

Pour une IA tournée vers l'avenir

Renforcer le rapport de la Commission IA en adoptant une approche d'anticipation proactive des développements technologiques futurs



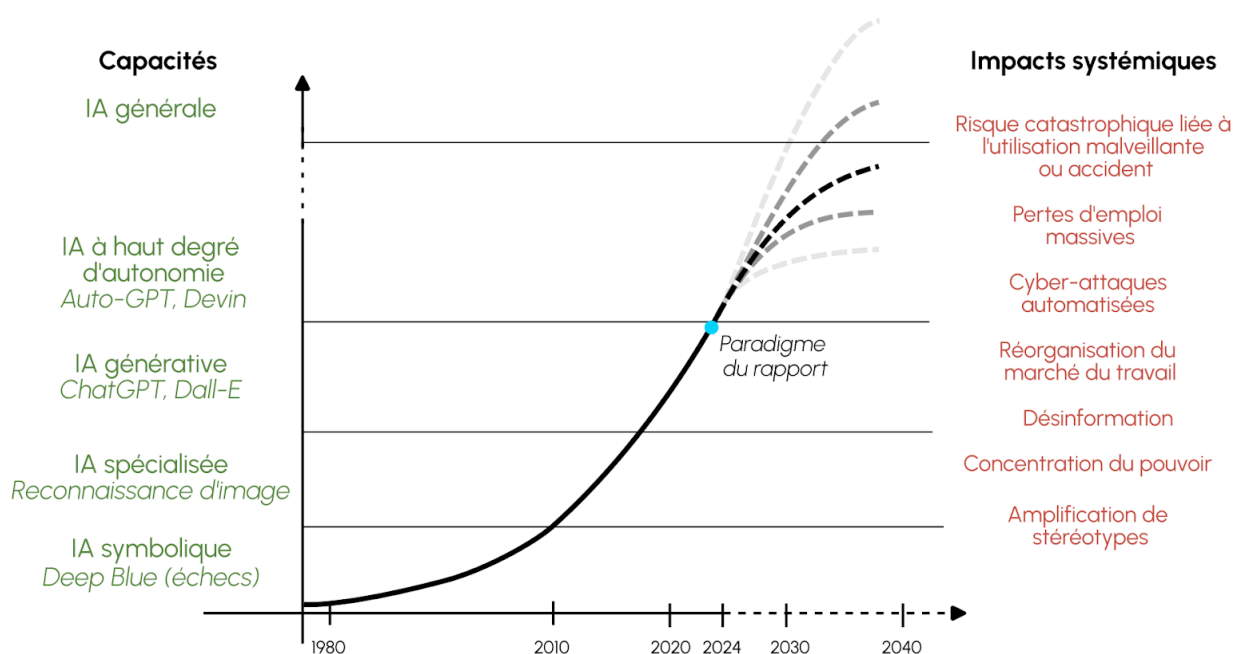
Introduction

Le 13 mars, la Commission de l'intelligence artificielle dirigée par Anne Bouverot et Philippe Aghion a rendu son premier rapport intitulé "[l'IA : notre ambition pour la France](#)" au Président. Nous applaudissons la Commission pour ce rapport détaillé, qui souligne avec justesse l'immense potentiel de l'IA et la nécessité d'investir dans l'innovation responsable, en accord avec nos valeurs humanistes, afin que la France et l'Europe ne soient pas laissées à l'écart.

Cependant, la rapidité des progrès en IA rend nécessaire l'adoption d'une approche proactive pour anticiper les développements futurs et leurs impacts systémiques, tel qu'illustré dans la figure ci-dessous. Sans cette perspective, nous prenons le risque de passer à côté de certains aspects technologiques, économiques et sociétaux essentiels, limitant notre capacité à préparer pleinement la France aux défis et opportunités à venir.

Nous proposons ici des recommandations concrètes pour renforcer ce rapport, notamment en initiant un travail prospectif permettant d'élaborer des scénarios technologiques à court et moyen terme, en œuvrant pour faire de la France un acteur clé dans l'évaluation de l'IA et en investissant dans la recherche en IA spécialisée.

Ne reproduisons pas avec l'IA l'erreur commise avec le problème climatique : ne laissons pas la perspective de gains immédiats occulter l'importance d'anticiper les défis futurs, même incertains.



L'avenir de l'IA est incertain; il est essentiel d'explorer l'étendue des ruptures technologiques possibles afin d'anticiper les impacts systémiques.

L'IA, moteur d'une transformation sociétale profonde

Nous saluons le travail remarquable des auteurs du rapport, avec lesquels nous partageons de nombreux points de consensus quant à l'impact transformateur de l'IA. L'arrivée des grands modèles de langage (LLMs) marque un véritable changement de paradigme, de par leur capacité à généraliser leur application à une vaste gamme de tâches.

Les bénéfices économiques potentiels de l'IA sont immenses. Comme le souligne le rapport, cette technologie a le potentiel de stimuler la croissance, améliorer la productivité, et nous aider à relever les défis majeurs de notre époque, de la [fusion nucléaire](#) à la [prédiction du climat](#), en passant par les voitures autonomes et l'aide au diagnostic médical. Sans être une solution miracle, elle peut néanmoins contribuer à résoudre des problèmes complexes, en complément d'approches sociétales et d'une réflexion sur nos modes de vie. Nous soutenons la proposition de fonds international pour l'IA au service de l'intérêt général présentée par le rapport.

Une gouvernance internationale de l'IA est la clé pour harmoniser et orienter son développement. Nous soutenons la mise en place des standards et audits indépendants harmonisés à l'échelle mondiale. Nous saluons ainsi la proposition d'Organisation mondiale de l'IA, et nous attendons avec intérêt les discussions à propos de sa création lors du sommet de l'IA à Paris.

Investissons dans une innovation européenne ambitieuse et responsable. Pour ne pas passer à côté de cette révolution technologique, il est primordial d'investir massivement dans les maillons clés de la chaîne de valeur où l'Europe peut faire la différence, ainsi que dans la recherche publique.

L'éducation doit être au cœur de notre stratégie IA, pour former les talents de demain et promouvoir une IA éthique et robuste et préparer la société aux nouveaux usages de l'IA. Nous soutenons la proposition de lancement immédiat d'un plan de sensibilisation et de formation.

L'IA évolue rapidement, créant des incertitudes quant à son impact futur sur la société. Nous partageons l'approche d'humilité et d'adaptabilité du rapport tout en soulignant l'importance d'anticiper les évolutions pour orienter les investissements et les politiques publiques. Nous soutenons la mission confiée à l'Organisation mondiale de l'IA visant à faire un état des connaissances sur l'évolution de l'IA et de ses impacts, à la manière du GIEC.

Ces points de consensus soulignent notre engagement pour un progrès technologique humaniste. Cependant, il nous semble essentiel de renforcer le rapport, en ajoutant une réflexion prospective sur les ruptures technologiques à venir. Apprenons des limitations du GIEC, qui a mis en lumière les risques climatiques sans toujours inciter à l'action immédiate. Adoptons une posture proactive face à l'IA, avec une anticipation dynamique pour naviguer les défis et saisir les opportunités à venir.

Anticiper les incertitudes pour mieux façonner l'avenir

Le progrès en IA évolue à un rythme effréné, comme le soulignent les auteurs eux-mêmes :

"Dans les mois et années à venir, nous devrions connaître de nouvelles avancées rapides et de grande ampleur. Les modèles seront progressivement capables d'être factuels, de mener des raisonnements, de comprendre le monde physique"

Dans cette direction, nous proposons d'ajouter à ce constat une analyse des dynamiques de l'IA qui accélèrent son développement :

- **Amélioration par passage à l'échelle** : les ["scaling laws"](#), confirmées à répétition, décrivent comment les performances brutes des modèles s'améliorent de manière prévisible avec l'augmentation de leur taille. Ainsi, l'émergence de nouvelles capacités d'IA à usage général, tel que ChatGPT, relève maintenant plus d'ingénierie à grande échelle que de percées scientifiques.
- **Explosion des investissements** : les investissements dans l'IA connaissent une croissance exponentielle, avec un [doublement de la puissance de calcul pour les plus grands modèles tous les neuf mois](#). Chaque nouveau modèle de frontière révèle des capacités auparavant inédites et imprévisibles.
- **Course à l'innovation et tragédie des communs** : la compétitivité du marché pousse les entreprises à développer des modèles toujours plus performants que leurs rivaux, parfois au détriment de la sécurité, comme l'a démontré [l'incident avec Bing Chat de Microsoft](#).

Le rapport se concentre principalement sur le paradigme actuel de l'IA générative, ce qui est compréhensible dans le contexte présent. Cependant, il semble essentiel de tenir compte de l'apparition récente des systèmes hautement autonomes, faute de quoi toute réflexion risque de devenir obsolète à court et moyen terme. Ces agents capables d'exécuter de longues séquences d'actions sans intervention humaine annoncent un nouveau changement de paradigme. Nous aspirons à ce que les développements de la technologie dans les cinq années à venir ne prennent pas au dépourvu les décideurs politiques, mais fassent au contraire partie des scénarios envisagés.

Aussi nous soulignons ci-dessous les aspects prospectifs les plus importants à anticiper:

1. **Les systèmes à haut niveau d'autonomie**. Contrairement aux IA actuelles qui sont de simples outils répondant à des requêtes spécifiques, les agents autonomes, en tirant parti de la puissance des modèles de langage précédents, sont capables de décomposer une tâche en une longue série d'actions permettant d'atteindre un objectif qui leur a été assigné, cela sans intervention humaine. Un riche écosystème d'applications se développe dans cette direction, tels que [Devin](#), [AutoGPT](#), [GPT-pilot](#), [Magic.dev](#) et d'autres, soutenus par des financements importants. Ces avancées rapides, comme en témoigne le [doublement des performances sur le benchmark General AI Assistant](#) en quelques mois, ouvrent la voie à la réalisation de tâches, comme la contribution à de grands projets logiciels, auparavant inaccessibles aux

assistants classiques. De nouveaux projets, à l'instar des [Large Action Models de Rabbit](#) et [SIMA](#) de DeepMind, exploitent les données d'actions humaines pour entraîner des modèles à les reproduire, marquant une rupture avec les approches basées uniquement sur le traitement du texte. Bien que ces technologies ne soient pas encore matures, initier une réflexion sur leurs impacts potentiels, notamment sur l'accélération de l'innovation, l'automatisation des tâches et les risques de cyberattaques, est essentiel.

- 2. Futur du travail.** L'analyse empirique présentée dans le rapport se concentre principalement sur les tâches automatisables par les systèmes d'IA actuels. Il reconnaît aussi qu'il est nécessaire de se préparer à une IA plus puissante que nous dans de nombreux domaines. Si à moyen terme (5 ans) la grande majorité des tâches deviennent automatisables, objectif affiché des laboratoires d'IA sur cette échéance, quelles en seront les implications ? Il semble critique d'évaluer la faisabilité de ces objectifs ainsi que leur temporalité. D'autre part, il serait utile d'initier dès maintenant un dialogue pour ébaucher de nouveaux modèles socio-économiques adaptés à de tels futurs.
- 3. Risques émergents.** Le rapport évoque des risques systémiques tels que "comportement émergent critique" et "accident systémique" dans un diagramme p. 32, mais ne développe ni leur nature, ni l'horizon temporel de leur réalisation. Ces risques, tels que ceux "liés aux modèles capables de se répliquer ou d'entraîner d'autres modèles", mentionnés dans le [AI Act européen](#), sont fréquemment mentionnés par les créateurs de ces modèles¹. Des réflexions sur ces risques et les mesures d'atténuation possibles existent, comme dans le récent article [Regulating advanced artificial agents](#) publié dans Science. Une discussion éclairée, ni utopique, ni catastrophiste, sur l'importance et la nature de ces risques doit faire partie du dialogue autour de l'évolution de l'IA. Alors que les systèmes d'IA sont sur une trajectoire qui les mènera à devenir bien plus que de simples outils, plus autonomes et plus capable que les humains dans de nombreux domaines, il est aussi nécessaire de compléter la description des risques liés à une utilisation malveillante par les problèmes de contrôle liés à l'alignement des systèmes d'IA avec les intentions humaines.

Recommandations

Le rapport de la Commission de l'intelligence artificielle représente un effort remarquable pour aborder les défis et les opportunités liés à l'IA. Nous souhaitons formuler quatre recommandations afin de compléter ce travail pour une stratégie nationale ambitieuse et tournée vers l'avenir.

- 1. Lancer un travail prospectif sur les avancées technologiques envisageables en IA à un horizon de 5 ans.** Sans prétendre à une prédiction exacte, ce projet qui pourrait prendre la forme d'un rapport aurait pour objectif d'explorer un large éventail de scénarios plausibles et d'en évaluer les implications potentielles, à la manière du

¹ Comme mentionné dans la [déclaration sur les risques de l'IA](#) du Center for AI Safety signée par dirigeants d'OpenAI, Anthropic et DeepMind entre autres.

rapport [Risques futurs des modèles frontières](#) du gouvernement britannique ou du groupe de travail [AI Futures](#) à l'OCDE. Il devra :

- a. Accorder une attention particulière à l'analyse des risques émergents, en s'appuyant sur les travaux existants et en approfondissant les aspects les plus critiques pour la société et l'économie.
 - b. Intégrer une dimension géopolitique en prenant en compte les stratégies et les capacités des principaux acteurs du développement de l'IA, qu'il s'agisse d'entreprises, de centres de recherche ou de puissances étatiques.
2. **Faire de la France un acteur clé de l'évaluation et la robustesse des modèles.** Le rapport prévoit seulement 0.05% du budget d'investissement total pour créer un écosystème d'évaluation de l'IA public-privé, insuffisant face à l'importance de l'évaluation des modèles, aux défis des risques émergents, et au potentiel économique du secteur. La France doit établir un cadre durable pour comprendre l'impact technologique, avec un investissement en recherche publique pour adresser les incertitudes liées aux risques émergents. Cette recherche s'inscrit en collaboration avec le secteur privé pour des évaluations détaillées des applications d'IA, exploitant un secteur économique prometteur. L'innovation en robustesse et supervision des modèles, où la France excelle déjà grâce à des initiatives comme Giskard, est essentielle pour maximiser les avantages de l'IA et positionner la France comme une référence en la matière. Comme le souligne le rapport, innovons pour assurer notre futur.
3. **Orienter les efforts de recherche et de développement vers les technologies d'IA spécialisées et à fort impact positif**, notamment celles permettant la création de nouveaux médicaments, la création de nouveaux matériaux, la prévision météorologique, le diagnostic médical, les véhicules autonomes ou encore le contrôle de la fusion nucléaire. Alors que l'attention se porte actuellement sur les outils d'IA à usage général, la France a une opportunité de se positionner comme un leader mondial dans le développement de systèmes spécialisés, peu risqués, attirant talents et investissements internationaux.
4. **Investir dans la recherche et le développement de contre-mesures techniques face aux risques liés à l'IA**, tels que la défense contre des agents autonomes malveillants avec des systèmes de supervision en temps réel, la cyberdéfense à l'image de l'[AI Cyber Defense Initiative de Google](#) ou la lutte contre la diffusion de deep fakes et la désinformation à grande échelle. Face à l'émergence de ces nouvelles menaces, il est impératif de se doter de moyens de défense efficaces et innovants, potentiellement utilisant eux-mêmes l'IA. C'est un domaine actuellement sous-estimé dans lequel la France peut se positionner en leader mondial.

À propos

Contact : hello@securite-ia.fr

Le [Centre pour la Sécurité de l'IA](#) (CeSIA - prononcé "Césia") est une jeune organisation française à but non lucratif dédiée à l'éducation, la recherche, et la diffusion d'informations sur les enjeux de l'intelligence artificielle.

L'intelligence artificielle se développe à une vitesse fulgurante, et offre d'énormes possibilités tout en soulevant des défis majeurs en matière de sécurité (jailbreaks, hallucinations, manque de transparence, enjeux de cybersécurité, risques systémiques). Devant une préparation mondiale qui semble insuffisante face à ces évolutions, il nous semble crucial de favoriser l'émergence et l'engagement d'une variété d'acteurs, chacun apportant ses propres compétences et spécificités.

Pour cette raison, le CeSIA vise à soutenir les acteurs clés de l'IA pour les préparer aux défis à venir, contribuant ainsi à l'établissement d'un écosystème français de l'IA solide et collaboratif. Notre cœur de métier consiste à apporter un éclairage technique sur les tendances de développement de l'IA, afin d'identifier les risques et enjeux actuels et d'anticiper ceux à venir.

Nos activités se concentreront sur trois axes : la recherche, qui inclut le développement d'outils open source pour superviser l'opération des modèles d'IA et détecter des modes de défaillance encore inconnus ; l'information du grand public, pour élever la conscience collective sur les usages et enjeux de l'IA ; et la formation des futures générations de chercheurs et de développeurs, centrée sur l'évaluation et la sécurisation des systèmes d'IA.