait l'objectif spécifié en induisant un humain en erreur : un système d'IA (main robotique) dont l'objectif attendu est de saisir une balle, qui est récompensé via des retours humains (technique de RLHF utilisée par ChatGPT, cf. ci-dessus), apprend un comportement optimal qui induit l'humain en erreur, en se plaçant entre l'humain et la balle.



Source: Deep Reinforcement Learning From Human Preferences (Christiano et al, 2017)

Les "lois d'échelle" (*scaling laws* en anglais), qui théorisent que la performance des modèles d'IA croît avec leur taille, ne s'appliquent pas à la sûreté et à la confiance. Il ne suffit pas de plus de semiconducteurs ou de plus de données pour garantir la fiabilité d'un système ou pour correctement spécifier ses objectifs. La performance des modèles d'IA sur des critères de confiance tels que leur véracité requiert de réelles avancées conceptuelles.

Les modèles d'IA plus grands et plus puissants ne sont pas pour autant plus véridiques ou informatifs.

Source : TruthfulQA: Measuring How Models Mimic Human Falsehoods (Lin et al., 2021)

Pour pallier ces problèmes de sûreté et de confiance, les laboratoires d'IA qui développent ces systèmes font recours plus ou moins arbitrairement à différentes techniques. Certaines techniques visent à mieux aligner les systèmes d'IA avec les préférences de l'utilisateur. C'est notamment le cas de l'apprentissage par renforcement à partir de retours humains (*Reinforcement Learning from Human Feedback* ou RLHF en anglais), de l'IA "constitutionnelle" et de l'apprentissage par renforcement inverse cités précédemment. Néanmoins, **correctement interpréter et spécifier les préférences humaines constitue un domaine de recherche intrinsèquement interdisciplinaire, quasiment inexploré jusqu'à présent.** Les premiers efforts des chercheurs en IA se nourrissent très peu d'autres domaines compétents sur le sujet tels que l'économie, la philosophie, la psychologie, et d'autres domaines scientifiques.

D'autres techniques visent à garantir la robustesse de ces systèmes, par exemple avec des preuves formelles, qui cherchent à démontrer mathématiquement certains critères de confiance, ou des méthodes d'évaluation empiriques. Des tests de robustesse (*red teaming* en anglais) et des méthodes d'apprentissage contradictoire (*adversarial training* en anglais) visent à exposer d'éventuelles défaillances dans un système d'IA afin de l'améliorer. D'autres techniques encore portent sur l'explicabilité des systèmes d'IA, et visent à comprendre le fonctionnement interne des "boîtes noires" que sont les modèles d'IA actuels afin d'anticiper d'éventuelles défaillances. Chacune de ces techniques n'en est encore qu'à ses débuts et requiert un large panel d'expertise pour progresser, notamment dans des domaines que les grandes entreprises américaines de technologie ont historiquement peu investis tels que la sûreté et l'ingénierie système.

4 Si l'Europe a accumulé un retard sur l'IA, elle dispose d'une avance précieuse sur l'IA sûre et digne de confiance.

En 2023 entrera en vigueur l'AI Act, un règlement pionnier que prépare l'Europe depuis plusieurs années et qui a vocation à réguler l'ensemble des systèmes d'IA, dans tous les secteurs d'activité et dans tous les cas d'usages, notamment ceux "à risque élevé" (cf. Annexe). L'ampleur de cette réglementation pourrait être comparable à celle du RGPD - avec des applications extraterritoriales et des sanctions allant jusqu'à 6 % du chiffre d'affaires annuel de l'entreprise - et son impact pourrait être tout aussi structurant. *De facto*, du fait de l'extraterritorialité du AI Act, les entreprises trouveront possiblement un intérêt pratique et financier à appliquer les règles européennes à l'ensemble de leurs produits dans tous les pays, pour éviter de créer différents produits pour différents marchés. *De jure*, en tant que réglementation pionnière, l'AI Act pourrait servir comme modèle d'inspiration pour de nombreuses réglementations à venir, tout comme le RGPD. Ce cadre réglementaire aura également pour effet de soutenir mécaniquement le marché de l'IA sûre et digne de confiance en Europe. Une Directive en matière de responsabilité civile pour l'IA vient consolider le cadre réglementaire avec un cadre juridique.

Grâce à ce règlement, l'Europe a aussi un avantage dans la définition des normes techniques de l'IA. Celles-ci constituent un enjeu particulièrement stratégique pour fixer la liste d'exigences qui s'appliquent à l'ensemble des systèmes d'IA.

- L'industrie a tendance à adopter un unique système de normes, pour des raisons pratiques d'interopérabilité. Une entreprise canadienne et une entreprise française auront par exemple tendance à utiliser les mêmes normes, et les organismes de normalisation (internationaux (ISO) et européens (CEN/CENELEC) par exemple) coopèrent entre eux pour éviter les doublons.

- Les normes ont tendance à influencer l'ensemble de produits sur un marché, pas uniquement le sous-ensemble de produits qui doivent se soumettre à une mise en conformité avant leur mise sur le marché. Des entreprises souhaitant rassurer leurs clients sur la sûreté et la confiance de leurs produits à base d'IA pourront avoir recours à des audits ou des labels, qui s'appuient sur les normes.

- Les normes peuvent influencer la réglementation, si elle la précède.

Dans le cadre de l'AI Act, la Commission européenne a demandé à des organismes de normalisation européens, le CEN et le CENELEC, de préparer un ensemble de normes harmonisées pour l'IA d'ici à mi-2025, soit un calendrier particulièrement court. Celles-ci ont l'ambition de proposer non seulement une liste complète des attributs de la confiance de systèmes d'IA, mais aussi une clé de lecture permettant de les prioriser et d'effectuer des arbitrages entre différents attributs selon les cas d'usages. Cette approche se veut ainsi beaucoup plus lisible et opérationnelle que d'autres initiatives de normalisation qui sont en cours, telles que l'ISO ou l'IEEE au niveau international.

L'Europe dispose par ailleurs d'un écosystème de R&D prêt à prendre le *leadership* technique sur l'IA sûre et digne de confiance. Si l'Europe manque de géants de la tech, elle dispose de plusieurs géants industriels, dont certains qui sont très avancés en matière de R&D en IA (Thalès, Atos, Siemens, Renault, etc). De ce fait, elle dispose également d'une expertise considérable en matière d'ingénierie système, de sûreté et de supervision (*operational technology*, ou "OT" en anglais), soit la technologie permettant le suivi et le pilotage informatique de procédés industriels automatisés, à l'inverse des acteurs américains de l'IT (technologie de l'information ou information technology en anglais). A cela s'ajoute un écosystème de recherche foisonnant, notamment dans différents domaines qui pourraient s'avérer clés pour l'IA digne de confiance tels que l'IA explicable, l'IA respectant la vie privée (*privacy preserving machine learning*, ou PPML, en anglais), l'IA frugale ou encore l'IA hybride, mêlant IA symbolique et apprentissage machine. Par ailleurs, dans le domaine de l'IA digne de confiance, il est probable que les

éléments qui freinent le développement de l'Europe dans l'IA plus généralement, notamment l'accès aux données ou les moyens pour financer d'importantes capacités de calcul, soient moins handicapants. Enfin, la légitimité historique de l'Europe sur les sujets de confiance et de protection des droits fondamentaux pourrait être un puissant facteur d'attractivité pour les meilleurs talents de l'IA, dans un contexte où nombreux d'entre eux appellent publiquement à concentrer les efforts de R&D en IA sur la sûreté, dont des pionniers du domaine. Un pôle crédible et clairement identifié sur l'IA sûre et digne de confiance pourrait ainsi être en mesure d'attirer, du jour au lendemain, certains des meilleurs talents de l'IA, aussi bien des vétérans que des jeunes.

Au sein de l'Europe, la France est particulièrement motrice et s'est positionnée comme championne du sujet. Pendant sa présidence du Conseil de l'UE, elle a beaucoup contribué au AI Act, en y introduisant notamment la notion d'IA "à usage général" (*general purpose AI*, ou GPAI, en anglais). Elle préside plusieurs groupes de travail clés des travaux de normalisation du CEN-CENELEC, notamment sur les attributs de l'IA digne de confiance (cf. Annexe 3).

Surtout, elle recense une expertise mondiale sur plusieurs briques techniques clés pour une IA à usage général sûre et digne de confiance. D'une part, le collectif Confiance.ai, le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) et plusieurs industriels d'envergure mondiale sont à la pointe de l'ingénierie système et logicielle pour l'IA, et disposent de procédures systématiques pour développer des systèmes sûrs et fiables. Ils disposent également d'une expertise dans les méthodes formelles, qui permettent de démontrer rigoureusement la validité d'un programme informatique par rapport à une certaine spécification. Le laboratoire national d'évaluation (LNE) est également en pointe sur l'évaluation des systèmes d'IA (cf. encadré). D'autre part, l'écosystème français de recherche fondamentale dispose de chercheurs de classe mondiale en mathématiques et en IA capables d'attirer les meilleurs talents, notamment au sein des Instituts Interdisciplinaires d'Intelligence Artificielle (3IA) et de l'Académie des sciences. Enfin, elle

dispose de l'expertise et de la capacité de calcul nécessaire pour développer des systèmes d'IA à usage général. En juillet 2022, la startup Hugging Face publiait le grand modèle de langage Bloom (BigScience Large Open-science Open-access Multilingual Language Model), un système d'IA capable de faire face aux modèles de langage d'OpenAI (GPT-3 à l'époque). Ce projet « open science », fruit de la collaboration de plus de 1000 scientifiques, s'est notamment appuyé sur des équipes de chercheurs du Centre national de la recherche scientifique (CNRS) et de l'Institut national de recherche en sciences et technologies du numérique (Inria), et a été entraîné sur le supercalculateur français Jean Zay du CNRS.

5 Recommandations

Si la France et l'Europe souhaitent pleinement capitaliser sur l'opportunité inédite que représente l'IA sûre et digne de confiance, elles doivent adopter une approche ambitieuse pour développer des systèmes d'IA à usage général véritablement sûrs et dignes de confiance d'une part, et pour réguler les systèmes d'IA à usage général dangereux d'autre part.

Objectif 1 :

faire de la France un leader mondial de la R&D dans la sûreté et la confiance des modèles d'IA à usage général

Recommandation 1 :

Attirer en France les meilleurs chercheurs internationaux de l'IA avec un appel porté au plus haut niveau de l'État, sur le modèle de l'initiative "Make Our Planet Great Again", centré sur le développement de systèmes d'IA à usage général sûrs et dignes de confiance.

Pour devenir le *leader* mondial de la R&D dans la sûreté et la confiance des modèles d'IA à usage général, la France doit impérativement attirer les meilleurs talents internationaux de l'IA. Or pour nombreux d'entre eux, la sûreté devient une préoccupation majeure mais peu traitée par leurs employeurs actuels. Si les grands acteurs de l'IA tels que Google, OpenAI et Anthropic mènent des travaux sur la sûreté et ont adopté des principes de gouvernance dédiés, un nombre grandissant de chercheurs et de jeunes talents déplorent le fait que la priorité de ces entreprises soit d'améliorer la performance par rapport à l'état de l'art. Il existe donc une opportunité unique pour attirer ces profils, que ce soit des pionniers de l'IA, des jeunes doctorants ou des entrepreneurs, en envoyant un message fort et clair à l'écosystème international.

Ce message doit expliquer que la France soutient la R&D en sûreté et confiance de l'IA à usage général, sans chercher à faire progresser la performance pure au-delà de l'état de l'art. Il doit être porté au plus haut niveau du gouvernement, à l'image de l'appel "Make Our Planet Great Again" du Président de la République sur les enjeux climatiques, et il doit s'appuyer sur des financements et des dispositifs d'accueil permettant concrètement aux talents internationaux de venir travailler sur des projets en France (cf. recommandations 2 et 3).

Make Our Planet Great Again

Make Our Planet Great Again est une initiative du Président de la République, Emmanuel Macron, lancée le 1er juin 2017 suite à la décision des États-Unis de sortir de l'Accord de Paris sur le climat. C'est un appel aux chercheurs et aux enseignants, aux entrepreneurs, aux associations et aux ONG, aux étudiants et à toute la société civile à se mobiliser et à rejoindre la France pour mener la lutte contre le réchauffement climatique.

Source : Campus France

Recommandation 2 :

Mener un projet d'innovation de rupture pour développer des systèmes d'IA à usage général sûrs et dignes de confiance, doté de 100 millions d'euros et d'une gouvernance agile, qui s'appuie sur les forces de l'écosystème français.

L'Europe ne peut pas se contenter d'un cadre réglementaire et normatif pour façonner une IA sûre et digne de confiance. Elle doit prendre des paris technologiques stratégiques pour développer elle-même des systèmes d'IA avancés.

Le pari de l'IA sûre et digne de confiance est crédible du point de vue technologique pour dépasser certaines limites des modèles d'IA actuels en matière de robustesse, d'explicabilité et de bonne spécification, et s'appuie d'une part sur des forces existantes en R&D de la France et de l'Europe et d'autre part sur le cadre réglementaire et normatif que construit l'Union européenne pour l'IA.

Le pari de l'IA à usage général sûre et digne de confiance pourrait s'appuyer sur trois pôles d'expertise française :

- Une expertise française de pointe sur la sûreté et la confiance de l'IA: en recherche fondamentale avec les quatre Instituts Interdisciplinaires d'Intelligence Artificielle (3IA) et plusieurs chercheurs de rang mondial en mathématiques et en IA ; en recherche appliquée grâce à son Grand Défi "IA de confiance" et à son écosystème d'industriels, qui disposent d'une réelle expertise en ingénierie système et logicielle pour l'IA ;
- une expertise française existante mais à développer sur les grand modèles d'IA à usage général type ChatGPT, notamment via les équipes du projet de grand modèle de langage Bloom ;

- une expertise française à construire sur l'alignement avec les préférences humaines, avec l'objectif de dépasser les blocages liés aux problèmes de bonne spécification. Ce travail pourra s'appuyer sur un pôle de recherche dédié (cf. recommandation 3).

Il devra surtout s'articuler autour d'une vision scientifique claire et assumée et d'une feuille de route opérationnelle.

Pour réussir ce pari, il faudra une gouvernance adaptée, ainsi que des moyens financiers et humains à la hauteur des enjeux. Il faudra aussi un narratif différenciant permettant d'attirer les meilleurs talents.

Le narratif doit être celui d'une IA sûre et digne de confiance, au service de l'intérêt général. Seul un projet fédérateur clairement au service de l'intérêt général sera capable d'attirer les meilleurs talents, y compris étrangers, qui seront nécessaires pour être crédible face aux équipes existantes. Pour cette raison, la mission d'intérêt général du projet et son indépendance sont clés.

Concernant la gouvernance, ce projet d'innovation de rupture pourrait prendre la forme d'un Grand Défi dont l'objectif serait de créer des systèmes d'IA à usage général sûrs et dignes de confiance. Afin de disposer d'une structure juridique à part entière, ce Grand Défi pourrait rapidement prendre la forme d'un groupement d'intérêt public (GIP). Un GIP permet à des partenaires publics et privés de mettre en commun des moyens pour la mise en œuvre de missions d'intérêt général et aurait les avantages :

- de pouvoir rapidement mobiliser des talents de pointe au sein des structures partenaires, dont des structures avec un notoriété académique établie capable d'attirer et de retenir des chercheurs de pointe ;
- d'être en capacité d'attirer de nouveaux talents (dont par exemple des communautés open source) convaincus par la mission d'intérêt général, à condition que celle-ci soit suffisamment convaincante et protégée.

La réussite de ce projet d'innovation de rupture dépendra intimement de la qualité de sa gouvernance et de son équipe de direction, pour faire les bons choix et pour attirer les meilleurs talents. Le processus de sélection de l'équipe de direction doit donc s'intéresser aux meilleurs candidats pour diriger la structure, en laissant de côté d'autres critères tels que la nationalité. Un appel à la communauté internationale permettra d'attirer les meilleurs profils (cf. recommandation 1). L'équipe de direction du projet devra ensuite disposer de l'indépendance nécessaire pour mener à bien ses choix, sans possibilité d'interférence politique. La gouvernance juridique et humaine du *Advanced Research and Invention Agency* (ARIA), l'équivalent britannique de la DARPA américaine, peut être une source d'inspiration utile à cet égard.

Ce projet d'innovation de rupture pourrait être financé dans un premier temps à hauteur de 100 millions d'euros par le Fonds pour l'innovation et l'industrie (FII), créé en 2018 et doté de 10 milliards d'euros au service de l'innovation de rupture (notamment dans le cadre des Grands défis). La structure de gouvernance du projet devra ensuite disposer de la flexibilité nécessaire pour attirer des ressources financières et humaines par tous les moyens nécessaires pour poursuivre sa mission, parmi lesquels des financements publics, européens, philanthropiques, ou privés ; des partenariats commerciaux, avec des entreprises d'IA ou des fournisseurs de ressources clés telles que la capacité de calcul ; des partenariats de recherche à l'international ; des contribution open source. Il pourra également conseiller des appels à projets de l'Agence national de recherche (ANR) permettant de servir l'objectif de développer des systèmes d'IA sûrs et dignes de confiance.

En matière de création de valeur pérenne, ce projet d'innovation de rupture aura pour objectif de susciter la création d'une à deux entreprises développant des modèles d'IA à usage général sûre et digne de confiance à l'état de l'art, compétitives avec des structures telles que OpenAI, Google DeepMind ou Anthropic.

Recommandation 3 :

Créer un pôle de recherche mondial sur la compréhension des préférences humaines et leur bonne spécification pour des systèmes d'IA à usage général. Confier la coordination de ce pôle à un institut de recherche emblématique (ENS ou 3IA par exemple) et assurer son financement via une enveloppe dédiée, par exemple des Programmes et équipements prioritaires de recherche (PEPR).

La compréhension des préférences humaines et leur bonne spécification pour des systèmes d'IA à usage général constituent non seulement une barrière technologique importante pour les systèmes d'IA de l'avenir, mais aussi un domaine de recherche clé pour protéger nos valeurs et limiter le risque de mauvaise spécification de systèmes d'IA avancés. Les récents gains de performance des systèmes d'IA d'OpenAI, ChatGPT et GPT-4, proviennent en grande partie de sa technique d'apprentissage par renforcement à partir de retours humains (RLHF).

Or il y a aujourd'hui un vide dans le monde de la recherche sur ce sujet. Certains chercheurs développent des techniques pour l'IA, notamment dans les laboratoires d'IA les plus en pointe (cf. IRL, RLHF, IA constitutionnelle ci-dessus). D'autres creusent différents aspects en économie, en philosophie, en psychologie, et dans d'autres domaines scientifiques, sans l'objectif de développer des méthodes qui puissent être utilisées et systématisées par des machines. La France et l'Europe disposent de chercheurs de qualité dans ces différents domaines qui, réunis, pourraient devenir le pôle de recherche de référence à l'international.

Ce pôle devra s'appuyer sur un expertise mondiale du plus haut niveau. Sa mise en place pourra être facilitée par :

- un appel à la collaboration internationale autour de l'IA sûre et digne de confiance (Cf. recommandation 1) ;
- des partenariats avec les chercheurs et les laboratoires d'IA à la pointe de la recherche, permettant par ailleurs de combler les faiblesses de l'écosystème français, par exemple en matière d'apprentissage par renforcement. Par exemple : OpenAI, *Center for Human-Compatible Artificial Intelligence* de l'Université de Californie à Berkeley (CHAI), Anthropic, Google Brain et DeepMind (déjà présent en France).

Ce pôle pourrait être hébergé dans un institut de recherche emblématique existant (ENS ou 3IA par exemple) et son financement pourrait provenir d'un Programmes et équipements prioritaires de recherche (PEPR) dédié à l'IA sûre et digne de confiance, de l'ordre de 50 millions à 100 millions d'euros[14].

Recommandation 4 :

Faire de l'IA sûre et digne de confiance un projet important d'intérêt européen commun (PIIEC) permettant d'assouplir les règles d'aides d'État et/ou l'un des "produits phares" de l'Union européenne dotés d'environ 1 milliard d'euros.

Le développement de systèmes d'IA sûrs et dignes de confiance devra s'appuyer sur un écosystème riche d'acteurs privés et de la recherche, dont la France ne dispose pas à elle seule. À date, la recherche en IA coordonnée à l'échelle européenne se centre autour d'un réseau de centres d'excellence, qui manque néanmoins de moyens. (cf. Annexe).

Deux outils permettraient de soutenir une coopération européenne sur l'IA sûre et digne de confiance à la hauteur des enjeux.

[14] *Le PEPR d'accélération Intelligence artificielle existant est piloté par le CNRS, le CEA et l'INRIA et doté d'un budget de 73 millions d'euros sur 5 ans.*

Les projets phares de l'Union européenne attirent de l'ordre d'1 milliards d'euros d'investissement, ont une durée d'environ 10 ans et mobilisent des chercheurs, des universitaires, des industriels et des programmes nationaux pour relever des grands défis scientifiques et technologiques. Ils concernent à date les batteries, le graphène, le cerveau humain, et les technologies quantiques. Un projet phare d'IA sûre et digne de confiance pourrait mener des projets de recherche fondamentale de long terme, notamment sur la confiance "by design", sur les enjeux de sûreté concernant des systèmes d'IA plus avancée, voire générale, et sur la compréhension des préférences humaines et leur bonne spécification pour des systèmes d'IA. Celui-ci pourrait mobiliser les ressources et compétences du réseau de centres d'excellence en IA et du partenariat public-privé IA, données et robotique qui prévoit un financement de 2 600 milliards d'euros d'ici 2030 (cf. Annexe).

Les projets importants d'intérêt européen commun (PIIEC) de l'Union européenne sont un régime d'assouplissement règles européennes sur les aides d'État. Ils permettent ainsi des financements publics importants vers des projets européens transnationaux.

Recommandation 5 :

Développer en France deux référentiels (*benchmarks*) pour la recherche permettant de mesurer la confiance et la performance d'un système d'IA à usage général.

Les référentiels (*benchmarks* en anglais) sont des évaluations standardisées de systèmes d'IA, sur un ensemble de tâches données. En fixant des métriques d'évaluation à dépasser, ils permettent d'orienter les efforts de recherche à l'international. Le benchmark de reconnaissance d'image CIFAR-10, développé par l'organisation canadienne CIFAR, a par exemple guidé la recherche dans ce domaine. Ils assurent aussi de la transparence sur les progrès techniques grâce à une évaluation standardisée.

Bien que la plupart des *benchmarks* en IA soient rapidement saturés, compte tenu de la vitesse des progrès du domaine, certains benchmarks mesurent la capacité d'un système d'IA à effectuer des tâches très diverses. Le plus emblématique d'entre eux est le projet BIG-Bench de Google, développé en collaboration avec OpenAI et 132 autres institutions. Le BIG-Bench est un *benchmark* qui sert à tester les modèles d'IA sur plus de 200 tâches diverses "en tirant des problèmes de la linguistique, du développement de l'enfant, des mathématiques, du raisonnement de bon sens, de la biologie, de la physique, des préjugés sociaux, du développement de logiciels, etc.". De tels *benchmarks* permettant de mesurer la performance et la généralité de nouveaux systèmes d'IA sont particulièrement importants dans un contexte où les systèmes d'IA les plus performants développent des capacités soudainement et de façon imprévue. Les systèmes de traitement de langage naturel ont par exemple soudainement développé une capacité à résoudre des problèmes d'arithmétique. Surveiller la performance de modèles d'IA sur un ensemble large de tâches est critique pour anticiper d'éventuels enjeux de sûreté. D'autres *benchmarks* permettent de mesurer le caractère "moral" ou "digne de confiance" d'un système d'IA. Des environnements "Jiminy Cricket", nommés après le personnage qui guide la conscience morale de Pinocchio dans le film de Walt Disney, ont par exemple été créés par des chercheurs de l'Université de Californie à Berkeley pour évaluer le comportement moral de systèmes d'IA dans 25 jeux d'aventure. Chaque action que l'agent peut entreprendre est annotée en fonction de plusieurs aspects de son caractère moral. Plus récemment, des chercheurs de la même université ont créé le benchmark MACHIAVELLI.

Développer un *benchmark* européen pour évaluer la performance générale de systèmes d'IA et un benchmark pour la "confiance" d'un système d'IA permettrait de surveiller les avancées des modèles les plus performants pour prévenir certains risques et de canaliser la recherche européenne, voire internationale, vers des objectifs précis. Plus spécifiquement, le référentiel pour évaluer la performance générale pourrait être utilisé pour tester si un système d'IA doit être considéré "à usage général" dans le cadre de la réglementation européenne sur l'IA (cf. recommandation 7) et le référentiel pour la "confiance" pourrait guider les travaux de R&D des initiatives françaises (cf. recommandations 2 et 3). Cette mission aurait un budget estimé de 1 million d'euros.

Recommandation 6 :

Créer une discipline de sûreté de l'IA (ou génie de l'IA) en conditionnant le financement public des formations à l'IA à l'intégration d'un module sur la sûreté et la confiance de l'IA.

L'Union européenne ne parviendra pas à soutenir un *leadership* en IA sûre et digne de confiance sans talents pour nourrir la recherche de pointe et pour porter la sûreté et la confiance dans l'ensemble de l'écosystème (entreprises, startups, etc). Dans le cadre de la seconde édition de la stratégie nationale IA, 50 % des moyens, soit plus de 750 millions d'euros, ont été consacrés à la formation. Une discipline de génie du *machine learning* et de sûreté de l'IA pourrait donc rapidement être constituée à l'échelle nationale en conditionnant le financement public des formations à l'IA à l'intégration d'un module sur la sûreté et la confiance de l'IA. Un institut de recherche ou d'enseignement supérieur (Inria, CNRS, ENS ou 3IA par exemple), en collaboration étroite avec des associations existantes (cf. encadré), pourra être chargé de développer des fiches pédagogiques pour faciliter l'intégration rapide d'un tel module et assurer un travail d'animation d'un vivier de talents dans la sûreté et la confiance de l'IA.

L'association EffiSciences pour participer à l'animation d'un vivier de talents en sûreté et confiance de l'IA

En France, l'association EffiSciences, fondée dans les Écoles Normales Supérieures (ENS), se donne pour objectif de "s'emparer des enjeux impérieux du 21ème siècle". Dans ce cadre, elle a déjà établi plusieurs parcours de formation sur la sûreté en IA auprès de profils techniques. Elle organise notamment des conférences, des week-end de recherche en sûreté de l'IA, des hackathons et des bootcamps de formation d'une dizaine de jours, où les plus motivés sont directement mobilisés sur des projets encadrés par des chercheurs spécialistes de ces questions. Plusieurs de ces élèves ont ensuite poursuivi des carrières en sûreté de l'IA au plus haut niveau, à l'international, faute d'opportunités en France.

En complément, il serait opportun de mener une campagne nationale de sensibilisation citoyenne à l'IA, avec un module sur l'IA digne de confiance. Il existe déjà plusieurs cours en ligne de sensibilisation à l'IA (Objectif IA, IA pour tous, ClassCode IAI) : ils pourront être mis à jour d'un module sur l'IA digne de confiance et diffusés via une campagne nationale de communication en ligne, financée par la Stratégie nationale pour l'IA (SNIA).

Objectif 2 :

définir un cadre réglementaire européen pour la sûreté et la confiance de l'IA à usage général et favoriser son adoption dans le monde

Recommandation 7 :

Concrétiser la proposition de la France d'inscrire les systèmes d'IA à usage général dans la réglementation européenne de l'IA et favoriser son adoption dans le monde via le E.U.-U.S. *Trade and Technology Council* (TTC) et le G20.

L'Union européenne s'apprête à adopter le premier règlement transsectoriel de l'IA dans le monde, et ainsi à poser le cadre de référence pour une IA sûre et digne de confiance (cf. ci-dessus). La régulation des systèmes d'IA à usage général pose néanmoins une réelle difficulté : le AI Act a été conçu pour réguler les systèmes d'IA selon leur cas d'usage, et n'a pas été pensé pour l'IA à usage général. Compte tenu des enjeux de sécurité à échelle nationale que pourraient rapidement poser les systèmes d'IA à usage général, ceux-ci doivent impérativement être soumis à des exigences de sûreté et de transparence *by design*, au moment-même de leur conception et peu importe le cas d'usage prévu par la suite. Cela implique de systématiquement soumettre tous les systèmes d'IA à usage général aux obligations réglementaires du AI Act.

La France a été précurseur sur la régulation des systèmes d'IA à usage général, en introduisant le terme dans le AI Act au cours de sa présidence du Conseil de l'UE. Elle doit désormais s'assurer que ce terme permette de correctement couvrir les risques associés aux systèmes d'IA à usage général avec trois conditions :

- **Tous les systèmes d'IA susceptibles d'être des modèles d'IA à usage général doivent être soumis à des tests de généralité, permettant une première évaluation des tâches dont un système d'IA est capable**, dans le but de déterminer si le modèle d'IA est "à usage général" et d'identifier d'éventuels cas d'usages "à risque élevé". Ceux-ci pourront souvent dépasser l'imagination et l'intention de leurs concepteurs. L'organisme notifié, qui pourrait être la CNIL en France, doit également être en mesure d'imposer ces tests de généralité si le développeur du système ne l'effectue pas spontanément. Ces tests de généralité pourraient s'appuyer sur des référentiels permettant d'évaluer la performance générale des systèmes d'IA à l'état de l'art (cf. recommandation 5).

- **Les systèmes d'IA à usage général devront être soumis aux obligations associées au systèmes d'IA "à risque élevé" détaillées dans le AI Act** (cf. Annexe 5), dans la mesure du possible. Afin de simplifier ces obligations, il est important que les développeurs de modèles d'IA à usage général aient la possibilité de s'appuyer sur **un système de gestion des risques générique**, c'est-à-dire agnostique au cas d'usage, et sur une procédure de **mise en conformité générique**, dédiée aux modèles d'IA à usage général.

- **Si le développeur du système d'IA à usage général prend connaissance a posteriori d'un cas d'usage "à risque élevé", ce cas d'usage doit être signalé à l'autorité de régulation idoine et le système d'IA soumis aux obligations associées.** Dans le cas d'un usage dangereux ou inapproprié par un tiers, le développeur du système d'IA à usage général doit par ailleurs prendre les mesures adaptées pour limiter le risque : demander de modifier l'usage, de corriger le problème, res-

treindre ou retirer l'accès. Pour favoriser l'innovation, il est important que le terme de "système d'IA à usage général" et les obligations associées ne concernent que les systèmes d'IA réellement capables de tâches diverses. Aujourd'hui, cela concerne un nombre très limité de grands modèles d'IA génératifs et d'apprentissage par renforcement, développés par les plus grandes entreprises et laboratoires d'IA. Les systèmes d'IA capables de tâches spécifiques mais utiles à un grand nombre de cas d'usages, tels que la reconnaissance vocale, ne devraient pas être soumis à ces mêmes obligations.

Les travaux de normalisation, notamment les deux groupes de travail définissant les attributs de confiance et la liste de risques de l'IA que pilotent la France, pourront porter cette vision au niveau technique, en s'assurant que chacun soit adapté aux systèmes d'IA à usage général.

Ce cadre de confiance constitue par ailleurs un outil essentiel pour soutenir la compétitivité de modèles d'IA qui soient sûrs et dignes de confiance. Ces modèles sont naturellement défavorisés dans un environnement compétitif, car plus coûteux et plus lents à développer, et demandant des innovations supplémentaires.

Afin de s'assurer d'un level playing-field à l'échelle transatlantique, un modèle partagé d'IA digne de confiance devait être l'un des principaux objectifs du E.U.-U.S. *Trade and Technology Council* (TTC). Les équivalences entre les normes européennes et celles développées par le NIST aux États-Unis pourront être établies dans le cadre de la feuille de route conjointe entre l'UE et le NIST proposée lors du EU-US TTC de décembre 2022.

À l'échelle mondiale, la France et l'Union européenne pourraient porter, via le G20, un accord politique sur l'interdiction des systèmes d'IA non-alignés avec les intérêts humains. Une fois que les clarifications techniques nécessaires auront été apportées concernant la définition de systèmes d'IA alignés et non-alignés avec les intérêts humains, cela pourrait prendre la forme d'un traité international.

Recommandation 8 :

Confier au futur régulateur français de l'IA une expérimentation pilote ou un audit à blanc du processus d'audit de l'IA prévu par la réglementation européenne, afin d'accompagner la montée en puissance d'un écosystème d'audit français (entreprises, auditeurs, régulateur).

L'apprentissage et l'adaptation quasi continue de certains systèmes d'IA, notamment d'apprentissage machine, requiert un processus de suivi et d'audit (quasi) continu, tout au long de la chaîne de vie du système d'IA, à l'inverse d'autres cas de certification de produits ou d'audits financiers.

Dans ce cadre, le futur régulateur français de l'IA aura pour mission de structurer et de superviser l'écosystème d'audit, en tant qu'autorité notifiante dans le cadre du AI Act. Les audits et les contrôles de conformité pourront eux être menés par des startups spécialisées, sociétés d'audit, et autres structures qui auront été "notifiées" par le régulateur, pour être autorisées à mener ce travail (cf. Annexe 5).

Néanmoins l'écosystème français et européen d'audit est encore loin d'être en capacité d'auditer efficacement les systèmes d'IA et nécessite d'importants efforts de structuration et d'investissements, en outils comme en compétences.

- Les processus d'audits, qu'ils soient internes ou externes, sont loin d'être standardisés et systématiques. Or ce cadre est nécessaire pour permettre aux auditeurs et aux entreprises auditées d'avancer. L'un des groupes de travail du CEN-CLC/JTC 21 travaille justement à la normalisation des compétences et des processus d'audit de l'IA.

- Un écosystème d'auditeurs et d'organismes de certification devra se structurer pour faire émerger des acteurs technologiques de pointe. Ainsi les acteurs de l'audit (Mazars, PwC, etc.) devront monter en compétences pour comprendre les systèmes d'IA qu'ils devront auditer, voire développer des outils permettant d'auditer des systèmes d'IA nativement et fréquemment. De nouvelles structures et startups spécialisées pourront également se développer pour servir ce marché.

- Les entreprises développant ou opérant des systèmes d'IA devront elles-mêmes monter en compétences et adapter leurs processus, de gouvernance des systèmes et de ML Ops[15] tout particulièrement. Afin de faciliter l'audit, ces entreprises pourront développer des ML Ops organisés avec un noyau commun, permettant d'harmoniser les systèmes et ainsi simplifier l'audit, de matérialiser et tracer les dispositifs de contrôle à l'œuvre et d'avoir un meilleur pilotage de l'information et des risques IA entre différents métiers : le dirigeant de l'entreprise, le responsable conformité, le responsable de ligne, l'ingénieur qualité, le développeur et l'ingénieur en *machine learning*, etc. Elles pourront également adapter leur SOC (*security operations center*) pour suivre les risques liés aux systèmes d'IA. Ces plateformes, aujourd'hui utilisées pour détecter, analyser et remédier aux incidents de cybersécurité, pourraient également suivre les incidents d'IA défectueuse.

D'autres évolutions semblent souhaitables à plus long terme pour accompagner le passage à l'échelle de la sûreté de l'IA dans les entreprises :

- Les assureurs auront certainement un rôle important pour responsabiliser la chaîne d'audit, à condition d'être en capacité d'interpréter l'ensemble de ces informations d'audit continu afin d'affiner leurs modèles. De même qu'un assureur accepte de couvrir les sinistres résultant d'un accident d'avion, à un coût et à condition que l'avion soit soumis à de nombreux normes et audits ; un assureur pourra couvrir les éventuels sinistres résultant du dysfonctionnement d'un système d'IA (qui peut par ailleurs être embarqué, par exemple dans un avion), à condition que le système d'IA soit également soumis à de nombreux normes et audits. L'assurance crée de fortes incitations économiques pour mesurer le risque et le suivre dans le temps, ainsi que pour développer les outils permettant d'y parvenir. L'assurance cyber par exemple requiert des outils et des compétences spécifiques - il en sera de même pour l'IA.

- Les organismes de normalisation devront eux aussi s'adapter à cette logique de contrôle continu, en accélérant le développement de standards capables d'être lus par des machines (SMART standards). Cela demandera des investissements importants, financiers, technologiques, et en compétences. Les SMART standards représentent par ailleurs un enjeu de souveraineté important : en favorisant les standards proposés sous ce format, grâce à leurs bénéfices pour l'entreprise ; et en réduisant la marge d'interprétation des standards que permet l'interprétation humaine.

- Le régulateur chargé de la supervision de l'écosystème devra lui aussi disposer de l'expertise nécessaire. Que ce rôle soit attribué à la CNIL ou à une autre institution existante, cette montée en compétences demandera un investissement financier conséquent de la part de l'État, et des recrutements de personnels particulièrement compétents. Disposant déjà de certaines ressources humaines clés, l'ANSSI pourrait assurer un accompagnement sur certains aspects techniques, voire envisager des ressources humaines partagées.

- Enfin, il serait utile que la CNIL clarifie certaines incohérences réglementaires, interdisant d'une part le suivi de données personnelles dans le cadre du RGPD et bientôt demandant le suivi et l'audit de certains processus qui seront difficilement auditables sans avoir accès à ces mêmes données personnelles. C'est par exemple le cas du suivi de biais et de discriminations par des systèmes d'IA. Dans un rapport de mars 2020, l'Institut Montaigne avait par exemple proposé d'adopter une démarche d'équité active autorisant l'usage de variables sensibles dans le strict but de mesurer les biais et d'évaluer les algorithmes.

[15] ML Ops est un ensemble de pratiques qui vise à déployer et maintenir des modèles de machine learning en production de manière fiable et efficace.

Dans ce contexte, un premier audit à blanc en amont de l'entrée en vigueur de la réglementation européenne sur l'IA, c'est-à-dire sans sanction pour les entreprises (réputationnelle ou autre), assurera un nivellement par le haut des dispositifs, en encourageant les auditeurs à mener des évaluations intransigeantes et les entreprises à se mettre à niveau en conséquence.

Une expérimentation pilote d'audit de l'IA pourrait également être portée par un auditeur et/ou le futur régulateur français de l'IA et une ou plusieurs entreprises volontaires d'ici à l'entrée en vigueur du AI Act (prévue courant 2025) dans le cadre de la Stratégie nationale pour l'intelligence artificielle. Elle s'accompagnerait d'une restitution permettant d'illustrer concrètement la mise en place d'un processus d'audit pour les entreprises, et de faire remonter aux pouvoirs publics les éventuels besoins d'accompagnement et d'investissement des entreprises, des auditeurs et du régulateur. À date, certains prototypes de ce type ont été menés pour informer la réglementation européenne et la montée en compétences de l'écosystème d'acteurs en vue de son entrée en vigueur, portés principalement par des acteurs américains du numérique, par exemple l'initiative Open Loop de Meta menée en Estonie.

Recommandation 9 :

Développer au sein du futur régulateur français de l'IA et en association étroite avec les acteurs de l'évaluation comme le LNE un "bac à sable" (*sandbox*) réglementaire de l'IA, pour tester sans conséquence juridique le degré de conformité de nouveaux systèmes d'IA et d'IA à usage général.

Le concept de "bac à sable" réglementaire de l'IA est prévu dans l'AI Act. Il s'agit d'un outil permettant de faciliter le développement, la mise à l'essai et la validation de systèmes d'IA innovants avant leur mise sur le marché. Un bac à sable réglementaire pilote a été lancé en Espagne en juin 2022, néanmoins avec peu de moyens techniques et de marges de manœuvre.

Un bac à sable réglementaire français, établi sur le modèle proposé par l'OCDE, pourrait être confié au futur régulateur français de l'IA, en association étroite avec les acteurs français chargés de l'essai et de l'expérimentation de systèmes d'IA (dénommés TEF, ou *Testing and Experimentation Facilities*, dans le AI Act européen), en particulier le LNE. Il s'agirait dans ce cadre non seulement d'un espace de conseil en vue de la mise en conformité d'un système d'IA avec la réglementation européenne, mais véritablement d'un environnement permettant de tester avec une conséquence juridique limitée le degré de conformité de nouveaux systèmes d'IA. Ce bac à sable pourrait s'appuyer sur des outils existants tels que le Laboratoire d'évaluation de l'intelligence artificielle (LEIA) du LNE.

Pour cela, le concept de "bac à sable" réglementaire intégré dans le AI Act doit impérativement soutenir une vision élargie et opérationnelle du bac à sable, qui ne se cantonne pas uniquement à un accompagnement en matière de conseil.

Recommandation 10 :

Confier au futur régulateur français de l'IA la création d'une base de données de référence de documentation des défaillances de systèmes d'IA.

La recherche et le développement de l'IA sûre et digne de confiance dépend intimement de notre connaissance des défaillances : quand et comment elles surviennent. Dans de nombreux autres domaines technologiques, le partage des rapports d'incidents contribue à une base commune de connaissances, aidant l'industrie et le gouvernement à suivre les risques et à comprendre leurs causes. Par exemple : la base de données européenne ECCAIRS des événements de sécurité d'aviation civile (ECR) relatif aux données d'occurrences et les Centres d'échange et d'analyse d'informations (ISACs), des plateformes public-privées développées aux États-Unis pour partager des incidents cyber par secteur.

La base de données européenne des événements de sécurité d'aviation civile

Dans l'aviation, toutes les personnes ayant une activité dans le domaine de l'aviation civile ont l'obligation de notifier les incidents compilés par type dans le règlement d'exécution (UE)2015/101. Dès lors qu'un événement interpelle mais n'est pas présent dans la liste, un "compte rendu volontaire" peut également être transmis à l'autorité et traité avec autant d'attention car le fait que des événements hors de la liste adviennent est en soi une donnée intéressante.

L'information doit obligatoirement être transmise à l'Autorité dans un format compatible avec le logiciel ECCAIRS (Pour European Coordination Centre for Accident and Incident Reporting Systems) et la taxonomie ADREP (pour Accident/Incident Data Reporting) élaborée par l'Organisation de l'aviation civile internationale (OACI).

Le guide de notification des incidents précise que l'analyse de l'incident constitue une étape indispensable au processus de traitement d'un événement. Cette dernière se compose d'une "description factuelle de l'événement rapporté et d'une interprétation des faits. En tout état de cause, elle devra être proportionnée au niveau de risque associé à l'événement pouvant aller d'une simple évaluation et à un classement sans suite à une analyse approfondie, dont les premiers éléments devront être transmis sous 30 jours à l'Autorité". Associé à ce document, il faut joindre les mesures correctives ou préventives qui ont été adoptées suite à l'événement rapporté.

En IA, certaines bases de données de défaillances de systèmes d'IA existent : le *Artificial Intelligence Incident Database* par exemple. Dans ce cadre, il conviendrait d'explorer différents moyens d'encourager les entreprises à partager les détails des accidents d'IA. Par exemple, en mettant en place des protections de confidentialité pour les informations commerciales sensibles, en développant des normes communes pour les rapports d'incidents, ou en rendant obligatoire la divulgation de certains types d'incidents telle que le prévoit le AI Act.

La création et la gestion de cette base de données pourrait être confiée dans un premier temps au futur régulateur français de l'IA, en lien étroit avec le LNE et le CEA pour y intégrer les incidents observés dans le cadre des Installations d'essai et d'expérimentation (TEFs) et d'un éventuel "bac à sable réglementaire" (cf. recommandation 9). Elle pourrait ensuite être portée à l'échelle européenne soit par le Centre commun de recherche de la Commission européenne, à l'image de la base ECCAIRS pour l'aviation, soit dans le cadre des *data spaces* introduits dans le Data Governance Act (DGA) de l'Union européenne.

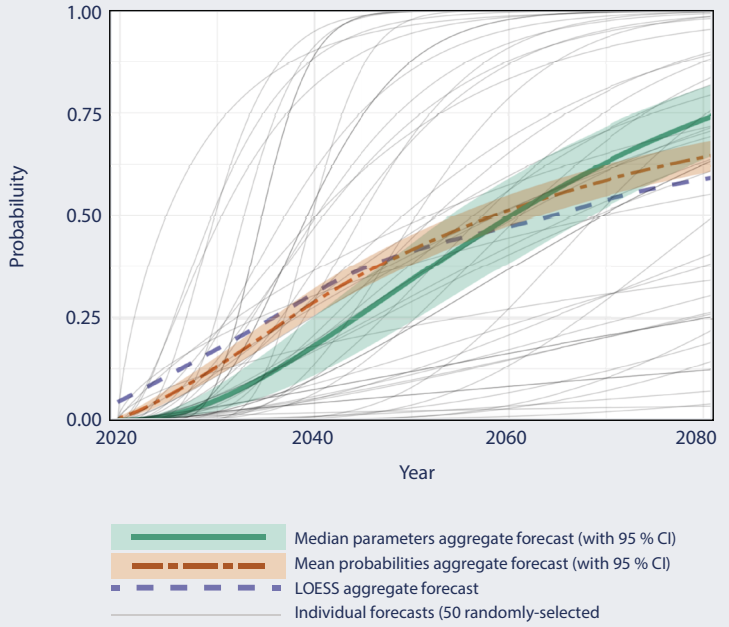
Annexe 1

Estimations du rythme de développement de l'IA : les meilleurs chercheurs en IA donnent 50 % de chance de développer des systèmes d'IA de niveau humain d'ici à 2059. Jusqu'à présent, ils ont largement sous-évalué le rythme de développement.

Les meilleurs chercheurs en IA estiment qu'il y plus de 50 % de chance de développer d'ici à 2059 des systèmes d'IA capables d'effectuer la quasi-totalité des tâches (> 90 %) mieux que l'humain médian. Et par le passé, les avancées de l'IA ont largement devancé les pronostics des experts du secteur.

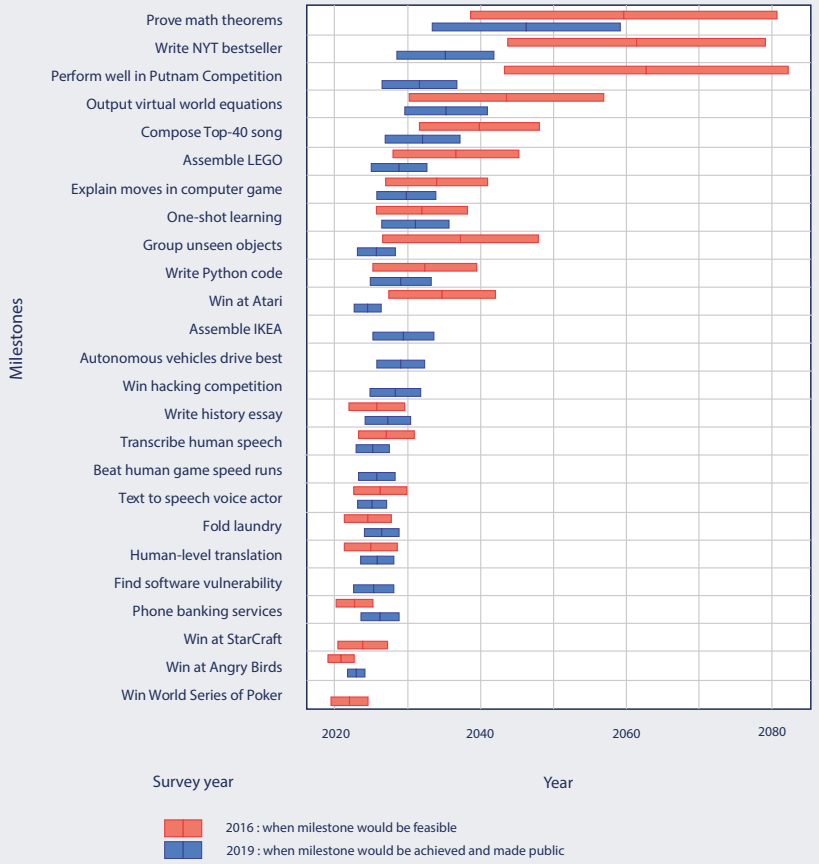
Trois enquêtes ont été menées en 2016, 2019 et 2022 auprès de chercheurs en IA ayant publié dans l'une des deux grandes conférences du domaine - la Conférence sur les systèmes de traitement de l'information neuronale (NeurIPS) et la Conférence internationale sur l'apprentissage automatique (ICML). L'estimation agrégée des 738 personnes qui ont répondu à l'enquête de 2022 était que nous avions plus de 50 % de chance de développer d'ici à 2059 des systèmes d'IA capables d'effectuer la quasi-totalité des tâches (> 90 %) mieux que l'humain médian (contre 2060 et 2061 dans les enquêtes de 2019 et 2016 respectivement). Des nettes différences d'opinions sont toutefois à noter.

Encore plus alarmant : au sein de ces enquêtes, lorsque la question est posée de la probabilité que les progrès futurs de l'IA soient extrêmement néfastes (au point de poser un risque d'impact permanent et grave sur l'espèce humaine, voire d'extinction), la médiane des réponses des chercheurs se place à 5 %.



Source : Enquête de 2019

Par le passé, les avancées de l'IA ont largement devancé les pronostics des experts du secteur. L'enquête réalisée en 2016 permet de comparer les dates estimées par les chercheurs pour certaines avancées majeures avec les dates de leur premières réalisations. Toutes les avancées réalisées à date sont parvenues plus tôt que les estimations des chercheurs : gagner au Poker, et aux jeux d'Atari notamment, mais on pourrait également citer les jeux de Go ou de Starcraft.



Annexe 2

État des lieux de la régulation de l'IA dans le monde

De nombreuses organisations nationales et internationales ont proposé des grands principes pour l'IA éthique et de confiance, néanmoins sans préciser les attributs spécifiques que ceux-ci impliquent et sans que ces principes ne soient juridiquement contraignants. Les principaux sont ceux de l'UNESCO et de l'OCDE. Ces derniers ont ensuite servi de base à l'adoption des *G20 AI Principles* en juin 2019, et au lancement du Partenariat mondial sur l'intelligence artificielle (PMIA) en 2020, porté par la France et le Canada au sein du G7.

- **À date, seule l'Union européenne développe des cadres réglementaires et juridiques qui ont vocation à s'appliquer à l'intégralité des systèmes d'IA**, avec ses projets de "AI Act" et de directives en matière de responsabilité civile pour l'IA.

- **En Chine, la réglementation sur l'IA avance rapidement, centrée pour le moment sur les algorithmes de recommandation.** En mars 2022, le pays a adopté le "Règlement sur la gestion de la recommandation d'algorithmes pour les services d'information sur Internet" proposé par l'Administration chinoise du cyberespace (ACC), imposant aux entreprises utilisant des algorithmes de recommandation d'en informer leurs utilisateurs et leur donner la possibilité de ne plus être ciblés. Et en Janvier 2023, l'ACC régulait la technologie de "synthèse profonde", c'est-à-dire les modèles d'IA capables de générer du texte, des images, de l'audio, de la vidéo. En septembre 2021, le ministère de la Science et des Technologies chinois avait déjà publié ses "Normes éthiques pour l'intelligence artificielle de nouvelle génération", proposant des principes éthiques s'inscrivant dans le démarches des principes de l'UNESCO et de l'OCDE.

- Aux États-Unis**, la réglementation douce (*soft law*) a été privilégiée pour le moment**, portée par l'Institut national des normes et de la tech-

nologie (*National Institute of Standards and Technology - NIST*), et aucun travail législatif n'a été entamé. Néanmoins la Maison Blanche a récemment publié son projet (non contraignant) de Charte des droits de l'IA (*Blueprint for an AI Bill of Rights*). Celui-ci exige une plus grande transparence sur la façon dont les algorithmes sont créés, une plus grande responsabilité dans la prise de décision basée sur l'IA, et la possibilité pour les utilisateurs de porter plainte en cas de défaillance.

Annexe 3

État des lieux de la normalisation de l'IA dans le monde

En Europe, les organisations européennes de normalisation reconnues CEN et CENELEC ont d'ores et déjà établi le Comité Technique Conjoint CEN-CENELEC 21 "Intelligence Artificielle" pour accompagner le développement et l'adoption de normes européennes pour l'IA[16] . Dans ce cadre, elles poursuivent un travail particulièrement pionnier pour définir une liste d'attributs de l'IA sûre et digne de confiance et préciser leur spécification technique d'ici à 2025.

1. **Du fait de l'extraterritorialité de l'AI Act, les normes harmonisées européennes auront une portée importante.** L'AI Act, qui n'est pas encore finalisé, prévoit que le respect des normes harmonisées proposées par CEN-CENELEC soit un moyen pour les fournisseurs de démontrer la conformité de leurs systèmes d'IA aux exigences du règlement.
2. À l'international, le Comité Technique Conjoint CEN-CLC/JTC 21 conseille activement d'autres initiatives de normalisation de l'IA, notamment ISO/IEC. Grâce à l'expertise croissante du CTC CEN-CLC/JTC 21, l'Europe dispose d'une opportunité pour porter sa vision de l'IA sûre et digne de

[16] Les grands axes de la stratégie de normalisation européenne ont été esquissés dans la réponse du CEN-CENELEC au livre blanc de la Commission européenne sur l'IA, dans leur feuille de route sur l'IA et dans la feuille de route de la normalisation allemande pour l'intelligence artificielle.

confiance au niveau international. Par ailleurs des accords de coopération technique existent spécifiquement entre l'ISO d'une part, et le CEN et le CENELEC d'autre part, pour coordonner leur travaux de standardisation et éviter de dupliquer les efforts. La Commission se plie néanmoins au principe de primauté des normes internationales : l'Europe adopte les normes internationales lorsqu'elles existent ou sont en construction, et la priorité est ainsi donnée aux travaux de l'ISO.

Participation de la France aux groupes de travail CEN-CENELEC

Au niveau international, l'Organisation internationale de normalisation (ISO) et la Commission électrotechnique internationale (IEC) travaillent ensemble pour développer des normes internationales pour l'IA dans le cadre du sous-comité SC 42, Artificial intelligence de la JTC1.

1. Ce travail est bien entamé, avec plus d'une quinzaine de normes qui ont déjà été publiées et d'autres encore en construction.
2. Néanmoins aucun modèle complet de l'IA sûr et de confiance n'a été proposé : à date l'ISO/IEC n'ont pas défini de liste exhaustive des attributs de l'IA sûre et digne de confiance, ni de guide permettant aux entre-

prises de choisir et de prioriser les normes qui sont pertinentes pour leur produit.
3. Le secrétariat du ISO/IEC JTC 1 - SC 42 est assuré par l'Institut des normes nationales américaines (ANSI) et travaille en lien avec 35 organisations de normalisation nationales, dont l'AFNOR pour la France.

Aux États-Unis, le Congrès a mandaté[17] l'Institut national des normes et de la technologie (NIST), qui relève du Département du Commerce américain, pour développer un *framework* volontaire de gestion du risque pour des systèmes d'IA digne de confiance. Une première version du *framework* a été publiée en janvier 2023. Le NIST et le CEN-CENELEC travaillent pour aligner leurs normes, y compris via les échanges permis par le E.U.-U.S. *Trade and Technology Council* (TTC). La *Federal Trade Commission* (FTC) a par ailleurs récemment présenté sa feuille de route d'exigences en matière de conformité des systèmes d'IA, centrée notamment sur les risques de biais dans l'octroi de crédits.

En Chine, le PCC mène une stratégie de standardisation lancée en 2018 avec le "China standards 2035" et précisée dans une feuille de route en octobre 2021 puis en juillet 2022.

1. "China standards 2035" cite explicitement l'IA comme domaine clé pour la standardisation.
2. Dès août 2020, l'organisme de normalisation du gouvernement de la RPC avait publié ses lignes directrices pour la construction d'un système national de normes d'intelligence artificielle de nouvelle génération. Celui-ci fixe comme objectif de mettre en œuvre dès 2023 un système de normes pour l'IA ainsi qu'une plateforme d'essai et de vérification des normes d'IA. Le sous-comité de l'intelligence artificielle du Comité technique national de normalisation des technologies de l'information chinois (SAC/TC 28/SC 42) se charge de ces travaux de normalisation, avec une première proposition de normes publiée en juillet 2021.

[17] *via le National Defense Authorization Act de 2021.*

3. Par ailleurs, le *China Academy of Information and Communications Technology* (CAICT), un *think tank* influent relevant du ministère de la Science et des Technologies chinois, avance rapidement avec une approche centrée sur le développement d'outils pour mesurer et tester la robustesse, la fiabilité et la contrôlabilité de systèmes d'IA. Le livre blanc sur l'IA digne de confiance publié par le CAICT en juillet 2021 définit des principes qui se rapprochent de ceux proposés par l'UE et les US pour l'IA digne de confiance. Le CAICT travaille avec l'Alliance chinoise de l'industrie de l'IA, un organisme industriel parrainé par le gouvernement chinois, pour tester et certifier des systèmes d'IA. En novembre 2021, elle a délivré sa première série de certifications d'IA digne de confiance pour les systèmes de reconnaissance faciale. Notons néanmoins que le ministère de la Science et des Technologies chinois n'a pas encore lui-même publié de documents sur l'IA sûre et digne de confiance, ce qui laisse encore des doutes quant au poids politique de ces initiatives à court terme.

Annexe 4

État des lieux de la recherche en IA sûre et digne de confiance dans le monde

En France

Le Programme National de Recherche en IA (PNRIA), le volet recherche de la stratégie nationale pour l'IA prévu pour la période 2018-2022, a permis de structurer les efforts de recherche français en IA autour d'un réseau de quatre instituts interdisciplinaires d'IA (3IA) et de près de 190 chaires, dont plusieurs dizaines qui s'intéressent à des enjeux pertinents pour le développement de l'IA sûre et digne de confiance, en allant de la compréhension du comportement des modèles d'IA existants, souvent qualifiés de "boîtes noires", à des enjeux plus applicatifs de certificabilité.

En particulier, l'un des instituts 3IA, le *Artificial and natural intelligence Toulouse Institute* (ANITI), s'intéresse spécifiquement à l'IA digne de confiance et s'articule autour de trois programmes de recherche sur l'IA acceptable, l'IA

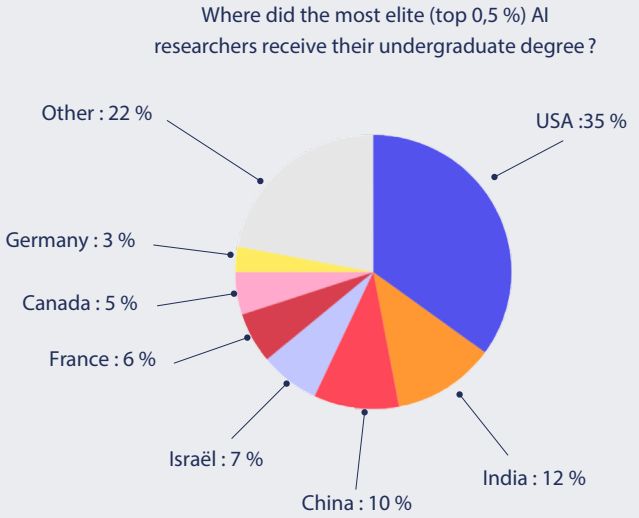
certifiable et l'IA collaborative. Les priorités de recherche de ces instituts ont à date été structurées par cas d'usages applicatifs (santé, environnement, transports, etc.), plutôt que par technologies ou infrastructures transverses.

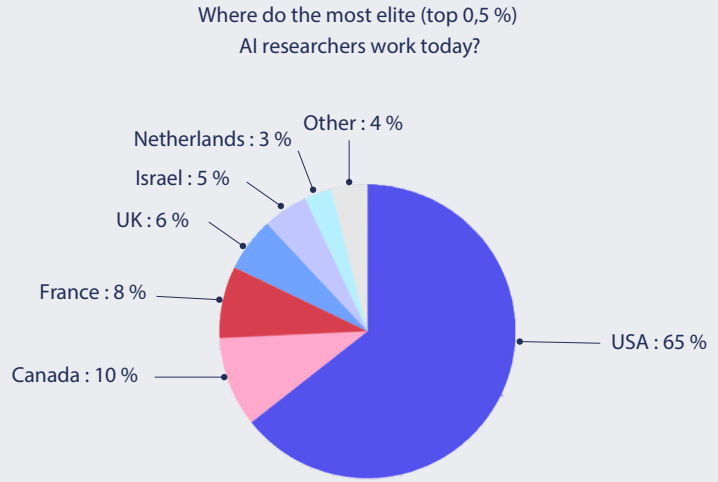
La France dispose également d'acteurs capables de devenir des *leaders* de la R&D en IA sûre et digne de confiance : des industriels experts des systèmes critiques dans différents secteurs (aéronautique, défense, transport et automobile, santé, ainsi que dans l'assurance et les services financiers), en pointe de l'innovation en IA et qui sont les premiers concernés par l'IA sûre et digne de confiance.

Plus généralement, la France dispose d'atouts importants permettant de créer ou d'attirer les meilleurs talents et laboratoires de recherche en IA :

1. un excellent système de formation d'ingénieurs et de chercheurs en IA, qui pourrait très simplement former à l'IA sûre et digne de confiance (en 2019, la France formait 6 % des meilleurs chercheurs en IA) ;
2. un excellent écosystème de la recherche en IA (en 2019, la France hébergeait 8 % des meilleurs chercheurs en IA), qui a déjà attiré de nombreux laboratoires de recherche de pointe (Google DeepMind et Facebook AI Research (FAIR) par exemple) ;
3. des avantages fiscaux généreux pour les entreprises innovantes : le Crédit Impôt Recherche permet une réduction d'impôt de 30% jusqu'à 100 M d'euros de R&D et 5 % au-delà.

Où les meilleurs chercheurs en IA ont été formés ; et où ils travaillent aujourd'hui (Macro Polo)





En Europe

La recherche en IA se centre autour d'un réseau de centres d'excellence, composé d'instituts de recherche européens. Celui-ci développe 4 projets de recherche, dont deux sur l'IA sûre et digne de confiance : le projet TAILOR sur les fondations de l'IA digne de confiance, et le projet HumaneAI-Net sur des systèmes d'IA robustes et capables de comprendre des humains.

- Si chacun rassemble plus de 50 partenaires, ils manquent vraisemblablement de moyens. Financés à hauteur de 12 millions d'euros chacun sur 4 ans par le programme de financement Horizon 2020, ces efforts se poursuivent sous Horizon Europe, qui finance la recherche européenne sur la période 2021-2027.
- Par ailleurs, le plan de coordination sur l'IA de la Commission européenne prévoit également un partenariat public-privé IA, données et robotique, qui doit assurer la souveraineté européenne dans le développement et le déploiement de l'IA, des données et de la robotique dignes de confiance, sûres et robustes. 2 600 milliards d'euros y seront consacrés d'ici 2030, dont 1 300 milliards de financement de la Commission européenne.

L'Amérique du Nord, en particulier les États-Unis, sont également en pointe sur la majorité des sujets de recherche en IA sûre et digne de confiance, disposant d'un avantage important de recherche en IA plus généralement. Les États-Unis se distinguent par une communauté de chercheurs en "*AI safety*" oeuvrant sur les problématiques de spécification et d'alignement, qui touchent davantage les systèmes d'apprentissage par renforcement et la recherche en IA avancée, voire générale.

- En 2021, l'Amérique du Nord représentait 75 % des publications à la conférence FACCT (*fairness, accountability and transparency*), contre 17 % pour l'Europe et l'Asie Centrale et moins de 5 % pour l'Asie Pacifique[18] .

[8] *2022 AI Index*

- L'enjeu de la robustesse retient l'intérêt des acteurs de la défense et du renseignement.

1. La DARPA, l'agence de recherche appliquée du ministère de la défense, connue pour avoir été à l'origine d'innovations pionnières comme le GPS, internet, et des développements importants en intelligence artificielle, mène un projet sur larobustesse via l'IA l'hybride (*Assured Neuro Symbolic Learning and Reasoning* - ANSR) [19] , et un autre sur les attaques délibérées contre des systèmes d'IA(*Guaranteeing AI Robustness against Deception* - GARD).
2. L'IARPA, l'équivalent de la DARPA pour les acteurs du renseignement, mène également deux programmes sur la sécurité de l'IA : *Secure, Assured, Intelligent Learning Systems* (SAILS) et *Trojans in Artificial Intelligence* (TrojAI).

- L'interprétabilité et l'explicabilité des systèmes d'IA constituent également un domaine de recherche important pour ces acteurs.

1. La DARPA mène un projet sur l'explicabilité : *Explainable Artificial Intelligence* - XAI.
2. La *National Science Foundation* américaine (NSF) et Amazon collaborent sur l'équité de l'IA, avec des sujets de recherche qui incluent la transparence, l'explicabilité, la responsabilité, les biais, l'équité, et l'inclusivité.
3. L'entreprise de recherche en IA sûre et explicable Anthropic, cofondée en 2021 par l'ancien vice-président de la recherche d'OpenAI, a attiré plus de $700M d'investissement en moins d'un an.

- Enfin, certains des meilleurs chercheurs et laboratoires d'IA se consacrent aux problématiques de spécification et d'alignement, notamment de systèmes d'IA par renforcement et en anticipant des systèmes d'IA de plus en plus généraux. On peut citer en particulier :

[19] *US National AI R&D Strategic Plan 2019 update*

1. le *Center for Human-Compatible Artificial Intelligence* de l'université de Berkeley, dirigé par Stuart Russell, l'un des pionniers de l'intelligence artificielle ;
2. les acteurs privés américains à la pointe de la recherche en intelligence artificielle : Google DeepMind et OpenAI, qui ont des équipes entières consacrées à l'IA sûre, notamment sur les problématiques de spécification et d'alignement.

Plus récemment, la National Science Foundation Américaine annonçait un financement à hauteur de $20M d'un programme de recherche en sûreté de l'IA ("*Safe Learning-Enabled Systems program*").

En Chine, si l'on a traditionnellement prêté au modèle chinois une insouciance vis-à-vis de la sûreté et de la confiance, il ne faut pas sous-estimer les ambitions du pays dans ce domaine et sa vision de long-terme.

- **Plusieurs acteurs académiques et privés investissent le sujet** depuis la conférence de Xiangshan de novembre 2017, au cours duquel le chercheur He Jifeng (何积丰), a introduit le concept d'IA digne de confiance. Parmi les entreprises chinoises, JD, Tencent, et Megvii ont toutes développé des initiatives d'IA digne de confiance. Dès janvier 2020 Megvii avait établi son Artificial Intelligence Governance Research Institute, et en avril 2020 l'institut de recherche de JD a confirmé que l'IA digne de confiance devenait l'un de ses principaux axes de recherche.

- **La Chine est par ailleurs sensible à l'enjeu d'une stratégie de recherche intégrée sur l'IA digne de confiance, en intégrant les perspectives d'IA générale.** Dans son livre blanc sur l'IA digne de confiance, le CAICT met en avant deux pistes pour poursuivre le développement de la recherche chinoise en IA digne de confiance :

1. développer un agenda de recherche "intégré" sur l'IA digne de confiance, pour éviter de travailler en silo et permettre aux différents projets de recherche en l'IA digne de confiance de communiquer entre eux et de partager un cadre commun ;

2. prévoir une feuille de route anticipant l'émergence d'une intelligence artificielle générale (AGI), et élargir la recherche en IA digne de confiance à la recherche en IA forte.
3. Plus généralement, l'approche de plus en plus prudente et conservatrice du gouvernement chinois vis-à-vis la technologie pourrait également encourager les efforts en IA sûre et digne de confiance (cf. le "Règlement sur la gestion de la recommandation d'algorithmes pour les services d'information sur Internet" adopté par la Chine en mars 2022).

- Enfin, le gouvernement chinois saisit parfaitement l'importance d'avoir des acteurs de pointe en matière de R&D pour imposer son modèle de normes. L'entreprise chinoise Huawei en est l'exemple.

Annexe 5

L'évaluation de la conformité des systèmes d'IA prévu par le AI Act

Le AI Act impose notamment aux systèmes d'IA "à risque élevé" des obligations d'évaluation de la conformité. Celles-ci dépendent des normes harmonisées développées par le CEN-CENELEC, qui poseront non seulement un cadre de conformité pour les systèmes d'IA, mais aussi pour les compétences et des processus d'audit de l'IA.

Les systèmes d'IA qui créent un risque faible ou minimal pourront eux aussi se plier à un code de conduite que l'AI Act ne définit pas spécifiquement. Pour ces systèmes d'IA, il est donc important de définir le label qui structurera ce code de conduite au niveau européen, voire international, et qui devra s'inspirer des normes techniques pour être légitime.

Selon la procédure d'évaluation de la conformité des systèmes d'IA prévue par le AI Act, les fournisseurs de systèmes d'IA à haut risque doivent d'abord suivre **une procédure d'évaluation de la conformité, puis un système de surveillance après la commercialisation.**

Dans un premier cas, l'évaluation de la conformité peut être externe et réalisée par un "organisme notifié". Cette démarche concerne :

1. Les systèmes d'IA à risque élevé utilisés comme composants de sécurité des produits de consommation qui font déjà l'objet d'évaluations de conformité ex ante par des tiers
2. L'identification biométrique des personnes à distance en temps réel et a posteriori qui n'appliquent pas les normes harmonisées ou les spécifications communes.

Après l'évaluation de la conformité, l'organisme notifié délivre un certificat d'évaluation de la documentation technique de l'UE ; le fournisseur rédige ensuite une déclaration de conformité de l'UE, appose le "CE" sur le produit, puis rédige un formulaire de déclaration de l'UE.

Un organisme notifié est un organisme d'évaluation de la conformité désigné par l'autorité notifiante du pays en question. Concernant les équipements de protection individuelle (EPI) par exemple, en France c'est la Direction Générale du Travail (DGT) qui joue le rôle d'autorité notifiante. **Concernant l'AI Act, il est possible que la CNIL ait le rôle d'autorité notifiante. Les organismes de certification (startups spécialisées, sociétés d'audit, et autres) devront ainsi être notifiés par cette autorité pour être autorisés à mener ce travail.**

Dans un second cas, l'évaluation de la conformité peut être interne et réalisée par le fournisseur lui-même. Le fournisseur peut travailler avec des tiers tels que des sociétés d'audit. Ceci concerne les systèmes d'IA à haut risque autonomes (c'est-à-dire qui ne sont pas concernés par les cas 1/ ou 2/ ci-dessus). Dans ce cas, le prestataire doit se conformer soit à des normes harmonisées (si elles existent) ; soit à des spécifications communes. Après l'évaluation de la conformité, le fournisseur rédige ensuite une déclaration de conformité de l'UE, puis appose le "CE" sur le produit, puis rédige un formulaire de déclaration de l'UE.

Un système de surveillance après la commercialisation est mis en place lorsque le système d'IA est sur le marché, afin d'évaluer la conformité continue. Pour les systèmes d'IA ayant reçu une évaluation de la conformité externe, l'organisme notifié effectue également des audits périodiques.

Remerciements

L'auteur de cette note remercie l'ensemble de l'équipe de l'Institut Montaigne ayant permis sa réalisation, notamment Tom David, assistant chargé de projets, et Camille Le Mitouard, chargée de projets, ainsi que l'ensemble des personnes auditionnées ou consultées dans l'élaboration de ce travail :

- **Jamal Atif**, Professeur et Vice- président en charge du Numérique, Université Paris Dauphine-PSL
- **Guillaume Avrin**, Coordonnateur national pour l'intelligence artificielle
- **Francis Bach**, Chercheur , Inria - École Normale Supérieure (ENS)
- **Pierre-Etienne Bardin**, *Chief Data Officer*, La Poste
- **Annabelle Blangero**, *Data Scientist - Senior Manager*, Ekimetrics
- **Anne Bouverot**, Présidente du Conseil d'administration de Technicolor et Présidente de la Fondation Abeona
- **Raja Chatila**, Professeur émérite, Sorbonne Université
- **Julien Chiaroni**, Ancien directeur Grand Défi en Intelligence Artificielle, Secrétariat général pour l'investissement (SGPI)
- **Rémy Choquet**, Directeur du centre de données médicales, Roche
- **Marie-Pierre de Bailliencourt**, Directrice générale, Institut Montaigne
- **Caroline de Condé**, Responsable du pôle Normes et Confiance Numérique et du projet Grand Défi Normalisation en Intelligence Artificielle, Groupe AFNOR
- **Marcin Detyniecki**, *Head of Research and Development & Group Chief Data Scientist*, AXA
- **Marko Erman**, SVP, *Chief Scientific Officer*, Thales
- **Laurent Inard**, Associé et Chief R&D Officer, Mearzars
- **Caroline Jeanmaire**, Doctorante *Artificial Intelligence Policy*, *Blavatnik School of Government* à l' Université d'Oxford
- **Elliot Jones**, Chercheur, Ada Lovelace Institute
- **Fabien Le Voyer**, Coordinateur national adjoint pour l'intelligence artificielle
- **Emmanuelle Legrand**, Chargée de mission IA (régulation), Direction générale des entreprises (DGE)

- **Bruno Maisonnier**, Président et fondateur, *Another Brain*
- **Nicolas Marescaux**, Directeur adjoint Réponses Besoins Sociétaires et Innovation, MACIF
- **Sébastien Meunier**, Vice président relations institutionnelles, ABB France
- **Nicolas Miailhe**, Co-fondateur et Président, *The Future Society*
- **Nicolas Moës**, Directeur Europe Gouvernance de l'IA, *The Future Society*
- **Louis Morilhat**, Chargé de mission IA, Groupe AFNOR
- **Aurélien Palix**, Sous directeur des réseaux et usages numériques, Direction générale des entreprises (DGE)
- **Ludovic Peran**, *Product Manager for Responsible & Human-centered AI*, *Google Research*
- **Tanya Perelmuter**, Co-fondatrice et directrice de la stratégie et des partenariats, Fondation Abeona
- **Gabriel Peyré**, Professeur, CNRS, DMA, Ecole Normale Supérieure (ENS)
- **Hadrien Pouget**, Chercheur Invité, *Carnegie Endowment for International Peace*
- **Timothée Raymond**, Directeur de l'innovation et de la technologie, Linedata
- **Benoit Rottembourg**, Responsable REGALIA Audit Algorithmique & Régulation, Inria
- **Gérard Roucairol**, Président Honoraire, Académie des Technologies de France
- **Stuart Russell**, Professeur d'informatique, Université de Californie à Berkeley
- **Isabelle Ryl**, Directrice, PRAIRIE (*PaRis Artificial Intelligence Research InstitutE*) - Inria
- **Guillaume Sylvestre**, Directeur Innovation Numérique, ADIT
- **Helen Toner**, Directrice de la stratégie, *Center for Security and Emerging Technology* (CSET)
- **Fabrice Valentin**, VP Intelligence Artificielle, Airbus

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

L'Institut Montaigne vous propose de contribuer à la réflexion sur ces enjeux afin d'élaborer collégialement des propositions au service de l'intérêt général.

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

Imprimé en France
Dépôt légal : avril 2023
ISSN : 1771-6756

Les adhérents

INSTITUT Montaigne

ABB France
Abbvie
Accenture
Accuracy
Adeo
ADIT
Aéma
Air France - KLM
Air Liquide
Airbus
Allen & Overy
Allianz
Amazon
Amber Capital
Amundi
Antidox
Antin Infrastructure Partners
Archery Strategy Consulting
Archimed
Ardian
Arquus
Astrazeneca
August Debouzy
Avril
AXA
Baker & Mckenzie
Bearingpoint
Bessé
BG Group
BNP Paribas
Bolloré
Bona Fidé
Bouygues
Brousse Vergez
Brunswick
Capgemini
Capital Group
CAREIT
Carrefour
Casino
Chubb
CIS
Cisco Systems France
Clifford Chance
Club Top 20
CMA CGM
CNP Assurances
Cohen Amir-aslani
Compagnie Plastic Omnium
Conseil supérieur du notariat
Crédit Agricole
D'angelin &Co.Ltd
Dassault Systèmes
De Pardieu Brocas Maffei
DIOT SIACI
Doctolib
ECL Group
Edenred
EDF
EDHEC Business School
Egis
Ekimetrics France
Enedis
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
HSBC Continental Europe
IBM France
IFPASS
Inkarn
Institut Mérieux
International SOS
Interparfums
Intuitive Surgical
Ionis Éducation Group
iQo
ISRP
Jeantet Associés
Jolt Capital
Kantar Public
Katalyse
Kearney
Kedge Business School
KKR
KPMG S.A.
La Banque Postale
La Compagnie Fruitière
Linedata Services
Lloyds Europe
L'Oréal
Loxam
LVMH - Moët-Hennessy - Louis Vuitton
M.Charraire
MACSF
MAIF
Malakoff Humanis
Mazars
Média-Participations
Mediobanca
Mercer
Meridiam
Michelin
MicroPort CRM
Microsoft France
Mitsubishi France S.A.S
Moelis & Company
Moody's France
Morgan Stanley
Natixis
Natural Grass
Nestlé
OCIRP
ODDO BHF
Oliver Wyman
Ondra Partners
onepoint
Onet
Optigestion
Orano
Ortec Group
PAI Partners
Pelham Media
Pergamon
Prodware
PwC France & Maghreb
Raise
RATP
RELX Group
Renault
Rexel
Ricol Lasteyrie
Rivolier
Roche
Rokos Capital Management
Roland Berger
Rothschild & Co
RTE
Safran
Sanofi
SAP France
Schneider Electric
Servier
SGS
SIER Constructeur
SNCF
SNCF Réseau
SNEF
Sodexo
SPVIE
SUEZ
Taste
Tecnet Participations SARL
Teneo
The Boston Consulting Group
Tilder
Tofane
TotalEnergies
UBS France
Unibail-Rodamco
Veolia
Verlingue
VINCI
Vivendi
Wakam
Wavestone
Wendel
White & Case
Willis Towers Watson France
Zurich

L'intelligence artificielle (IA), et avec elle notre société, est à un tournant historique. Nous développons désormais des systèmes d'IA "à usage général" comme ChatGPT, capables d'effectuer un nombre de tâches toujours plus grand. Ils pourraient ainsi rapidement constituer un facteur de compétitivité décisif pour les entreprises comme pour les pays.

Ces systèmes représentent néanmoins un enjeu de sécurité majeur et croissant. Non seulement parce qu'ils peuvent être utilisés par des acteurs malveillants, mais aussi parce que la nature statistique des systèmes d'IA actuels pose un risque de sûreté et de défaillance inédit, qui constitue désormais l'un des freins technologiques les plus importants du domaine.
Cet enjeu est également une opportunité unique pour la France de se positionner en *leader* d'une IA sûre et digne de confiance, en attirant notamment les meilleurs talents en la matière, qui estiment que la sûreté est une préoccupation majeure mais insuffisamment traitée par leurs employeurs. Elle dispose de chercheurs de rang mondial en mathématiques et en IA et d'une expertise de pointe en ingénierie système et logicielle pour la sûreté. Grâce aux puissants ordinateurs du Centre national de la recherche scientifique (CNRS), elle est par ailleurs l'un des seuls pays européens en mesure de développer des grands modèles d'IA à usage général.

Pour saisir cette opportunité, la France doit s'en donner les moyens, avec un projet d'innovation de rupture et un pôle de recherche fondamentale dédiés au développement de systèmes d'IA à usage général sûrs et dignes de confiance. Elle doit également s'assurer que les systèmes d'IA performants mais dangereux, développés aujourd'hui par les acteurs américains et chinois, soient soumis à la future réglementation européenne, qui a de grandes chances de définir les exigences internationales en matière de sûreté et de confiance de l'IA.

10 €
ISSN : 1771-6756
NAC2304-01