IA : NOS CRAINTES POUR LA FRANCE

Contre-expertise du Rapport de la Commission de l'IA

PAUSE IA

Résumé Exécutif

La Commission de l'intelligence artificielle [1], créée par le gouvernement français et présidée par Anne Bouverot et Philippe Aghion, a pour mission de positionner la France en leader face aux enjeux de l'IA pour les années à venir. Son rapport final [2], publié le 13 mars 2024 et intitulé « IA : Notre ambition pour la France», s'avère être une occasion manquée d'engager une véritable réflexion sur les enjeux et les risques du développement et du déploiement de l'intelligence artificielle. La question revêt pourtant une importance croissante pour l'avenir technologique et économique de la France. Notre contre-expertise, soutenue par 14 experts et 5 organisations, a pour objectif de corriger cette trajectoire, en mettant en lumière les nombreuses lacunes et biais méthodologiques du rapport de la Commission.

Le rapport de la Commission incite à faire vivre un débat public. Dans cette perspective, notre analyse se concentre délibérément sur ses aspects les plus problématiques. Cette approche critique vise à combler les manques que nous avons identifiés, sans pour autant nier l'existence d'éléments pertinents dans le rapport original.

Au cœur de notre critique se trouve le non-respect des standards scientifiques par le rapport, qui écarte sans justification l'avis d'une grande partie des experts en IA. Face à une situation où de nombreux chercheurs éminents, dont une majorité des experts en sécurité de l'IA, alertent sur des risques actuels et potentiellement catastrophiques à court et moyen terme, **la Commission fait le choix de les ignorer**. Cette approche compromet la capacité de la France à anticiper et gérer les défis majeurs posés par l'IA, mettant ainsi en péril la sécurité nationale et l'avenir de notre société.

Notre analyse révèle des défaillances systématiques dans le rapport, que nous pouvons résumer en sept critiques principales :

1. Silence injustifié sur les avertissements des experts en IA et sur la littérature scientifique existante.
2. Aucune anticipation des développements futurs de l'IA, même à court terme.
3. Négligence de l'impact des IA de recommandation sur le déclin démocratique globalisé et du rôle qu'elles ont joué dans plusieurs massacres de masse [3], [4], [5].
4. Omission des considérations principales sur la sécurité de l'IA.

5. Sélection des données et des faits pour soutenir une vision excessivement optimiste.
6. Minimisation systématique des risques identifiés, notamment en matière d'emploi et de cybersécurité.
7. Rigueur insuffisante par comparaison aux rapports similaires d'autres pays.

Ces défaillances sont illustrées à de nombreuses reprises, dont voici quelques exemples :

- **Emploi :** Le rapport ne retient que les données impliquant un impact positif de l'IA. Dans un premier temps, il reconnaît que sa première approche n'est pas concluante, mais affirme quelques phrases plus loin qu'elle montre un « effet positif sur l'emploi ». Ensuite, il cite une étude prévoyant un risque de remplacement de 30 % du travail par l'IA, mais l'ignore immédiatement, pour conclure sans justification à un effet positif.
- **Cybersécurité :** Alors que GPT-4 surpasse déjà 88 % des pirates humains dans certaines compétitions, le rapport disqualifie ce risque en deux phrases, sans aucune référence.
- **Risques existentiels :** Le rapport tourne en dérision les avertissements sur les risques existentiels de l'IA sans aucune justification scientifique, tout en démontrant une méconnaissance du sujet en allant jusqu'à confondre la déclaration du Center for AI Safety [6] signée par des centaines d'experts de premier plan avec la lettre ouverte [7] d'une autre institution.
- **Création artistique :** Sans citer aucune source ni référence, le rapport affirme que « L'IA ne met pas en danger l'originalité de la création ». Cette déclaration ignore les rapports alarmants qui émergent sur l'impact de l'IA dans le domaine artistique, et occulte les transformations profondes qu'elle impose dès aujourd'hui à ce secteur.
- **Absence d'anticipation :** Le rapport évalue systématiquement les risques en se basant uniquement sur les capacités passées des IA, supposant implicitement un arrêt soudain du progrès technologique. Cette approche amplifie tous les problèmes précédents. Les risques déjà sous-estimés pour l'emploi, la cybersécurité et la création artistique sont décuplés si l'on considère l'évolution rapide et continue des capacités des IA.

Ces manquements s'expliquent en grande partie par la composition même de la Commission, marquée par **des conflits d'intérêts majeurs** et un **manque de diversité d'opinions et d'expertises.** La Commission ne compte aucun expert

en sécurité de l'IA et est dominée par des représentants de l'industrie favorables à, et favorisés par, un développement accéléré et peu régulé de l'IA.

Face à cette situation inquiétante, nous appelons à :

1. Un remaniement de la Commission pour éliminer les conflits d'intérêts et garantir une diversité d'expertise.
2. La consultation urgente d'experts en sécurité de l'IA.
3. À la suite de cette consultation, la rédaction d'un addendum traitant des risques ignorés.
4. L'ouverture d'un débat public éclairé sur l'avenir de l'IA en France.

L'accueil par la France du Sommet pour l'action sur l'IA début 2025 offre l'occasion de prendre une position de leadership sur les enjeux de sécurité de l'IA. **Ces enjeux ne représentent rien de moins que l'avenir de notre société et, potentiellement, de l'humanité toute entière.** Il est impératif que la France prenne la mesure réelle des défis et des dangers que pose cette technologie, et agisse en conséquence.

Cette contre-expertise se veut une réponse concrète à l'appel au débat public mentionné dans le rapport de la Commission (p. 7), initiant ainsi un dialogue critique et constructif sur l'avenir de l'IA.

Table des matières

1 Introduction

L'intelligence artificielle (IA) transforme rapidement notre société, avec des implications profondes pour l'avenir [8], [9]. Les décisions prises aujourd'hui en matière de développement et de régulation de l'IA auront des impacts concrets sur l'économie, la sécurité, la vie sociale et politique, l'influence internationale de la France, ainsi que sur son industrie, sa recherche scientifique, l'emploi et la vie privée des citoyens. Il est donc essentiel d'examiner attentivement les bases sur lesquelles ces décisions sont prises.

Prenant acte de l'importance de ces impacts, le gouvernement français crée en septembre 2023 le Comité — depuis devenu Commission [10] — de l'intelligence artificielle. La Commission est sommée, sous six mois, de produire un rapport censé guider les politiques nationales en matière d'IA pour les années à venir [11]. Ce rapport est publié le 13 mars 2024 et intitulé *IA : Notre ambition pour la France* [12].

Le présent document propose une analyse critique approfondie du rapport de la Commission, s'inscrivant dans la lignée d'autres critiques notables, notamment celles exprimées par l'UGICT-CGT [13] et le Centre pour la Sécurité de l'IA [14]. Nous mettons en évidence des lacunes majeures et des conflits d'intérêts qui, s'ils restaient ignorés, pourraient avoir des implications considérables pour l'avenir de la France et de l'humanité. **Notre objectif principal est de créer les conditions d'un débat sain et équilibré autour de l'IA, qui ne soit plus capturé par des lobbys puissants et des intérêts privés et s'appuie sur les connaissances scientifiques les plus récentes en IA.**

Notre analyse se divise en trois parties principales :

1. Analyse critique du rapport de la Commission
2. Évaluation de la composition et du fonctionnement de la Commission
3. Recommandations pour adresser les lacunes identifiées

Dans la première partie, nous examinerons en détail :

a. Les omissions critiques du rapport
b. La minimisation et la distorsion des risques
c. La négligence des préoccupations citoyennes
d. La différence de qualité d'analyse comparée à d'autres rapports nationaux et internationaux

La seconde partie se concentrera sur l'analyse de la composition de la Commission, mettant en évidence les potentiels conflits d'intérêts et leurs implications sur les conclusions du rapport.

Enfin, nous proposerons des recommandations concrètes visant à combler les lacunes identifiées et à améliorer le processus d'élaboration des politiques en matière d'IA.

Cette contre-expertise se veut une contribution critique et constructive, et une réponse concrète à l’appel au débat public national mentionné dans le rapport de la Commission (p. 7). En mettant en lumière les faiblesses du rapport de la Commission et en proposant des pistes d'amélioration, nous cherchons à promouvoir une approche plus rigoureuse et transparente des enjeux de l'IA, dans l'intérêt à court, moyen et long terme de la société française.

2 Méthodologie et approche

Cette contre-expertise a été réalisée entièrement par un groupe de citoyens bénévoles possédant diverses expertises, sous la direction de Maxime Fournes, expert en Intelligence Artificielle. Notre approche a été principalement qualitative, axée sur une analyse critique approfondie du rapport de la Commission de l'IA.

Engagement bénévole :

- Ce travail a été effectué intégralement par des bénévoles sur leur temps libre (voir section « A propos des auteurs »).
- L'effort collectif représente environ 400 heures de travail, réparties sur plusieurs mois.
- Nous avons visé une date de publication suffisamment en amont du sommet de l'IA, afin de catalyser une réflexion et un débat éclairés.

Objectifs :

- Évaluer la rigueur scientifique et l'exhaustivité du rapport de la Commission.
- Identifier les lacunes et les biais potentiels.
- Proposer des recommandations pour une approche plus complète et équilibrée.

Sources et collecte de données :

- Analyse détaillée du rapport de la Commission de l'IA.
- Revue de la littérature scientifique pertinente sur la sécurité de l'IA.
- Consultation de chercheurs en sécurité de l'IA.

Processus d'analyse :

1. Lecture critique du rapport de la Commission.
2. Identification des points clés et des lacunes potentielles.
3. Comparaison avec la littérature scientifique existante.
4. Consultation d'experts pour valider nos observations.
5. Synthèse des résultats et formulation de recommandations.

Validation :

À l'issue de la rédaction initiale, le document a été soumis à 14 experts indépendants pour relecture et validation. **Le processus de validation s'est concentré sur la partie analytique du rapport, excluant les recommandations qui n'engagent que les auteurs principaux**.

Niveaux de participation :

1. **Validation :** L'expert confirme l'exactitude et la pertinence de l'analyse présentée. Cette validation peut s'appliquer à l'ensemble du document ou à des sections spécifiques relevant de l'expertise de l'expert.
2. **Soutien :** L'expert approuve les principales conclusions de l'analyse, tout en exprimant des réserves sur certains points spécifiques. Le soutien peut concerner l'ensemble du document ou des sections spécifiques.

Les experts ont eu la possibilité de choisir les sections qu'ils souhaitaient examiner, en fonction de leur domaine d'expertise. Chaque expert a eu l'opportunité d'ajouter un bref commentaire personnel pour expliciter sa position ou ses éventuelles réserves.

Limitation :

- Cette contre-expertise a été réalisée de manière indépendante, sans accès aux données brutes ou aux délibérations de la Commission de l'IA.

3 Analyse critique du rapport

3.1 Vue d'ensemble

Le rapport de la Commission de l'IA [15] présente des lacunes profondes et multiples qui compromettent gravement sa crédibilité et sa pertinence en tant que document d'orientation stratégique pour la France. Notre analyse révèle une approche manquant de rigueur scientifique et ignorant certains risques majeurs liés au développement rapide de l'IA.

Premièrement, le rapport souffre d'omissions critiques. Il passe sous silence le domaine scientifique de la sécurité de l'IA, établi depuis plus de quinze ans [16]. Il ignore les avertissements de très nombreux experts de premier plan concernant les risques existentiels, y compris ceux émis par des figures emblématiques comme Stuart Russell [17], Yoshua Bengio [18], [19] et Geoffrey Hinton [20]. De plus, le rapport fait preuve d'un manque important d'anticipation, se concentrant principalement sur les impacts des technologies actuelles sans considérer sérieusement les défis potentiels posés par les innovations à venir. Enfin, le rapport néglige largement les risques liés aux technologies d'IA déjà déployées, comme les algorithmes de recommandation, pourtant déjà omniprésents et influents. Cette omission est particulièrement préoccupante étant donné l'impact profond de ces systèmes sur la formation de l'opinion publique, le fonctionnement de nos démocraties et la santé mentale des utilisateurs.

Deuxièmement, le rapport minimise systématiquement les risques qu'il aborde. Cette minimisation se manifeste à travers une déformation du contexte historique et actuel de l'IA, créant une fausse impression de continuité et de maîtrise technique. Le traitement des risques spécifiques, notamment en matière d'emploi et de cybersécurité, est particulièrement biaisé. Le rapport emploie des formulations orientées ou des arguments rhétoriques pour discréditer les scénarios les plus alarmants sans en adresser la substance, et façonne ainsi une perception des enjeux rassurante mais trompeuse.

Troisièmement, l'analyse comparative avec d'autres initiatives similaires met en lumière les déficiences méthodologiques du rapport. La sélection des experts, la formulation des conclusions, et l'incapacité à explorer divers scénarios crédibles témoignent d'une approche peu rigoureuse et potentiellement orientée.

Les lacunes identifiées ne sont pas de simples erreurs. Elles reflètent une défaillance systématique dans l'évaluation des risques et des opportunités liés à l'IA. Il est impératif que la France prenne la mesure réelle des défis posés par l'IA et adopte une approche rigoureuse, éthique et responsable pour guider son développement futur.

3.2 Omissions Critiques

Le rapport de la Commission est marqué par des omissions récurrentes qui compromettent sa crédibilité et son utilité. Ces lacunes ne semblent pas être de simples oublis, mais plutôt s'inscrire dans une démarche cohérente visant à minimiser les dangers et les défis posés par l'IA.

3.2.1 Absence totale de mention de la sécurité de l'IA

Ce champ de recherche, établi il y a plus de quinze ans, se consacre spécifiquement à l'étude des risques liés au développement de l'IA, aux moyens de les atténuer, ainsi qu'aux questions fondamentales de compréhension et de contrôle des intelligences artificielles [16]. Des institutions renommées telles que le Centre for the Study of Existential Risk à Cambridge [21], le Center for AI Safety (CAIS) [22], et le Center for Human-Compatible AI (CHAI) à Berkeley [23] sont à l'avant-garde de ces recherches. De plus, les principaux laboratoires d'IA comme OpenAI [24], [25], Google DeepMind [26], [27], [28] et Anthropic [29], [30] ont constitué des équipes dédiées à la sécurité de l'IA, soulignant l'importance croissante de ce domaine. Cette discipline est désormais enseignée dans les universités à la pointe de l'intelligence artificielle à travers le monde, notamment à Stanford [31], au MIT [32], à Berkeley [33], et à l'ETH Zurich [34], ainsi que dans des écoles parmi les plus prestigieuses en France, telles que l'École Normale Supérieure [35] et l'École Polytechnique [36].

Parmi les accomplissements majeurs de ce domaine, on peut citer :

1. L'identification et l'évaluation d'un large éventail de risques liés aux systèmes d'IA actuels et futurs [37], [38], [39]. Ces analyses ont notamment mis en lumière :
    a. Les vulnérabilités fondamentales des systèmes d'apprentissage, particulièrement pour les modèles de très grande dimension [40], [41].

b. La susceptibilité à la manipulation par des données empoisonnées, un problème omniprésent notamment pour les IA de recommandation [42], [43].
c. La mémorisation non désirée de données sensibles ou protégées par des droits d'auteur [44], [45].
d. La vulnérabilité systématique au « *jailbreaking* », c'est-à-dire à des manières de prompter ses systèmes qui rendent leur comportement imprévisible, voire dangereux [46], [47], [48].
e. L'identification de problèmes fondamentaux comme le « *reward hacking* » [49], permettant de mieux comprendre les défis de l'alignement.
f. Les risques liés aux algorithmes de recommandation déjà déployés [50], [51], [52] (voir section 3.2.4).

2. Le développement de techniques d'alignement du comportement sur les valeurs humaines, notamment :

    a. La gouvernance collaborative algorithmique des IA [53], [54], [55].
    b. L'apprentissage par jugements comparatifs [56], [57].
    c. L'apprentissage par filtrage collaboratif [58], [59].
    d. L'apprentissage par renforcement à partir de feedback humain (*RLHF*) [60], [61], aujourd'hui un des mécanismes principaux permettant de rendre les grands modèles de langage utilisables et plus alignés avec les intentions humaines.

3. Des avancées significatives dans l'interprétabilité mécanistique, ouvrant la voie à une compréhension plus profonde du fonctionnement interne des réseaux de neurones artificiels [62], [63], [64].

Pour une vue d'ensemble plus complète sur la sécurité de l'IA, nous renvoyons le lecteur aux rapports du NIST [65] et de l'ANSSI [66], ainsi que le site web du Centre pour la Sécurité de l'IA [67].

Cependant, nous faisons les constats suivants :

1. Aucune mention n'est faite de ce domaine de recherche ni de ses conclusions.
2. Aucun des experts en sécurité de l'IA que nous avons consultés n'a été approché pour contribuer à ce rapport.

Ces omissions constituent une lacune stratégique significative. En ne tenant pas compte de connaissances essentielles pour l'évaluation des risques liés à l'IA, le rapport compromet la pertinence de ses propres recommandations.

Cette approche soulève des inquiétudes quant à l'exhaustivité de l'expertise mobilisée et suggère que la stratégie nationale en matière d'IA pourrait ne pas prendre en compte certains des risques les plus sérieux identifiés par la communauté scientifique spécialisée.

3.2.2 Omission des avertissements de la communauté scientifique sur les risques systémiques et catastrophiques liés à l'IA

Sections du rapport concernées : 1.2

Malgré le silence du rapport à leur sujet, ces préoccupations ne sont pas marginales. Au contraire, des études récentes montrent qu'une part significative des experts du domaine considère que les IA présentent un risque réel d'extinction pour l'humanité [68]. Cette position est soutenue par des centaines de figures emblématiques du domaine [6] telles que Yoshua Bengio et Geoffrey Hinton, co-récipiendaires du prestigieux Prix Turing pour leurs travaux sur les réseaux de neurones profonds [69], [70].

Les risques catastrophiques et existentiels

Les risques existentiels liés à l'IA sont des menaces potentielles qui pourraient conduire à l'extinction de l'humanité ou à des dommages catastrophiques à l'échelle mondiale. Ces risques comprennent la création involontaire de systèmes d'IA incontrôlables ou mal alignés avec les valeurs humaines, l'utilisation malveillante de l'IA pour concevoir des armes biologiques ou des cyberattaques dévastatrices, et l'escalade de conflits géopolitiques assistée par IA. Selon une enquête menée en 2022 par AI Impacts [66], 48 % des experts interrogés estiment qu'il y a au moins 10 % de chance d'un résultat extrêmement négatif lié au développement de l'IA. Plusieurs éminents chercheurs en intelligence artificielle comme Geoffrey Hinton ou Yoshua Bengio ont fait part de craintes similaires. Geoffrey Hinton a récemment estimé le risque d'une catastrophe induite par l'IA à plus de 50 % [69]. Ces estimations alarmantes soulignent l'importance cruciale de la recherche en sécurité de l'IA et de la mise en place de garde-fous robustes pour guider le développement responsable de cette technologie.

En mai 2023, le Center for AI Safety a publié une déclaration [6], signée par plus de 350 experts de premier plan, affirmant que « l'atténuation du risque

d'extinction par l'IA devrait être une priorité mondiale ». Cette déclaration a suivi une autre lettre ouverte, *Pause Giant AI Experiments* [7], qui a recueilli plus de 33 000 signatures, entre autres d'experts en intelligence artificielle venant du milieu académique comme industriel, démontrant l'ampleur de ces préoccupations au sein de la communauté scientifique et au-delà.

La section « 1.2 Faut-il avoir peur de l'IA ? » (p. 31) du rapport tourne en ridicule ces craintes sans aucune justification ou argument, et déforme les faits, confondant différentes lettres ouvertes et minimisant leur portée. Par exemple, le rapport fait référence à la déclaration du Center for AI Safety en affirmant qu'elle n'a recueilli que 60 signatures d'experts, tout en la confondant avec la lettre ouverte sur la pause des expérimentations géantes en IA (p. 33, note de bas de page numéro 9).

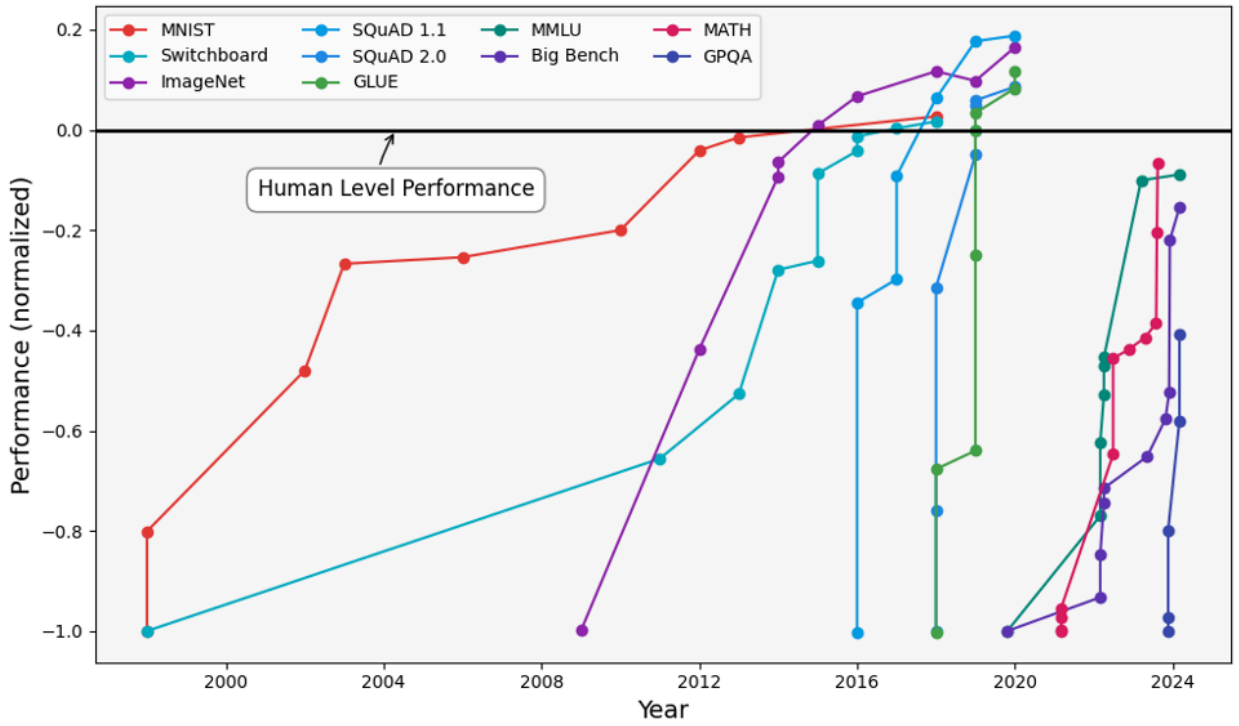
Il est certes juste de noter qu'il existe également des scientifiques qui considèrent ces inquiétudes comme exagérées. Cependant, face à une telle controverse impliquant des risques potentiellement catastrophiques, il est important d'étudier explicitement et rigoureusement les deux positions. Le rapport, en ignorant complètement l'un des côtés du débat, ne permet pas une évaluation équilibrée et approfondie des enjeux.

Cette représentation erronée et ce rejet des préoccupations légitimes de la communauté scientifique soulèvent une question fondamentale : comment un rapport censé guider la politique nationale en matière d'IA peut-il esquiver délibérément les avertissements sonores d'une grande partie des experts du domaine ? En écartant ces voix sans aucune justification valable, le rapport prive les décideurs et le public d'informations essentielles pour comprendre l'ampleur réelle des défis posés par l'IA avancée.

3.2.3 Aucune anticipation des développements de l'IA, même à court terme

Sections du rapport concernées : 1.4

Bien que la Commission reconnaisse que l'IA connaîtra « de nouvelles avancées rapides et de grande ampleur » (p. 12), son analyse se limite principalement aux impacts des technologies actuelles, négligeant ainsi les défis potentiels posés par les innovations à venir. Cette lacune, déjà pointée par le Centre pour la Sécurité de l'IA (CeSIA) dans une analyse détaillée [14], n'a reçu aucune réponse de la part de la Commission.



Performance des modèles d'IA sur divers benchmarks de 2000 à 2024, comprenant la vision par ordinateur (MNIST, ImageNet), la reconnaissance vocale (Switchboard), la compréhension du langage naturel (SQuAD 1.1, MNLU, GLUE), l'évaluation générale des modèles de langage (MMLU, Big Bench, et GPQA) et le raisonnement mathématique (MATH). De nombreux modèles dépassent le niveau de performance humaine (ligne noire solide). Kiela, D., Thrush, T., Ethayarajh, K., & Singh, A. (2023) « Plotting Progress in AI ».

L'IA évolue à une vitesse sans précédent [71], [72], et le rapport récent du UK AI Safety Institute [73] reconnaît la possibilité que cette progression « extrêmement rapide » (p. 9) se poursuive dans un avenir proche. Une étude menée en 2023 auprès de 1714 experts a cherché à déterminer quand des machines autonomes seraient capables d'accomplir toutes les tâches mieux et à moindre coût que des travailleurs humains. Les prévisions globales de cette étude estiment qu'il y a 50% de chances que cela se produise avant 2047 [74], soit 13 ans plus tôt que les prévisions d'une étude similaire réalisée en 2022 [75]. Ne pas adopter une approche prospective dans un tel contexte n'a aucun sens et compromet sérieusement la pertinence, même à court terme, des recommandations du rapport. Cette absence d'anticipation est particulièrement remarquable dans le cas du changement de paradigme vers des systèmes d'IA plus autonomes, un développement majeur que la Commission a omis de prendre en compte. Ces systèmes, capables d'exécuter de longues séries d'actions avec très peu de supervision humaine, soulèvent des questions d'ordre sociétal et des risques qualitativement différents des outils d'IA générative actuels [76]. Des projets comme AutoGPT [77], Devin [78], Genie [79] ou encore AI Scientist [80], qui étaient prévisibles depuis plus d'un an (c'est-à-dire bien avant la constitution même de la Commission), illustrent cette

transformation fondamentale. Pour plus d'informations à ce sujet, nous invitons à consulter la note du CeSIA [14].

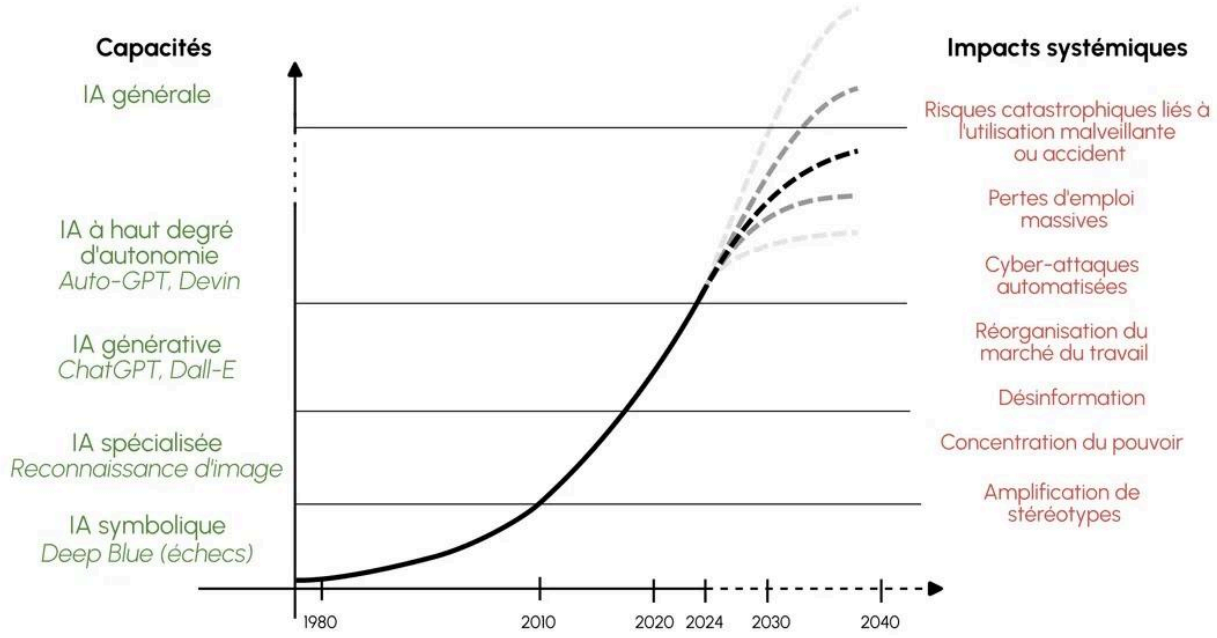


Figure illustrative extraite de l'article « Pour une IA française tournée vers l'avenir » du Centre pour la Sécurité de l'IA (CeSIA) [14]. Cette figure illustre le changement de paradigme actuel en IA, en comparant l'approche statique du rapport à l'évolution rapide des systèmes autonomes et leurs impacts systémiques potentiels, soulignant ainsi l'importance d'anticiper ces développements.

Notons que cette approche non prospective présente un intérêt stratégique : celui de minimiser les risques liés à l'IA par omission. En effet, tous les risques sont amplifiés par le développement des capacités des modèles. Si les avancées s'arrêtaient aujourd'hui, comme le suppose implicitement la Commission, alors les risques seraient effectivement bien moindres. Cependant, cette hypothèse est en contradiction avec la réalité du domaine [71], [72] et avec les propres déclarations de la Commission sur les « avancées rapides et de grande ampleur » (p. 12) à venir.

Cette absence d'anticipation se reflète particulièrement dans l'analyse des effets de l'IA sur l'emploi. La Commission cherche à anticiper ces effets en s'appuyant sur une enquête de l'INSEE [81] comparant des entreprises ayant adopté l'IA à celles ne l'ayant pas fait entre 2016 et 2021. Or, cette période précède l'explosion de popularité de l'IA générative, survenue en 2022 avec ChatGPT. Il est peu probable que l'adoption de l'IA générative par les entreprises engendre les mêmes conséquences que l'intégration des technologies d'IA dite « étroite » ces dernières années. De plus, l'impact de l'adoption de technologies d'IA encore plus avancées dans les années à venir pourrait être bien plus important.

Paradoxalement, bien que le rapport de la Commission souligne que les études sur l'impact de l'IA sur l'emploi « ne permettent pas encore de conclure à un effet sur un horizon de quelques années », elle affirme dans le même temps « un effet positif de l'IA sur l'emploi » (p. 44, 45). Cette conclusion hâtive, basée sur des données obsolètes et sans prise en compte des développements futurs, illustre clairement les dangers d'une approche non prospective dans un domaine en évolution aussi rapide que l'IA.

3.2.4 Négligence de l'impact dévastateur des IA de recommandation

Sections du rapport concernées : 1.7

Le rapport de la Commission néglige largement les risques posés par les IA de recommandation, omniprésentes dans notre paysage numérique actuel. Ces systèmes, qui déterminent quels contenus sont présentés aux utilisateurs sur les plateformes sociales, les moteurs de recherche et les sites de streaming, ont un impact profond sur la formation de l'opinion publique, le fonctionnement de nos démocraties et la santé mentale des utilisateurs.

Alors que le rapport se concentre principalement sur les IA génératives, il passe sous silence les dangers bien réels et immédiats des systèmes de recommandation. Cette omission est d'autant plus frappante que ces algorithmes sont déjà largement déployés et influencent quotidiennement plusieurs milliards d'utilisateurs. La section 1.7 du rapport, intitulée « L'IA peut-elle nuire à la qualité de l'information ? », effleure à peine la surface du problème, se concentrant principalement sur la création de fausses informations plutôt que sur la diffusion de propagande par des IA de recommandation.

Les impacts sociétaux de ces systèmes sont considérables et bien documentés :

- **Polarisation politique et radicalisation :** De nombreuses études [82], [83], [84] ont mis en évidence une corrélation entre l'usage intensif des réseaux sociaux et une augmentation de la polarisation politique. Les algorithmes de recommandation, en favorisant les contenus provocateurs et émotionnellement chargés, contribuent à créer des vagues de haine où les utilisateurs sont principalement exposés à des prises de position qui renforcent leur mépris pour les croyances opposées [84], [85], [86].

- **Amplification des discours haineux :** Comme l'illustre tragiquement le cas du génocide des Rohingyas en Birmanie [87] (voir section 4.2.1, encart « L'éthique contestable de Meta »), les systèmes de recommandation peuvent amplifier dangereusement les discours haineux, les appels à la violence et les incitations au meurtre [88], [89]. Le rapport ne mentionne ni ce risque pourtant avéré, ni ses conséquences potentiellement catastrophiques.
- **Désinformation et manipulation de l'opinion publique :** Les algorithmes de recommandation sont particulièrement vulnérables aux manipulations par des acteurs malveillants [90], [91]. Le rapport ignore complètement le rôle de ces systèmes dans la propagation de la désinformation, notamment lors d'événements critiques comme les élections [83]. Par exemple, lors des élections présidentielles américaines de 2016, une agence russe a pu obtenir une visibilité comparable à celle des candidats officiels en exploitant ces algorithmes, pour un coût mille fois moindre [92], [93], [94].
- **Impacts sur la santé mentale :** Le rapport ne mentionne pas les effets néfastes de l'exposition prolongée aux contenus recommandés sur la santé mentale des utilisateurs, particulièrement chez les jeunes [95]. Les « Facebook Files », révélés par la lanceuse d'alerte Frances Haugen (voir section 4.2.1, encart « L'éthique contestable de Meta »), ont mis en lumière que Meta (anciennement Facebook) était conscient des effets néfastes d'Instagram sur la santé mentale des adolescentes, sans pour autant prendre des mesures significatives pour y remédier [96].
- **Déclin démocratique :** Le V-Dem Institute et d'autres organismes indépendants ont observé un déclin global de la qualité des démocraties depuis l'avènement des réseaux sociaux [97]. Bien que la causalité directe soit difficile à établir, l'influence des algorithmes de recommandation sur ce phénomène mérite une attention sérieuse que le rapport ne lui accorde pas.
- **Cybersécurité et manipulation à grande échelle :** Le rapport sous-estime gravement la vulnérabilité des systèmes de recommandation aux manipulations par le cybercrime [98], [99]. L'utilisation de faux comptes et de bots [100] pour influencer ces algorithmes représente une menace sérieuse pour l'intégrité de l'information en ligne, un aspect essentiel de la sécurité nationale à l'ère numérique.

L'omission de ces risques dans le rapport est d'autant plus préoccupante que les algorithmes de recommandation sont déjà omniprésents et influents. « En 2017, YouTube estimait que, sur le milliard d'heures de vidéo que l'humanité consommait quotidiennement, 70 % était dû à des recommandations de son IA. 700 millions d'heures quotidiennement recommandées par l'IA de YouTube

correspondent à la durée totale d'enseignement que dispenseraient 25 000 professeurs au cours de leur carrière » [101]. De plus, depuis 2016, il y a plus de vues sur YouTube que de recherches sur Google, soulignant le rôle croissant de la recommandation par rapport à la recherche active d'information.

Le rapport néglige également les implications géopolitiques de ces technologies. L'utilisation différenciée des algorithmes de recommandation par certains pays, comme la Chine avec ses plateformes TikTok et Douyin, soulève des questions cruciales sur la manipulation de l'opinion publique à l'échelle internationale que le rapport ignore complètement.

Notons par ailleurs que les risques liés aux algorithmes de recommandation existants et ceux associés aux IA génératives ne sont pas mutuellement exclusifs, mais au contraire, se renforcent mutuellement. Les IA génératives offrent la possibilité de créer du contenu encore plus personnalisé pour chaque utilisateur, amplifiant ainsi le pouvoir d'influence des algorithmes de recommandation. Cette synergie potentielle entre les deux technologies pourrait exacerber tous les risques mentionnés précédemment, de la polarisation politique à la manipulation de l'opinion publique, en passant par les impacts sur la santé mentale.

En ne traitant pas adéquatement ces enjeux, le rapport de la Commission manque une opportunité d'alerter sur les risques immédiats et concrets posés par les algorithmes de recommandation, ainsi que sur leur potentielle amplification par les IA génératives. Cette lacune compromet la capacité de la France à élaborer une stratégie cohérente et complète face aux défis posés par l'IA dans toutes ses formes, actuelles et futures.

3.3 Minimisation et déformation des risques

Le rapport de la Commission de l'IA ne se contente pas d'omettre des informations essentielles ; il présente également une vision déformée et minimisée des risques qu'il traite. Cette section analyse en détail les différentes techniques employées pour sous-estimer l'ampleur et la gravité des défis posés par le développement rapide de l'IA. Nous examinerons d'abord comment le rapport déforme le contexte historique et actuel de l'IA, créant ainsi une fausse impression de continuité et de maîtrise. Ensuite, nous mettrons en lumière le traitement biaisé de risques spécifiques, notamment en matière d'emploi et de cybersécurité. Enfin, nous analyserons les techniques de rhétorique utilisées pour discréditer les scénarios les plus alarmants et façonner une perception rassurante mais trompeuse des enjeux.

3.3.1 Déformation historique et contextuelle

Sections du rapport concernées : Introduction p. 17

Le rapport de la Commission présente une perspective déformée de l'histoire et du contexte actuel de l'IA, ce qui conduit à une minimisation des risques et des défis posés par les développements récents.

Présentation orientée de l'histoire de l'IA

Le rapport décrit l'IA comme une technologie arrivée à maturité à l'issue d'une longue histoire, ce qui est trompeur. Cette perspective confond l'histoire du champ de recherche avec celle des technologies elles-mêmes. En réalité :

1. L'IA en tant que champ de recherche existe depuis les années 1950 [102].
2. Les paradigmes actuels, basés sur l'apprentissage profond et les grands modèles de langage (*LLM*), sont extrêmement récents — à peine plus d'une décennie pour le succès de l'apprentissage profond qui a provoqué un renouveau de la recherche en IA [103], et seulement quelques années pour les *LLM* [104].

> ***LLM***
>
> Les *Large Language Models* (*LLM*) sont des modèles d'intelligence artificielle entraînés à prédire le prochain élément dans un texte en fonction d'un contexte, générant ainsi du contenu cohérent. Ils utilisent des estimations probabilistes et sont capables de comprendre et de produire du langage humain, ce qui les rend particulièrement polyvalents pour diverses applications. Cependant, leur complexité et leur opacité posent des défis en termes de compréhension et de contrôle. Pour en savoir plus sur les *LLM*, consultez l'annexe D.

Cette présentation biaisée minimise le tournant constitué par les développements actuels. Les modèles d'IA contemporains représentent une rupture fondamentale avec les paradigmes précédents :

- Ils reposent sur des approches radicalement différentes (apprentissage profond vs approches symboliques).
- Ils manifestent des capacités émergentes, c'est-à-dire des aptitudes qui n'ont pas été explicitement programmées ou entraînées [105], [106].

- Leur vitesse d'évolution et d'amélioration est incomparable avec les systèmes précédents [72], [107].

Représentation trompeuse des modèles actuels

Le rapport sous-estime la complexité et l'imprévisibilité des modèles d'IA actuels, en particulier des *LLM*. Cette représentation trompeuse se manifeste de plusieurs façons :

1. Surestimation de notre compréhension :
    - Les *LLM* sont fondamentalement opaques [108]. Les algorithmes menant d'une entrée à un résultat sont invisibles, **même pour leurs créateurs**.
    - Un nouveau champ de recherche, l'interprétabilité [62], [63], a dû être créé pour tenter de comprendre le fonctionnement interne de ces modèles. L'interprétabilité n'en est qu'à ses balbutiements et les experts en interprétabilité mécanistique rappellent que leurs travaux encore exploratoires ne devraient pas être confondus avec une solution finie au problème de l'opacité [62].
2. Exagération de notre capacité de contrôle :
    - Les *LLM* ne sont pas « programmés » au sens traditionnel, mais plutôt obtenus à la suite d'un processus d'entraînement sur d'énormes quantités de données. Contrairement à un logiciel classique où chaque fonction est explicitement codée, les comportements d'un *LLM* émergent de manière complexe et imprévisible à partir de son apprentissage [105].
    - Ils développent des capacités émergentes souvent découvertes après leur mise à disposition au public [105].
    - Certains comportements potentiellement dangereux et indésirables ont été observés, comme le mensonge [109], la manipulation [110], et le piratage informatique [111], [112], [113], [114].
3. Minimisation de la complexité et de l'imprévisibilité :
    - Les capacités des *LLM* augmentent rapidement et de manière non linéaire [115], menant vers des risques d'augmentation brusque et incontrôlée de leurs capacités.
    - Les expériences du passé éclairent peu sur les capacités futures de ces systèmes. Il est, d'une part, difficile d'anticiper quelles

capacités émergentes vont survenir lors des prochaines générations de modèles. D'autre part, la continuité des *scaling laws [116]*, les lois prédisant l'amélioration des *LLM* en fonction de leur taille et de la quantité de calcul investie pour leur entraînement, fiable jusqu'à présent, doit nous faire anticiper une amélioration des capacités des modèles plutôt qu'une stagnation [117].
    - Les futures générations de *LLM* vont probablement déborder du paradigme initial en intégrant des méthodes de résolution de problème et de raisonnement développées dans d'autres branches de l'IA [118]. La fusion de méthodes fait croître la complexité. Elle réduit considérablement la prévisibilité.

Le rapport de la Commission présente les IA actuelles comme de simples « outils » utilisés depuis des décennies, alors qu'elles résultent d'un paradigme extrêmement récent, marqué par un effort sans précédent pour créer des machines cognitives autonomes, potentiellement capables de surpasser l'intelligence humaine [119]. Cette représentation trompeuse minimise les risques et les défis uniques posés par ces nouvelles technologies, dans un contexte de forte concurrence internationale qui fait peu de cas des dangers inhérents à l'intégration rapide de ces systèmes dans de nombreux aspects de la société.

3.3.2 Traitement biaisé des risques spécifiques

3.3.2.1 Emploi

Sections du rapport concernées : 1.4

L'analyse des effets de l'IA sur l'emploi présentée dans le rapport de la Commission est fondamentalement erronée, au vu de deux problèmes majeurs :

Hypothèse erronée d'un arrêt du progrès technologique

Le rapport suppose implicitement que le progrès en IA s'arrêtera au niveau actuel, ne considérant pas l'impact de modèles plus avancés que GPT-4. Cette hypothèse conduit à une sous-estimation systématique des risques :

- Elle ignore la rapidité des avancées en IA et les objectifs affichés des laboratoires de recherche, qui visent à créer une intelligence artificielle générale (*AGI*) [119]. Une *AGI*, par déduction d'après la définition

d'OpenAI, aurait la capacité d'automatiser l'essentiel du travail humain, cognitif dans un premier temps, et manuel à terme.

- Bien qu'il n'y ait pas de consensus sur le calendrier de développement de l'*AGI*, certains experts avancent des estimations très rapprochées, allant jusqu'à évoquer un horizon de 3 ans [120], [121]. Si ces projections s'avéraient exactes, cela pourrait impliquer une automatisation massive et soudaine de l'emploi. Même en l'absence de consensus sur ces questions, l'existence de telles préoccupations parmi une partie des experts justifierait que le rapport traite explicitement de ces scénarios potentiels.
- Elle néglige l'effet multiplicateur des futures avancées sur l'automatisation des tâches et des emplois. Une IA légèrement plus avancée pourrait non seulement effectuer des tâches existantes plus efficacement, mais aussi combiner ces tâches de manière nouvelle et automatiser des processus entiers. Ainsi, une petite avancée technologique pourrait déclencher une vague d'automatisation bien plus importante que prévu.
- Le couplage en cours des *LLM* et des robots indique également la potentialité d'un saut important dans l'automatisation des activités industrielles [122], [123].

Cette approche non prospective rend toutes les conclusions sur l'emploi excessivement optimistes.

Présentation orientée et sélection des données

En plus d'admettre sans fondement l'hypothèse d'un arrêt des progrès en IA, le rapport emploie des techniques rhétoriques et une sélection des données pour présenter une vision artificiellement positive :

- Il affirme en introduction, en gras : « Notre propre analyse empirique suggère un effet positif de l'IA sur l'emploi » (p. 41), puis « Dans 19 emplois sur 20, il existe des tâches que l'IA ne peut pas accomplir » (p. 41) sans arguments ni sources. Cette conclusion est en contradiction avec les données présentées par la suite.
- En effet, la première approche, basée sur des études passées, présente des résultats mitigés et non concluants. Les auteurs remettent eux-mêmes en cause la validité à court terme des études en raison d'un manque de recul historique, particulièrement pour l'IA générative : « ces études ne permettent pas encore de conclure à un effet sur un horizon de quelques années » (p. 44). Pourtant, de manière surprenante et sans justification, le rapport affirme ensuite : « Comme la précédente, cette approche conduit à prédire un effet positif de l'IA sur l'emploi. » (p. 45).

Cette conclusion est en contradiction avec la prudence et les nuances exprimées dans l'analyse des données.
- La seconde approche cite deux études aux résultats divergents : l'une suggère un potentiel de remplacement de 5,1 % des emplois [124], l'autre de 30 % [125]. Le rapport minimise cette disparité pourtant alarmante et conclut sans plus de justifications à un effet positif.
- Face à cette incertitude et ces risques, le rapport n'évoque à aucun moment la nécessité d'appliquer un principe de précaution.

L'analyse des effets de l'IA sur l'emploi présentée dans ce rapport est doublement biaisée : elle repose sur une hypothèse irréaliste d'un arrêt du progrès technologique, et même dans ce cadre favorable, elle déforme les données pour présenter une conclusion injustifiée et excessivement optimiste. Cette approche pourrait conduire à une dangereuse sous-estimation des risques de déstabilisation massive du marché du travail.

3.3.2.2 Cybersécurité

Sections du rapport concernées : 2 phrases p. 32 et 59

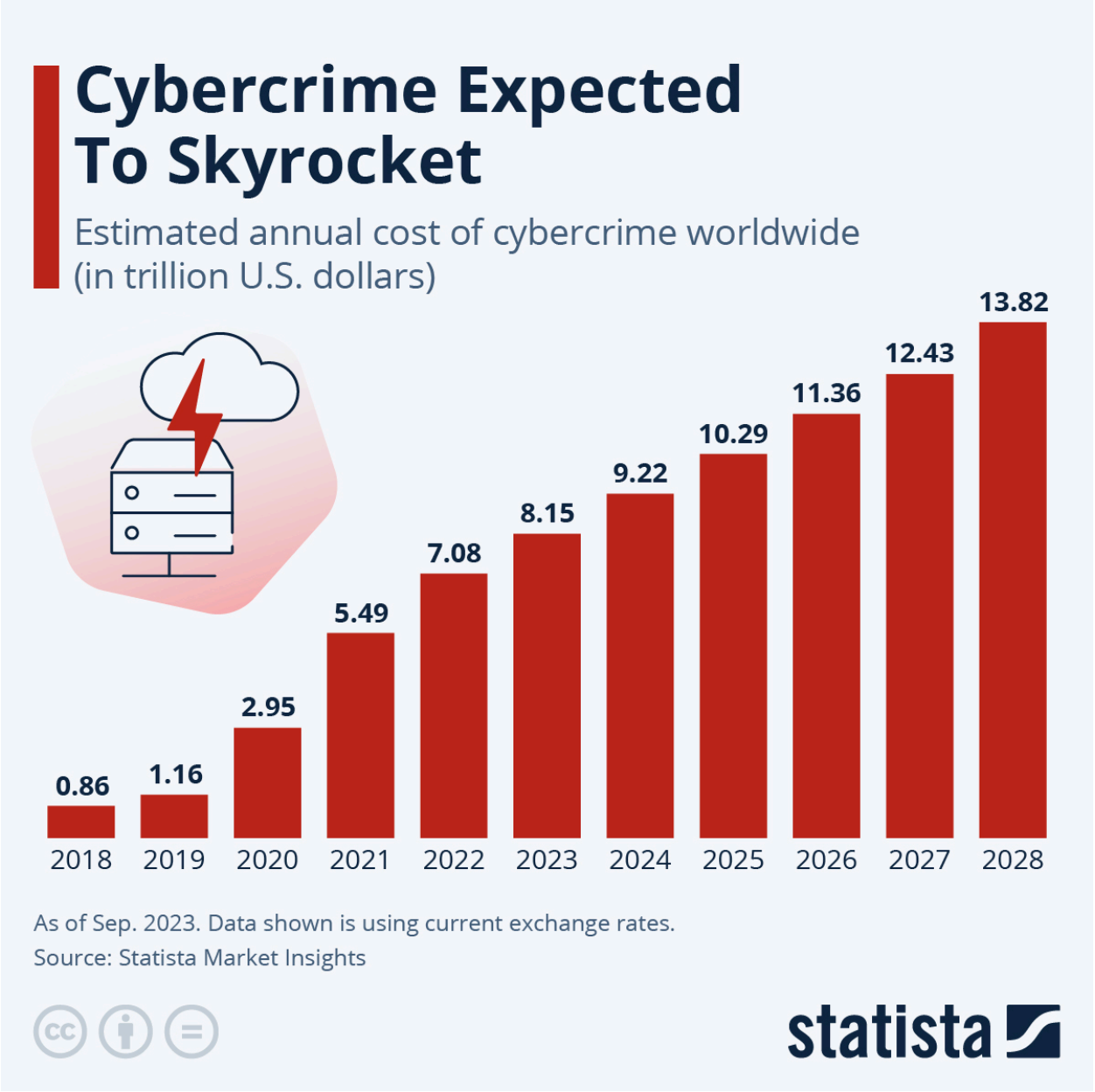
L'IA présente des défis nouveaux et significatifs en matière de cybersécurité, largement ignorés par le rapport de la Commission. De nombreuses études récentes [111], [113], [114], [126] mettent en évidence les risques de cybersécurité démultipliés par l'IA, ce qui aurait dû justifier un traitement approfondi de ces risques dans le rapport. Un état des lieux doit être effectué pour mesurer précisément l'ampleur de ces risques.

État des lieux des risques de cybersécurité liés à l'IA

- Capacités d'attaque améliorées : L'IA permet le développement de techniques d'attaque plus sophistiquées, notamment de phishing personnalisé à grande échelle [127], de *deepfakes* convaincants [128], et de malwares adaptatifs.
- Surface d'attaque élargie : L'adoption généralisée des systèmes d'IA introduit de nouvelles vulnérabilités, comme les attaques adversariales [129] ou l'extraction de données sensibles [130].
- Automatisation et évolution rapide des menaces : D'après le NCSC [131], l'IA permet aux cybercriminels d'automatiser leurs opérations, notamment la reconnaissance, l'ingénierie sociale et le développement de logiciels malveillants, ce qui rend les attaques plus efficaces et plus rapides.
- Découverte et exploitation automatisées de vulnérabilités : Les systèmes d'IA avancés peuvent analyser le code, trouver des failles et les exploiter

de manière autonome. GPT-4, par exemple, surpasse déjà 88 % des pirates humains dans certaines compétitions [114].
- Le NCSC prévoit avec une quasi-certitude une augmentation du volume et de l'impact des cyberattaques au cours des deux prochaines années.

Ces risques sont pris très au sérieux par les experts du domaine. 93 % des professionnels en cybersécurité [132] estiment qu'un « événement cyber catastrophique de grande ampleur est probable dans les deux prochaines années » et 97 % [133] estiment que leur organisation va subir un incident de sécurité causé par une IA.



Cette figure montre l'évolution projetée du coût annuel mondial de la cybercriminalité, exprimé en trillions de dollars américains. Les données révèlent une croissance accélérée entre 2018 et 2023. Les coûts devraient passer de 0,86 trillion USD en 2018 à 13,82 trillions USD en 2028.

Traitement de la cybersécurité dans le rapport

Le rapport de la Commission n'accorde qu'une attention minimale et biaisée à la question de l'impact de l'IA sur la cybersécurité, dont elle reconnaît pourtant l'existence. Le mot « cybersécurité » n'apparaît que deux fois dans les 130 pages du rapport :

- « Rien n'indique toutefois que l'IA changera durablement le rapport de force entre cybercriminels et ceux chargés de nous protéger, à condition que ces derniers puissent s'emparer de ces technologies. » (p. 32)
  Une étude du *Centre for the Governance of AI* (GovAI) [134] contredit cette affirmation, notant que « L'équilibre entre l'attaque et la défense risque de pencher davantage en faveur de l'attaque à mesure que les modèles de fondation deviennent de plus en plus performants. ». Il existe une asymétrie fondamentale entre l'attaque et la défense en cybersécurité [135], [136] : un défenseur doit sécuriser l'ensemble des points vulnérables de son infrastructure, tandis qu'un attaquant n'a besoin de n'en exploiter qu'un seul pour réussir. À ce jour, aucun mécanisme ne permet d'utiliser l'IA pour corriger systématiquement toutes les failles potentielles à travers la multitude d'écosystèmes existants, bien qu'il existe des spéculations selon lesquelles l'IA pourrait être utilisée pour automatiser la réaction aux attaques [137].
- « Pour les risques biologiques et cyber, rien n'indique que les modèles ouverts posent plus de risques que des modèles fermés. » (p. 59)
  Cette assertion n'est étayée par aucune preuve ou référence dans le rapport, notamment en ce qui concerne les risques cyber. Bien au contraire, tout indique que les modèles ouverts posent plus de problèmes que les modèles fermés, comme nous le détaillons dans la section 3.3.2.3 « Open source ». Alors que les modèles fermés peuvent encore être contrôlés par leurs développeurs pour prévenir des usages malveillants, les modèles open source échappent fondamentalement à tout contrôle, ce qui permet aujourd'hui déjà à des utilisateurs malintentionnés de les exploiter [138].

Le rapport ignore ainsi complètement l'ampleur et la complexité des menaces posées par l'IA dans le domaine de la cybersécurité et faillit à sa mission d'informer adéquatement les décideurs et le public sur les défis critiques auxquels nous sommes confrontés.

3.3.2.3 Open source

Sections du rapport concernées : 1.9

L'open source permet le partage libre et collaboratif de code, favorisant l'innovation et l'accessibilité technologique. Bien que bénéfique dans de nombreux domaines, cette approche peut poser des risques graves lorsqu'elle est appliquée à des technologies dangereuses. Dans un contexte où les experts du domaine alertent sur des risques potentiellement catastrophiques à court terme, la question de l'open source des modèles d'IA les plus avancés devient un enjeu majeur de sécurité nationale et internationale.

Les risques liés à l'open source des modèles d'IA avancés sont particulièrement préoccupants, comme le souligne une étude approfondie de GovAI [134].

- **Capacités dangereuses et menaces pour la sécurité** : L'open sourcing de modèles hautement compétents pourrait entraîner la diffusion de capacités dangereuses, susceptibles de causer des dommages physiques importants ou de perturber des fonctions sociétales essentielles. Des acteurs malveillants pourraient exploiter ces modèles pour faciliter la diffusion de désinformation à grande échelle et la surveillance coercitive de la population, lancer des cyberattaques contre des infrastructures critiques, ou encore faciliter le développement d'armes biologiques et chimiques [139].
- **Désalignement rapide et perte de contrôle** : Une fois un modèle rendu public, il échappe au contrôle de ses créateurs, car il est rapidement modifié pour supprimer les restrictions de sécurité imposées. En moyenne, une version sans restrictions apparaît dans les quelques jours suivant la publication d'un modèle open source, rendant inefficace toute tentative de limiter ou contrôler son utilisation malveillante. Ce processus est accessible à tous et ne coûte qu'environ 200 dollars [138].
- **Caractère irréversible de l'open sourcing** : Une fois un modèle rendu public, il devient impossible de l'éliminer ou de corriger des failles de sécurité découvertes ultérieurement, ce qui amplifie le risque d'exploitation malveillante de manière irréversible.

Face à ces préoccupations majeures, le traitement de la question de l'open source par la Commission de l'IA est alarmant par sa légèreté. Le rapport s'aligne sur la position promue par la société Mistral et le groupe Meta pour affirmer d'emblée que l'ouverture des modèles d'IA ne pose « pas de risque supplémentaire significatif » (p. 59), une conclusion en contradiction avec l'étude de GovAI. Cette position repose sur des hypothèses non justifiées et des omissions graves.

- La Commission ignore complètement les risques de cybersécurité liés à l'open source, se contentant d'affirmer leur inexistence sans aucune justification.
- Concernant les biorisques, le rapport se base sur une seule étude [140] montrant l'absence de risque avec les modèles actuels, ignorant ainsi la possibilité de risques futurs. Comme nous l'avons montré plus haut, le rapport suppose implicitement un arrêt soudain des progrès en IA : voici un exemple clair montrant que cela donne lieu à une sous-estimation dangereuse des risques.
- Le rapport admet que les modèles open source réduisent de 70 % le coût de production de la désinformation, mais il minimise ce risque en le considérant comme déjà inévitable. Il justifie cette position en affirmant que ces modèles sont déjà disponibles en open source, sans considérer que l'amélioration continue de ces technologies va certainement amplifier ce problème.

La Commission traite l'IA comme une technologie figée, sans danger significatif, et en conclut que l'open source ne présente pas de dangers supplémentaires. Cette approche est d'autant plus étonnante qu'elle ne serait certainement pas appliquée à d'autres technologies potentiellement dangereuses. On imagine mal, par exemple, des recommandations similaires pour rendre open source les plans de construction d'armes biologiques, nucléaires ou de virus informatiques.
Il est important de noter que cette analyse biaisée est certainement influencée par les conflits d'intérêts au sein de la Commission, de nombreux membres ayant des intérêts financiers directs dans le développement de l'IA open source que nous analysons en profondeur en section 4.2.

3.3.2.4 Création artistique

Sections du rapport concernées : 1.6

L'impact de l'IA sur la création artistique est un sujet de préoccupation croissante dans le monde de l'art et de la culture. Des mouvements d'artistes en colère, tels que ArtistsHate [141], émergent en réaction à la menace perçue de l'IA sur leurs métiers et leur créativité. Ces inquiétudes ne sont pas infondées, comme le montre une étude récente menée par CVL Economics [142] sur l'impact de l'IA générative dans l'industrie du divertissement.

État des lieux : une disruption massive en cours

L'étude *Future Unscripted* [139] révèle l'ampleur de la transformation en cours :

- 72 % des entreprises du divertissement sont des adopteurs précoces de l'IA générative, contre seulement 3,9 % dans l'économie globale.
- 203 800 emplois américains dans le divertissement devraient être perturbés d'ici 2026, principalement par consolidation ou remplacement.
- L'industrie du film, de la télévision et de l'animation sera la plus touchée, avec 21,4 % de sa main-d'œuvre impactée.
- Le secteur du jeu vidéo affiche le taux d'adoption le plus élevé, près de 90 % des entreprises utilisant déjà l'IA générative.

Ces chiffres préoccupants contrastent fortement avec le traitement superficiel et partial de la question par le rapport de la Commission de l'IA.

Une analyse déficiente et trompeuse

Le rapport de la Commission traite la question de l'impact de l'IA sur la création artistique de manière particulièrement problématique :

- Minimisation des impacts négatifs : Bien que le rapport reconnaisse brièvement les inquiétudes du secteur, il les balaie rapidement pour se concentrer sur les aspects positifs. Il n'y a aucune analyse approfondie des conséquences potentiellement dévastatrices pour certains métiers artistiques, en contradiction totale avec les données présentées dans l'étude de CVL Enonomics [142].
- Absence totale de données et d'études : Alors qu'une étude de Goldman Sachs [143] prévoit une automatisation de 26 % des tâches dans le domaine artistique rien qu'avec les IA actuelles, le rapport de la Commission ne s'appuie sur aucune donnée chiffrée ni aucune étude pour soutenir ses affirmations. Cette absence de rigueur scientifique est particulièrement frappante et compromet sérieusement la crédibilité de l'analyse.
- Vision à court terme et hypothèse d'arrêt du progrès : Comme dans le reste du rapport (voir section 3.2.3), l'analyse de la Commission souffre d'une vision à court terme qui suppose implicitement un arrêt brusque des progrès de l'IA. Cette approche ignore complètement la rapidité des avancées dans le domaine et les projections à moyen terme. Le rapport affirme par exemple que « L'IA ne met cependant pas en danger l'originalité de la création en elle-même et ses processus de sélection »

(p. 53), sans aucune justification ni prise en compte des développements futurs potentiels de l'IA. Cette affirmation péremptoire contraste fortement avec les inquiétudes légitimes exprimées par de nombreux artistes.

En conclusion, le traitement de la question de l'impact de l'IA sur la création artistique par la Commission est symptomatique des problèmes qui entachent l'ensemble du rapport : minimisation des risques, manque de rigueur scientifique et vision à court terme. Cette approche faillit gravement à la mission d'éclairer les décideurs et le public sur les défis réels posés par l'IA dans le domaine artistique et culturel, et pourrait conduire à une dangereuse sous-estimation des bouleversements à venir dans ce secteur crucial de notre société.

3.3.3 Utilisation stratégique du langage

Le rapport de la Commission de l'IA utilise des choix linguistiques qui orientent le lecteur pour minimiser les risques liés à l'IA générative et façonner une perception rassurante mais trompeuse des enjeux. Ces procédés rhétoriques se manifestent à travers les choix lexicaux, les connotations émotionnelles, la discréditation des scénarios catastrophiques et l'utilisation d'amalgames et de fausses analogies.

Le choix des termes et des expressions tout au long du rapport révèle une volonté délibérée de désamorcer les inquiétudes et de favoriser la confiance du public. Le titre de la première partie, « Dédiaboliser l'IA, sans pour autant l'idéaliser », illustre parfaitement cette approche. Il suggère que les craintes liées à l'IA sont exagérées et doivent être tempérées, tout en prétendant adopter une position équilibrée.

La discréditation des scénarios critiques de l'IA est particulièrement apparente dans le traitement des avertissements émis par de nombreux scientifiques. Le rapport qualifie ces scénarios « d'épouvante », une expression chargée émotionnellement qui les assimilent à des fantasmes irrationnels plutôt qu'à des préoccupations légitimes basées sur des analyses scientifiques. Cette caractérisation contraste fortement avec les expressions plus neutres comme « effets indésirables » ou « actes malveillants » utilisées pour décrire les risques reconnus par la Commission. Cette disparité dans le choix des termes vise clairement à miner la crédibilité des préoccupations les plus graves.

Le rapport s'appuie sur une comparaison fallacieuse entre l'IA et l'électricité. Cette analogie, présentée sans justification approfondie, sert à banaliser les

risques de l'IA en l'assimilant à une technologie familière et largement acceptée. Cependant, cette comparaison ignore les différences fondamentales entre ces deux technologies, notamment en termes de complexité, d'autonomie et d'impact sur la société. L'utilisation de cette fausse analogie révèle une tentative de simplification excessive des enjeux liés à l'IA.

Le rapport fait également des amalgames et de fausses analogies pour banaliser les risques liés à l'IA. La formulation « Faut-il avoir peur de l'IA ? Non, mais il faut être vigilant comme avec tout outil » (p. 31) est un exemple frappant de cette technique. En assimilant l'IA à « un simple outil », le rapport ignore la nature fondamentalement différente et potentiellement autonome des systèmes d'IA avancés, notamment les risques liés aux systèmes d'IA désalignés.

La structure même du rapport contribue à cette manipulation sémantique. Chaque chapitre commence par un scénario positif, créant d'emblée un cadre optimiste. Bien que le développement qui suit nuance parfois ce tableau, l'impression initiale positive persiste, influençant la perception globale du lecteur. Cette technique rhétorique est bien illustrée dans la partie traitant de la perte d'emploi. Alors que l'analyse présentée dans le corps du texte soulève des préoccupations sérieuses quant à l'impact de l'IA sur le marché du travail, le chapitre est encadré par des conclusions résolument optimistes. Il s'ouvre sur une vision positive des transformations à venir et se clôt sur des perspectives encourageantes, créant ainsi un effet de « sandwich optimiste » qui atténue la gravité des problèmes soulevés. Cette structure crée un décalage entre le contenu analytique, qui reconnaît des défis majeurs, et le message global véhiculé, qui minimise ces mêmes défis.

Cette approche linguistique révèle un biais d'optimisme prononcé et une volonté de traiter les enjeux de l'IA de manière superficielle. En minimisant systématiquement les risques à travers le langage utilisé, le rapport de la Commission échoue à fournir une analyse équilibrée et approfondie des défis posés par l'IA, compromettant ainsi sa crédibilité et sa pertinence en tant que document d'orientation stratégique.

3.4 Négligence des préoccupations citoyennes et approche non-démocratique

3.4.1 Perception publique de l'IA : un mélange de préoccupations et d'attentes

Selon plusieurs sondages récents [144], [145], [146], les Français expriment des inquiétudes significatives concernant l'IA :

- 72 % sont préoccupés ou ambivalents vis-à-vis de l'IA, le taux le plus élevé parmi 21 pays sondés.
- 78 % s'inquiètent de la sécurité des données, de la vie privée et des droits d'auteur.
- 91 % se sentent mal informés des risques et implications de l'IA.
- 68 % sont favorables à l'établissement de règles contraignantes au niveau étatique.

Parallèlement, les Français reconnaissent le potentiel positif de l'IA dans certains domaines comme la santé et la recherche, mais expriment une forte méfiance quant à son utilisation pour des décisions critiques (justice, transport, diagnostic médical).
D'autre part, l'utilisation des IA dans le contexte des réseaux sociaux est reconnue comme néfaste tant du point de vue de la santé des utilisateurs que de la désinformation et des chambres d'écho dans lesquelles ils se retrouvent enfermés.

3.4.2 Décalage entre les préoccupations publiques et l'approche de la Commission

Le rapport de la Commission reconnaît l'existence d'une perception négative de l'IA par le public français. Cependant, au lieu d'explorer en profondeur les raisons de ces inquiétudes et d'envisager une possible divergence fondamentale entre les aspirations du public et la direction actuelle du développement de l'IA, la Commission semble privilégier une approche paternaliste, visant à modifier cette perception.

Le rapport mentionne bien l'importance du débat public :

« Nous recommandons de lancer immédiatement un plan de sensibilisation et de formation de la nation. Pour y parvenir, nous devons

d'abord créer les conditions d'une appropriation collective de l'IA et de ses enjeux. Cela suppose d'animer en continu des débats publics dans notre société [...] » (p. 7)

Cependant, les propositions concrètes semblent davantage orientées vers la familiarisation et l'acceptation de l'IA que vers une consultation citoyenne sur les orientations à prendre :

« [...] susciter la création de lieux d'expérimentation et d'appropriation de la technologie (les « cafés IA »), de mettre à disposition un outil numérique d'information ou encore de lancer un concours de cas d'usages positifs de l'IA. » (p. 7)

Cette approche, bien qu'elle mentionne le dialogue, privilégie une vision de communication unilatérale descendante, de certains experts vers le public, où l'objectif principal est de former et de convaincre le public plutôt que de l'impliquer véritablement dans les décisions concernant le développement et le déploiement de l'IA.

De plus, depuis la publication du rapport, l'accent semble avoir été mis principalement sur les « cafés IA », une initiative qui vise avant tout à réconcilier le public avec l'IA, plutôt que sur l'organisation de véritables débats démocratiques publics permettant d'aborder les préoccupations citoyennes de manière approfondie. Il est d'ailleurs à noter que l'occultation et la minimisation des dangers par le rapport (voir sections 3.2.2 et 3.3) contribue à un débat public mort dans l'œuf, car les citoyens seraient privés des informations nécessaires pour prendre des décisions éclairées.

Dans un contexte de grande incertitude sur les risques et bénéfices potentiels de l'IA, il convient de garder l'esprit ouvert et de consulter la population française sur les principes qui guideront la recherche et le développement de l'IA, afin de respecter ses priorités sans chercher à imposer une vision de l'IA avant qu'un consensus scientifique et sociétal s'affermisse sur le sujet.

Cette approche soulève des questions sur la légitimité démocratique des décisions recommandées par le rapport. Les inquiétudes du public et les avertissements des experts reflètent des appréhensions légitimes face à cette technologie en rapide évolution. En l'absence de consensus scientifique, il convient d'informer tant sur les risques que sur les bénéfices potentiels, et par respect des opinions citoyennes, une gouvernance plus inclusive de l'IA doit prendre en compte ces préoccupations pour assurer une gestion démocratique et éclairée de cette technologie.

3.4.3 Proposition controversée sur l'accès aux données

Le rapport suggère de « faciliter l'accès aux données à caractère personnel » (p. 10) pour l'entraînement des modèles d'IA, par exemple en supprimant les procédures d'autorisation préalable à l'accès des données (p. 101). Cette recommandation va à l'encontre des préoccupations majeures du public concernant la protection de la vie privée, mais soulève également de sérieuses préoccupations éthiques et pratiques.

L'intégration de données personnelles dans les modèles de langage (*LLM*) présente des risques significatifs souvent sous-estimés :

- Les *LLM* ont tendance à mémoriser une grande partie des données sur lesquelles ils sont entraînés. Ces informations se retrouvent encodées dans les milliards de paramètres du modèle.
- Des techniques de *« jailbreaking »* peuvent potentiellement être utilisées pour extraire ces données personnelles du modèle. Cela signifie qu'un attaquant pourrait, en théorie, récupérer des informations privées qui n'auraient jamais dû être accessibles [147].
- L'utilisation de données personnelles pour l'entraînement de modèles d'IA soulève des questions importantes sur le consentement éclairé et la transparence envers les individus dont les données sont utilisées.

Le rapport ne mentionne aucun des risques propres à l'IA pour les données personnelles. Cette approche révèle une propension inquiétante à privilégier le développement technologique au détriment des droits fondamentaux à la vie privée et à la protection des données personnelles. Elle nécessite un débat public approfondi et une évaluation rigoureuse des risques avant toute mise en œuvre.

Face à ces lacunes dans l'approche de la Commission, **notre contre-expertise se positionne comme une réponse à l'appel au débat public mentionné dans le rapport**, tout en allant bien au-delà. Nous visons à initier un dialogue véritablement démocratique et inclusif sur l'IA, en fournissant une analyse critique approfondie et en encourageant la participation d'une diversité d'experts et de citoyens.

3.5 Analyse comparative

3.5.1 Introduction

L'essor spectaculaire des modèles de fondation en intelligence artificielle a incité de nombreux pays et organisations internationales à anticiper les bénéfices et les risques liés à leur développement et à leur déploiement.

En novembre 2023, le gouvernement britannique a lancé une initiative pionnière de coordination internationale à Bletchley, près de Londres [148]. Ce sommet a réuni des représentants de nombreux gouvernements et organisations, aboutissant à une déclaration conjointe, signée par la France, soulignant l'importance de développer des IA de manière sécurisée. Cette déclaration reconnaît, en particulier, le potentiel de dommages catastrophiques causés par des modèles de pointe en raison de leurs capacités avancées, citant les risques en biosécurité, cybersécurité et désinformation [149].

Depuis lors, plusieurs pays et groupes de réflexion ont mandaté des experts pour dresser un état des lieux de l'intelligence artificielle et anticiper ses trajectoires de développement potentielles. Cette démarche est cruciale pour éclairer les décisions futures dans ce domaine.

Le rapport de la Commission française sur l'IA se voulait, en principe, aligné sur cette approche. Cependant, il se démarque par ses lacunes comparé à d'autres initiatives similaires. Le document échoue à traiter la question avec le sérieux qu'elle mérite, éludant la plupart des risques sans les examiner en profondeur. Il ne cite aucun expert dans les domaines de risques les plus couramment anticipés, tels que les biorisques ou la sécurité informatique, et ne propose aucune solution politique ou institutionnelle crédible face à ces enjeux.

En négligeant d'aborder les risques liés à l'IA avec la rigueur nécessaire, le rapport se trouve en contradiction avec les résolutions co-signées par la France à Bletchley.

Des exemples de rapports plus complets et rigoureux existent pourtant. C'est pourquoi nous analysons ici les points forts d'autres initiatives comparables, dont les conclusions diffèrent nécessairement et significativement de celles du rapport de la Commission.

Pour évaluer la qualité du rapport de la Commission française sur l'IA, nous l'avons comparée à quatre initiatives récentes, reconnues pour leur rigueur et leur influence dans le domaine de la gouvernance de l'IA avancée. Ces

documents représentent diverses approches, allant de l'analyse scientifique approfondie à la législation concrète, et couvrent des régions géopolitiques clés.

1. **Le *Gladstone Report*** *[150]* : Commandé par le Département d'État américain, ce rapport offre une évaluation approfondie des risques liés à l'IA avancée. Il se distingue par son analyse détaillée des risques catastrophiques et propose un plan d'action gouvernemental inédit, structuré autour de cinq axes d'effort.
2. **L'*International Scientific Report on the Safety of Advanced AI* (*ISR*)** [151] : Cette initiative scientifique mondiale réunit des experts de 30 pays, de l'UE et de l'ONU. Ce rapport intérimaire se concentre sur les systèmes d'IA à usage général, comme les grands modèles de langage. Il examine en profondeur les capacités actuelles et futures de l'IA, les méthodologies d'évaluation, les risques et les approches techniques pour les atténuer.
3. **La loi européenne sur l'IA** [152] : C'est la première réglementation complète sur l'intelligence artificielle proposée par un régulateur majeur. Initié par la Commission européenne, cet acte législatif vise à établir un cadre juridique pour le développement et l'utilisation de l'IA au sein de l'Union européenne. Il adopte une approche basée sur les risques, catégorisant les applications d'IA selon leur niveau de danger potentiel.
4. **L'*Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (*Executive Order*)** américain sur l'IA [153] : Signé par le président Joe Biden, ce décret exhaustif couvre un large éventail de domaines, de la sécurité nationale à la protection des consommateurs. Il établit des principes directeurs pour le développement responsable de l'IA, impose de nouvelles exigences aux entreprises développant des systèmes d'IA puissants, et mandate de nombreuses actions spécifiques des agences fédérales.

Les deux rapports (*Gladstone Report* et *ISR*) abordent en profondeur de nombreux risques fondamentaux. Ils font appel à de nombreux spécialistes et à une littérature abondante, évaluent les arguments pour et contre différents scénarios, et **expriment leurs conclusions en partageant leurs degrés d'incertitude**. Les deux textes législatifs (loi européenne sur l'IA et *Executive Order*), quant à eux, font un effort d'anticipation des risques très important tout en cherchant à respecter la libre entreprise.

La comparaison de ces textes avec le rapport de la Commission française montre que ce dernier ne remplit pas son rôle d'information concernant les différents risques majeurs d'un développement mal maîtrisé de l'IA. Notre évaluation est résumée dans la figure suivant

Risque	Rapport français	*Gladstone Report*	*ISR*	*Loi européenne sur l'IA*	*Executive order*
Cybersécurité	2 phrases 0 références (p. 32, 59)	Plusieurs mentions et paragraphes, interview de plusieurs experts en cybersécurité 7 références	2 pages (4.1.3) 18 références	Récital 110, et plusieurs chapitres pour encadrer les systèmes à haut risque	26 mentions concernant les risques 2 concernant le potentiel pour la défense
Bioterrorisme	3 phrases 1 référence (p. 32, 59)	Plusieurs mentions et discussions 3 références	3 pages (4.1.4) 20 références	Récital 110, et plusieurs chapitres pour encadrer les systèmes à haut risque	12 mentions
Perte d'emploi	7 pages 10 références mais analyse biaisée et trompeuse.		3 pages (4.3.1) 37 références	Non mentionné	9 mentions et demande d'un rapport par le *Council of Economic Advisers* et le *Secretary of Labor*
Perte de contrôle	1 phrase 0 référence Risques ridiculisés	Extrêmement étayé 20 références	3 pages (4.2.3) 45 références	Chapitre 3 et plusieurs autres chapitres pour encadrer les systèmes à haut risque	1 mention et plusieurs systèmes d'évaluation pour les capacités dangereuses non spécifiquement conçus pour le risque de perte de contrôle
Guerre informationnelle	1,5 pages 1 référence		2 pages (4.2.1) 34 références	Récitals 132 et 133	
Armes autonomes	Non mentionné				
Violations de confidentialité	2 phrases 0 référence		2 pages (4.3.5) 20 références	Article 15	37 mentions et demande au *Secretary of Commerce* de faire une évaluation des meilleures pratiques dans le domaine sans imposer de restrictions particulières aux développeurs des modèles
Biais	Non mentionné		2 pages, (4.2.2) 45 références	Récital 67	29 mentions et ordonne l'évaluation de ces risques au *Secretary of Health and Human Services* et au *Secretary of Labor*
Deepfakes	2 phrases 0 référence		Plusieurs mentions (4.2.1 et 4.3.1) 17 références	Récitals 132 et 133	Non mentionné
Escalade des conflits internationaux	Non mentionné	Le rapport reconnaît les risques et propose des solutions intégrées à des LOE pour mitiger les risques d'escalade.		Non mentionné	
Impact environnemental	2 pages 6 références		2 pages 12 références	Non mentionné	Non mentionné

Légende
Hors du champ d'application déclaré du rapport
Aucune mention ou mention très brève sans analyse
Mentionné avec une analyse superficielle et des références peu nombreuses ou biaisées. Le rapport évoque le risque, mais l'analyse est limitée ou trompeuse, avec des références rares ou sélectionnées pour soutenir un point de vue particulier.
Analyse modérée avec quelques références variées, mais potentiellement incomplète. Le rapport fournit une analyse substantielle, bien que certains aspects ou points de vue importants puissent être sous-représentés.
Analyse approfondie avec des références solides et diverses, offrant une image équilibrée des débats ou controverses existants. Le rapport cite une variété de sources représentant différents points de vue sur le sujet.

3.5.2 Les risques éludés par le rapport français

3.5.2.1 Perte de contrôle des modèles d'IA avancés

Le *Gladstone Report (p. 24 et 36-37)* et *l'ISR (section 4.2.3)* examinent sérieusement la possibilité de perdre le contrôle d'IA de niveau humain ou supérieur tandis que la loi européenne sur l'IA (chapitre 3) et l'*Executive Order* américain proposent des mesures concrètes pour évaluer ces risques et contrôler le développement des modèles.

En revanche, le rapport de la Commission française se contente d'une phrase pour écarter ces préoccupations. Il utilise une longue analogie comparant l'IA à l'adoption de l'électricité, ce qui est trompeur car l'IA représente un changement bien plus fondamental et d'une toute autre nature (voir section 3.3.3).

3.5.2.2 Biorisques et cybersécurité

Le *Gladstone Report (p. 21)* et *l'ISR (section 4.1.4)* décrivent des scénarios où un modèle de fondation futur serait utilisé pour transmettre des informations critiques dans la fabrication d'armes biologiques, voire d'automatiser leur production. Les risques de facilitation des cyberattaques sont également reconnus par l'*ISR*, qui reste serein concernant les modèles publiés au moment de la rédaction du rapport (4.1.3). Le *Gladstone Report* prend ce risque au sérieux (ex. p. 24) et recommande la création d'institutions surveillant ces capacités, comme un Observatoire de l'IA (p. 51). La loi européenne sur l'IA classe ces modèles comme présentant des « risques systémiques » (*Recital* 110) et exige des contrôles stricts. L'*Executive Order* reconnaît ces dangers et propose des mesures pour les évaluer et les atténuer.

Le rapport français affirme simplement que « rien n'indique » que les modèles actuels soient particulièrement dangereux à cet égard. Cette affirmation est non seulement fausse concernant les cyberattaques (voir section 3.3.2.2), mais elle ignore surtout la trajectoire rapide des progrès en IA.

3.5.2.3 Deepfakes et désinformation

L'*ISR* reconnaît les risques liés aux *deepfakes* et à la désinformation (p. 41-43), précisant qu'aucune technique robuste n'existe pour les prévenir (p. 76-77, 82-83). Le *Gladstone Report* souligne l'inquiétude des chercheurs concernant la manipulation de l'opinion publique (p. 35, 43, Annexe F) et propose des systèmes de détection en amont (p. 273). La loi européenne sur l'IA interdit le

déploiement de modèles d'IA utilisant des techniques de désinformation et reconnaît les risques liés aux *deepfakes* (*Recitals* 132 et 133).

Le rapport français sous-estime ces risques. Sa principale solution, la labellisation des contenus générés par IA, est insuffisante et techniquement inachevée (ISR p. 77), alors même que les risques liés à la désinformation en ligne sont désormais bien documentés et pourraient se voir rapidement amplifiés avec l'IA générative (voir section 3.2.4).

3.5.2.4 Confidentialité des données

L'*ISR* souligne que les modèles de langage actuels sont peu sécurisés et peuvent divulguer des informations confidentielles (p. 60-61), précisant qu'aucune technique n'est connue pour empêcher totalement ces fuites (p. 81-82). La loi européenne sur l'IA ordonne aux développeurs de rendre leurs modèles plus sûrs (Article 15).

Le rapport français ne traite pas sérieusement cette question. Au contraire, il suggère d'adapter la CNIL pour permettre une utilisation plus large des données d'utilisateurs, sans reconnaître l'absence de solution technique pour garantir leur confidentialité (voir section 3.4.3).

3.5.2.5 Autres risques non traités

Le rapport français omet plusieurs risques importants :

- L'escalade des conflits internationaux due à une course à l'armement en IA (*Gladstone Report* p. 19, 85).
- La création et la mise à disposition du public de capacités de création d'armes autonomes (*Gladstone Report* p. 84, *ISR* p. 12, 16).
- Les biais discriminatoires que les modèles peuvent intégrer (*ISR* p. 49-51, loi européenne sur l'IA, *Recital* 67, *Executive Order*).

Le *Gladstone Report* préconise la mise en place d'une Agence Internationale de l'IA et un contrôle international sur la chaîne d'approvisionnement globale de l'IA (p. 19). La loi européenne sur l'IA (chapitre III et IX) et l'*Executive Order* recommandent la mise en place de régulations et de systèmes de vérification des données d'entraînement et du comportement des modèles.

Cette comparaison montre que le rapport français ne traite pas de manière adéquate de nombreux risques liés au développement de l'IA, contrairement aux initiatives internationales qui adoptent une approche plus complète et prudente.

3.5.3 Les risques traités superficiellement

Le rapport de la Commission française n'aborde en profondeur que deux catégories de risques, non catastrophiques. Malgré leur importance, le traitement reste insuffisant comparé aux autres rapports internationaux.

3.5.3.1 Pertes d'emploi

Le rapport consacre 7 pages à la question de la perte d'emploi, en faisant le risque le plus discuté. Cependant, comme détaillé dans la section 3.3.2.1, l'analyse présente plusieurs faiblesses :

- Les données présentées suggèrent que peu d'études ont été réalisées.
- Les résultats sont mitigés et ne suggèrent pas un effet positif évident de l'IA sur l'emploi.
- Les études sont basées sur des données passées, ne prenant pas en compte l'évolution rapide des modèles d'IA.
- Le rapport n'étudie pas l'effet économique d'une potentielle automatisation à bas coût d'une grande partie du travail humain.

En contraste, l'*ISR* adopte une approche plus prudente, notant l'incertitude et l'absence de consensus parmi les économistes sur ce sujet (voir section 4.3.1 du rapport de l'*ISR*). Il présente une vision plus riche des différentes difficultés liées à cette situation inédite.

L'*Executive Order* américain va plus loin en demandant à son administration de préparer un rapport spécifique sur les conséquences de l'IA sur le marché du travail, et de proposer des solutions aux risques identifiés.

3.5.3.2 Impact environnemental

Malgré un effort réel d'estimation, le rapport de la Commission minimise l'impact environnemental de l'IA. Il adopte une approche qui manque de rigueur scientifique en considérant l'impact de l'IA comme nécessairement positif, ce qui lui permet de le comparer favorablement en regard de son coût environnemental.

En revanche, l'*ISR* consacre deux pages à une discussion plus approfondie et arrive à une conclusion radicalement différente (p. 59-60) :

- Malgré les incitations à rendre les modèles plus économiques en énergie, l'augmentation de la demande excède largement les améliorations dans l'efficacité des entraînements.

- Les entraînements des modèles d'IA augmentent très rapidement en taille et en nombre, ce qui accroît leur impact environnemental.

Il est important de souligner que si l'IA peut potentiellement offrir des solutions aux problèmes environnementaux dans le futur, ces bénéfices restent pour le moment purement spéculatifs, et ne pourront être obtenus que si l'on parvient à éviter les risques sous-estimés par le rapport de la Commission. En revanche, les coûts environnementaux actuels sont bien réels et considérables.

Par exemple, l'entraînement de GPT-3 aurait consommé autant d'électricité que 100 foyers en un an [154]. Le fonctionnement de ChatGPT nécessite autant d'énergie qu'une petite ville [155]. Les centres de données qui hébergent ces modèles consomment des millions de litres d'eau par jour pour leur refroidissement, parfois au détriment des ressources en eau potable locales [156]. De plus, la production et le remplacement fréquent du matériel informatique nécessaire impliquent l'utilisation de nombreux produits chimiques toxiques [157], [158] et contribuent à l'épuisement des ressources naturelles.

3.5.4 Conclusion

N'envisageant pas de scénarios crédibles et pluriels pour le développement de l'IA, la Commission ne propose logiquement pas de réponses concrètes pour faire face aux défis potentiels.

La mission confiée par le gouvernement à la Commission est cruciale et survient à un moment très important de notre histoire. Les biais et le manque de rigueur dans la production du rapport sont nombreux et préoccupants. On note l'emploi récurrent d'une rhétorique de la dérision pour qualifier les tenants d'une posture prudente et argumentée.

L'approche de la Commission contraste fortement avec les résolutions co-signées par la France lors du sommet de Bletchley, ce qui ne peut qu'affecter sa crédibilité internationale sur les questions d'IA. Cela réduit également nos chances d'anticiper et de prévenir les risques les plus importants liés au développement rapide de l'intelligence artificielle.

4 Analyse de la composition de la Commission de l'IA

Dans un contexte où l'IA soulève des enjeux sociétaux majeurs, l'impartialité des instances guidant les politiques publiques est impérative. Comme nous l'avons montré dans la partie précédente, le rapport de la Commission de l'IA présente des omissions sur les risques et un enthousiasme marqué pour l'open source. Ces positions coïncident fortement avec les intérêts de certaines entreprises représentées au sein de la Commission, notamment Meta et Mistral, ce qui soulève des questions légitimes sur la composition de la Commission et son influence sur les recommandations produites.

Cette section vise à mettre en lumière les dysfonctionnements dans la composition de la Commission qui pourraient expliquer les lacunes du rapport. Nous examinerons le manque de diversité, les conflits d'intérêts majeurs et les actions controversées de certains membres, soulignant l'importance d'une composition plus neutre pour des recommandations plus équilibrées.

Points clés à retenir :

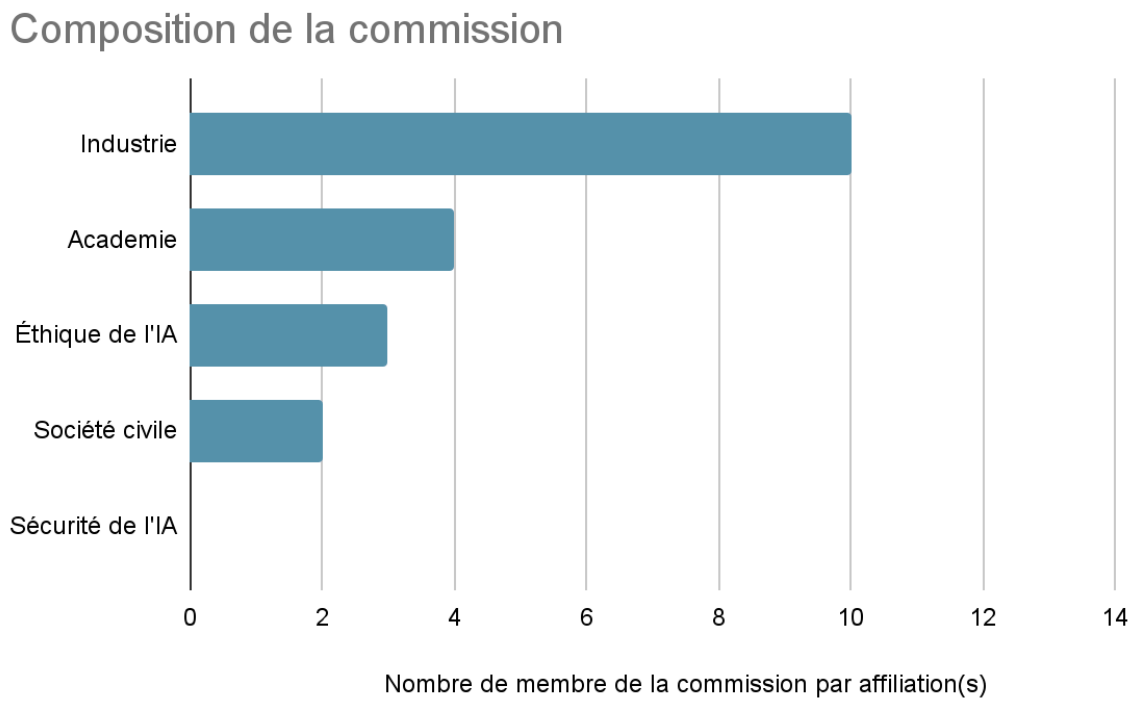
- Une surreprésentation significative de l'industrie par rapport aux experts indépendants, notamment en éthique et sécurité de l'IA.
- Des conflits d'intérêts majeurs chez plusieurs membres influents, liés à des géants technologiques américains et start-ups d'IA.
- Des actions de lobbying et des prises de position controversées remettant en question l'impartialité du processus.
- Ces éléments fournissent une explication potentielle aux lacunes identifiées dans le rapport de la Commission.

4.1 Manque de diversité

La composition d'une Commission chargée d'élaborer des recommandations sur l'IA influence directement la qualité et l'exhaustivité de ses conclusions. La présence d'experts en éthique et en sécurité de l'IA est particulièrement importante, car ces disciplines se penchent spécifiquement sur les impacts sociétaux et les risques liés à cette technologie.

Pour analyser la composition de la Commission, nous avons choisi de nous concentrer sur des axes qui nous semblaient particulièrement pertinents : la

représentation de l'industrie par rapport au monde académique, les représentants de la société civile, et la présence d'experts en éthique et en sécurité de l'IA. Cette catégorisation s'est basée sur les affiliations principales de chaque membre au moment de la Commission. Une description détaillée de la méthodologie est disponible en annexe.



Nombre de membres de la Commission par affiliation(s). La plupart des membres viennent du monde de l'industrie tandis que les académiques, la société civile et les experts en éthique de l'IA sont sous représentés et qu'aucun expert en sécurité de l'IA n'en fait partie.

Ce graphique illustre clairement la surreprésentation du secteur privé par rapport au monde académique, ainsi que l'absence d'experts en sécurité de l'IA et la sous-représentation des spécialistes en éthique de l'IA.

Éthique et sécurité de l'IA

Bien que la distinction entre éthique et sécurité de l'IA soit parfois contestée, elle est utile pour notre analyse. Nous définissons l'éthique de l'IA comme la discipline examinant les implications morales et sociétales de l'IA, en établissant des principes pour son développement et son utilisation responsables, tandis que la sécurité de l'IA vise à développer des systèmes d'intelligence artificielle sûrs et fiables, en minimisant les risques de dommages non intentionnels.

Cette composition déséquilibrée a des implications directes sur les lacunes du rapport :

1. L'absence totale d'experts en sécurité de l'IA explique en partie l'omission des risques les plus sévères et le manque d'anticipation des progrès futurs.
2. La sous-représentation des experts en éthique contribue à la minimisation des risques d'utilisations malveillantes.
3. La surreprésentation de l'industrie, dont les intérêts penchent vers moins de régulation, renforce ces tendances.

Dans un entretien [159], Anne Bouverot, co-présidente de la Commission, justifie la forte présence de l'industrie par le fait que la majorité de l'innovation en IA provient du secteur privé. Cependant, cet argument ne tient pas compte de la nature transversale de l'IA et de ses impacts sociétaux qui dépassent largement le cadre de l'innovation technologique.

Nous déplorons également le manque de transparence de la Commission concernant les experts consultés durant le processus. Aucun des experts en sécurité de l'IA que nous avons nous-mêmes consultés n'avait été approché par la Commission.

Pour une Commission plus équilibrée et à même de produire des recommandations tenant compte de tous les enjeux, nous aurions souhaité voir au minimum une parité entre représentants de l'industrie, du monde académique et de la société civile, avec au moins la moitié des membres représentant l'éthique et la sécurité de l'IA pour contrebalancer les intérêts privés.

L'absence de certains types d'experts (spécialistes de l'alignement, chercheurs en sécurité de l'IA) est particulièrement préoccupante. Bien que la présence de figures éminentes de l'industrie comme Yann LeCun soit compréhensible, il est essentiel d'assurer un équilibre des perspectives. Par analogie, une Commission sur l'énergie incluant des représentants de l'industrie pétrolière devrait également intégrer des experts en environnement et en énergies renouvelables. De même, une Commission sur l'IA devrait inclure des experts en sécurité et en éthique de l'IA pour contrebalancer les perspectives de l'industrie.

4.2 Conflits d'intérêts

Les conflits d'intérêts au sein d'une Commission chargée de guider les politiques publiques sont toujours préoccupants, mais ils prennent une dimension particulière dans le domaine de l'intelligence artificielle. Un conflit d'intérêts survient lorsqu'un individu est en position d'exploiter sa capacité professionnelle ou officielle d'une manière qui pourrait bénéficier à ses intérêts personnels ou corporatifs au détriment de l'intérêt général.

L'IA, en tant que technologie transversale, impacte un large éventail de secteurs industriels, ce qui signifie qu'une grande partie des représentants de l'industrie pourraient être considérés comme ayant un certain degré de conflit d'intérêts. La surreprésentation de l'industrie au sein de la Commission, démontrée dans la section précédente, accentue ce risque et souligne l'importance d'intégrer des perspectives diverses pour équilibrer les enjeux de l'IA.

Notre analyse se concentre sur les conflits les plus prononcés, impliquant les représentants des entreprises créatrices de modèles d'IA, qui seraient les plus directement impactés par d'éventuelles régulations. Nous avons également considéré les efforts de lobbying anti-régulation, notamment de Meta, Microsoft, Google et Mistral, pour identifier les cas où les intérêts personnels ou corporatifs pourraient influencer les recommandations de la Commission.

Activités de lobbying des entreprises concernées :

1. Meta (anciennement Facebook) déploie des efforts massifs et une stratégie coordonnée pour influencer la régulation de l'IA. En 2023, l'entreprise a investi 14,6 millions de dollars en lobbying direct auprès du Congrès américain et de l'administration Biden sur l'IA [160]. Meta finance l'American Edge Project, un groupe de pression qualifié de « chien d'attaque anti-réglementaire de Facebook » [161], qui a reçu 85,5

millions de dollars entre 2019 et 2022. Ce groupe a dépensé plus de 150 000 dollars en publicités Facebook contre la régulation de l'IA entre février et juin 2024 [162].

2. Google investit des ressources considérables dans le lobbying contre la régulation de l'IA. En 2023, l'entreprise a dépensé 9,2 millions de dollars en lobbying auprès des législateurs américains sur des questions liées à l'IA et à la propriété intellectuelle [160]. L'entreprise soutient que la doctrine du « *fair use* » protège l'IA des infractions au droit d'auteur. Google s'oppose activement à la proposition de loi californienne SB 1047 visant à réguler l'IA [163], tout comme d'autres géants technologiques.
3. Mistral AI a joué un rôle crucial dans le lobbying contre la loi européenne sur l'IA [164]. L'entreprise a rapidement obtenu un accès privilégié aux plus hauts niveaux de décision, notamment grâce à Cédric O, son principal lobbyiste et membre de la Commission de l'IA. Ancien secrétaire d'État au Numérique, O a activement fait pression pour diluer les exigences sur les développeurs d'IA à usage général. Ce lobbying intensif a eu pour effet de faire changer la position française au sujet de l'AI Act. (Pour plus de détails sur ce sujet, voir la section 4.3.1)

Sur cette base, nous avons identifié les personnes suivantes pour une analyse approfondie de leurs potentiels conflits d'intérêts : Yann LeCun (Meta), Arthur Mensch (Mistral), Cédric O (Mistral), Joelle Barral (Google).

Dans les sections suivantes, nous examinerons en détail les conflits d'intérêts potentiels de ces membres de la Commission, en analysant comment leur affiliation à ces entreprises pourrait influencer leurs positions et recommandations. Nous considérons uniquement les conflits d'intérêts pendant la durée d'exercice de la commission, c'est-à-dire entre septembre 2023 et mars 2024.

Une liste complète des activités professionnelles de tous les membres de la Commission est disponible en annexe.

4.2.1 Yann LeCun

- Vice-président & scientifique en chef de l'IA chez Meta (anciennement Facebook).
- Cofondateur, conseiller à Element Inc. Entreprise R&D de reconnaissance biométrique qui s'associe à des institutions financières, des gouvernements et des organismes de santé pour transformer la manière dont ils fournissent des services.

Yann LeCun est une figure emblématique dans le domaine de l'intelligence artificielle. Lauréat du prestigieux prix Turing en 2018, souvent considéré comme un équivalent du prix Nobel en informatique, il est également membre de l'Académie des sciences française [165] et Chevalier de la Légion d'honneur [166]. Ces distinctions témoignent de l'importance de ses contributions scientifiques et de sa reconnaissance au plus haut niveau de la communauté scientifique et de l'État français.

L'influence de LeCun s'étend bien au-delà du monde académique. Avec plus de 800 000 *followers* sur Twitter [167] et 782 000 sur LinkedIn [168], sa voix est prédominante dans le débat public sur l'IA. Sa présence médiatique est considérable, avec des centaines d'apparitions et mentions dans les médias en 2023. Cette influence a été reconnue par le magazine TIME, qui l'a inclus dans sa liste « Time 100 AI » [169] des personnes les plus influentes dans le monde de l'IA. Son rayonnement atteint les sphères politiques et économiques les plus élevées, comme en témoignent ses rencontres avec le Président Emmanuel Macron [170] et sa participation au Forum économique mondial de Davos [171]. Cette stature exceptionnelle dans les milieux scientifiques, médiatiques et politiques suggère que les opinions de LeCun pourraient avoir un poids significatif au sein de la Commission.

Malgré son expertise incontestable, la présence de Yann LeCun au sein de la Commission soulève des questions importantes quant aux conflits d'intérêts potentiels. En effet, LeCun occupe actuellement le poste de Vice-Président et Directeur Scientifique en IA chez Meta (anciennement Facebook), une entreprise américaine qui a été au cœur de nombreux scandales éthiques majeurs ces dernières années.

L'éthique contestable de Meta

Meta (anciennement Facebook) a été au cœur de plusieurs scandales éthiques majeurs :

1. Le scandale Cambridge Analytica (2018) [169] a révélé l'exploitation massive de données personnelles de millions d'utilisateurs à des fins de manipulation politique.
2. Les « *Facebook Files* » (2021) [170], une enquête du Wall Street Journal basée sur des documents internes, ont exposé de nombreuses pratiques antisociales de l'entreprise, notamment :
    a. L'amplification algorithmique de contenus toxiques et de désinformation.
    b. L'aggravation des problèmes de santé mentale, particulièrement chez les adolescents.
    c. Le laxisme face au trafic de drogue, au trafic d'êtres humains et aux activités mafieuses sur ses plateformes.
3. Amnesty International a accusé Meta d'avoir contribué au génocide des Rohingyas en Birmanie [84] et à la violence ethnique en Éthiopie [171], en amplifiant massivement des appels à la violence sur ses plateformes.

Ces révélations suggèrent que Meta a souvent privilégié la croissance et les profits au détriment de la sécurité des utilisateurs [172]. Dans le contexte de l'IA, Meta, disposant d'immenses bases de données et de ressources informatiques, a un intérêt évident à limiter la régulation.

Yann LeCun, occupant un poste de direction stratégique durant ces scandales, a systématiquement défendu l'entreprise [172], [173], [174], soulevant des questions sur sa capacité à évaluer objectivement les risques éthiques de l'IA au sein de la Commission.

Il décrit lui-même son rôle comme étant « focalisé sur la direction scientifique, la stratégie IA, et l'évangélisation externe » [172]. Or, Meta développe activement des modèles d'IA à la frontière technologique, précisément le type de modèles dont le développement rapide est considéré comme risqué par de nombreux experts. De plus, Meta promeut l'approche open source pour ses modèles d'IA, une stratégie qui soulève des inquiétudes en termes de sécurité et de contrôle (voir section 3.3.2.3). LeCun est également cofondateur et conseiller d'Element Inc., une entreprise spécialisée dans la reconnaissance biométrique, ce qui ajoute une couche supplémentaire de conflit potentiel,

notamment sur les questions de protection de la vie privée et d'éthique liées à ces technologies.

Il est également important de noter que LeCun est connu pour ses opinions controversées sur les risques liés à l'IA et pour sa tendance à recourir à des attaques *ad hominem* envers ses opposants. Il a notamment qualifié de membres d'une « secte apocalyptique » [173] ceux qui ne partagent pas sa vision optimiste des dangers de l'IA, y compris ses co-lauréats du prix Turing, Yoshua Bengio et Geoffrey Hinton. Cette attitude soulève des questions quant à sa capacité à considérer objectivement les risques de l'IA dans le cadre de son rôle au sein de la Commission.

4.2.2 Arthur Mensch

- Cofondateur et PDG de Mistral AI.

Arthur Mensch est cofondateur et PDG de Mistral AI [174], une start-up française devenue en peu de temps un leader dans le développement de modèles d'intelligence artificielle générative, notamment open source. En juin 2024, Mistral AI était valorisée à 6 milliards d'euros [175], la positionnant comme une « pépite » française face aux géants technologiques américains.

Le conflit d'intérêts de Mensch au sein de la Commission de l'IA est évident :

- En tant que PDG, Mensch a un intérêt direct à minimiser la régulation de l'IA pour favoriser le développement rapide de sa start-up.
- Mistral AI, sous la direction de Mensch, a activement fait pression contre certaines dispositions de la loi européenne sur l'IA, notamment celles concernant la régulation des modèles de fondation (voir section 4.3.3).
- Lors d'une audition au Sénat, Mensch a tenu des propos trompeurs minimisant les risques liés à l'IA et remettant en question la nécessité d'une régulation stricte (voir section 4.3.3).
- Le récent accord avec Microsoft, un acteur majeur de l'IA, soulève des questions sur l'indépendance de Mistral et ses positions futures en matière de régulation.

La présence de Mensch dans la Commission, combinée à ses intérêts commerciaux directs et ses positions publiques, crée un risque significatif que les recommandations de la Commission soient influencées en faveur d'une régulation minimale de l'IA, potentiellement au détriment de considérations éthiques et de sécurité plus larges.

4.2.3 Cédric O

- Conseiller cofondateur de Mistral AI.
- Membre du conseil d'administration d'Artefact.

Cédric O, ancien secrétaire d'État au Numérique, présente des conflits d'intérêts significatifs :

1. Rôle chez Mistral AI : Conseiller, cofondateur et principal lobbyiste de Mistral AI. Selon Capital [176], O aurait investi dans Mistral AI via son agence de conseil, un investissement valorisé à 23 millions d'euros en décembre 2023.
2. Changement radical de position : Après avoir défendu des réglementations strictes en tant que secrétaire d'État, O plaide désormais pour une dérégulation de l'IA, alignant ses positions sur les intérêts de Mistral AI.
3. O a été au cœur d'un scandale de lobbying contre la loi européenne sur l'IA (voir section 4.3.1).
4. Membre du conseil d'administration d'Artefact, une société de conseil dédiée à l'accélération de l'adoption de l'IA, renforçant son intérêt pour une régulation minimale.

La gravité de ces conflits d'intérêts est soulignée par la décision de la Haute Autorité pour la Transparence de la Vie Publique, **qui a interdit à O, pour une durée de trois ans, de faire du lobbying auprès du gouvernement** ou de détenir des actions dans des entreprises technologiques [177].

4.2.4 Joëlle Barral

- Directrice principale de la recherche et de l'ingénierie chez Google DeepMind.
- Directrice de l'ingénierie chez Google.

Google DeepMind, filiale de Google, est l'un des leaders mondiaux dans le développement d'IA avancées, ayant notamment créé des modèles avancés emblématiques comme AlphaGo et AlphaFold.

Les positions de Barral soulèvent des questions de conflits d'intérêts potentiels. En tant que représentante de Google et DeepMind, elle pourrait être incitée à favoriser des politiques bénéficiant à ces entreprises. Étant donné les efforts de lobbying de Google contre la régulation de l'IA, sa présence dans la Commission pourrait influencer les recommandations dans un sens favorable à

l'industrie. Son double rôle pourrait ainsi compromettre son impartialité dans l'évaluation des risques et bénéfices de l'IA avancée.

4.3 Actions et opinions controversées

Certains membres de la Commission ont été impliqués dans des actions ou ont exprimé des opinions qui soulèvent des questions quant à leur impartialité et leur capacité à évaluer objectivement les enjeux liés à l'IA. Ces controverses, détaillées ci-dessous, mettent en lumière des positions potentiellement biaisées en faveur d'un développement peu régulé de l'IA, au détriment d'une approche plus prudente prenant en compte les risques.

Les actions et opinions controversées que nous avons retenues incluent :

1. **Lobbying de Cédric O :** L'ancien secrétaire d'État au Numérique, devenu conseiller de Mistral AI, a été accusé de conflit d'intérêts dans son opposition à la loi européenne sur l'IA.
2. **Lettre au Sénat Américain :** Yann LeCun et Arthur Mensch ont signé une lettre faisant la promotion de l'open source et contenant des affirmations scientifiquement erronées.
3. **Déclarations d'Arthur Mensch devant le Sénat :** Lors d'une audition, le cofondateur de Mistral AI a tenu des propos trompeurs sur le contrôle des modèles d'IA et la régulation du secteur.

4.3.1 Lobbying de Cédric O

4.3.1.1 Contexte

Cédric O, ancien secrétaire d'État au Numérique et membre de la Commission de l'IA, a joué un rôle clé dans l'élaboration de réglementations européennes rigoureuses comme le Digital Services Act et le Digital Markets Act. Après son mandat, il est devenu conseiller, cofondateur et principal lobbyiste de Mistral AI, une start-up française d'intelligence artificielle.

4.3.1.2 Actions controversées

Depuis son arrivée chez Mistral AI, O a radicalement changé de position, plaidant pour une dérégulation de l'intelligence artificielle. En octobre 2023, il affirme [178] que la loi européenne sur l'IA pourrait « tuer » son entreprise. Selon Médiapart, il aurait réussi à influencer le gouvernement pour qu'il adopte cette nouvelle approche [179], [180], [181]. Il est considéré comme l'architecte principal de l'opposition française à la loi européenne sur l'IA [182], [183], [184].

O a notamment organisé une lettre ouverte [185] signée par plus de 150 dirigeants d'entreprises, dont plusieurs membres de la Commission de l'IA, mettant en garde contre une réglementation trop stricte des modèles de fondation. Avec l'appui de l'Allemagne et de l'Italie, la France s'est opposée à toute règle contraignante pour les fournisseurs de modèles de fondation [186].

Catherine Morin-Desailly, vice-présidente de la Commission des Affaires Européennes, a déclaré : « O n'a pas déclaré Mistral auprès de la Haute Autorité pour la Transparence de la Vie Publique, qui l'avait déjà mis en garde de ne pas se faire embaucher par Atos ou toutes autres sociétés technologiques. » [184]

4.3.1.3 Réactions et critiques

Ce revirement a suscité de vives critiques. Le député Philippe Latombe a exprimé ses inquiétudes : « Le fait que la Haute Autorité ait demandé à Cédric O de déclarer toutes ses prises de participations dans le secteur de la tech soulève des questions sur son actionnariat au sein de Mistral et, par conséquent, au sein du Comité pour l'intelligence artificielle générative. » [184]

Le commissaire européen Thierry Breton a également critiqué cette position, déclarant que Cédric O défendait « tout sauf l'intérêt général » [187].

Face à ces accusations, O s'est défendu dans une réponse publique en affirmant : « Je n'ai pas changé de position depuis mes précédentes fonctions » [188]. Cependant, la Haute Autorité pour la Transparence de la Vie Publique lui a interdit [177], pour trois ans, de faire du lobbying auprès du gouvernement ou de détenir des actions dans des entreprises technologiques.

4.3.1.4 Implications

Ces controverses soulèvent des questions importantes sur l'impartialité de la Commission de l'IA. Le double rôle de Cédric O, à la fois membre de la Commission et lobbyiste pour une entreprise d'IA, ainsi que son investissement personnel dans Mistral AI (valorisé à 23 millions d'euros en décembre 2023 [176]), créent un conflit d'intérêts très prononcé.

4.3.2 Lettre au Président des Etats-Unis

4.3.2.1 Contexte

Yann LeCun et Arthur Mensch, membres de la Commission de l'IA, ont cosigné une lettre adressée au Président Américain Joe Biden au sujet de son *Executive Order* sur l'IA [189], faisant la promotion de l'open source dans le développement de l'IA.

4.3.2.2 Affirmation trompeuse

La lettre contient une déclaration scientifiquement erronée :

> « *Although advocates for AI safety guidelines often allude to the "black box" nature of AI models, where the logic behind their conclusions is not transparent, recent advancements in the AI sector have resolved this issue, thereby ensuring the integrity of open-source code models.* » [189].
>
> (« Bien que des partisans de recommandations en sécurité de l'IA fassent souvent allusion à la nature « boîte noire » des modèles en IA, où la logique derrière leurs conclusions n'est pas transparente, des avancées récentes dans le secteur de l'IA ont résolu ce problème, garantissant ainsi l'intégrité des modèles de code en open source. »)

Cette affirmation est également présente, pratiquement au mot près, dans une lettre adressée par le groupe Andreessen Horowitz (a16z), une société de capital-risque à la Chambre des Lords britannique [190] et dans une autre lettre de Martin Casado, partenaire de Andreessen Horowitz, adressée au Sénat américain [191]. Elle est en contradiction directe avec le consensus scientifique actuel [192], [193], [194]. Bien que notre capacité à analyser certains aspects des modèles d'IA s'améliore, il est incorrect de dire que la question de la nature de « boîte noire » de l'IA a été résolue, même partiellement. Les réseaux de neurones profonds, notamment les modèles de langage, restent opaques, même avec un accès libre à leurs paramètres. Ces modèles, comportant souvent des milliards de paramètres, présentent des interactions complexes non interprétables par les experts. Les outils d'interprétabilité, comme les cadres d'IA explicable, ont permis des avancées, mais nous n'en sommes qu'aux balbutiements de cette discipline. De plus, les récents progrès en IA, en rendant les modèles plus grands et plus complexes, exacerbent ce problème.

Cette déclaration erronée est utilisée pour soutenir l'idée que les modèles open source sont sans danger, une position qui favorise les intérêts des entreprises impliquées.

4.3.2.3 Réactions et critiques

De nombreux chercheurs et experts en interprétabilité ont vivement critiqué cette affirmation [195]. Voici quelques exemples :

- Neel Nanda (Google Deepmind) [196] : « *WTF?! This is massively against the scientific consensus.* » (« Pardon ?! C'est absolument contraire au consensus scientifique ! »)
- Max Kesin (Meta) : « *it's either deep stupidity/ignorance or more likely a blatant lie in service of their thesis* » (« c'est soit de la stupidité/de l'ignorance, soit plus probablement un mensonge éhonté pour servir leur argumentaire »)
- Darren McKee : « *No. And how odd.* » (« Faux. Et comme c'est bizarre [de dire ça]. »)

Martin Casado, l'auteur de la lettre, s'est excusé pour cette déclaration, admettant qu'elle était incorrecte [197]. John Carmack, un des signataires, a reconnu que l'affirmation était « clairement incorrecte » [198].

Cependant, Arthur Mensch et Yann LeCun n'ont pas présenté de réponse ou de rétractation.

4.3.2.4 Implications

Le fait que ces membres influents de la Commission aient signé une lettre adressée au président des Etats-Unis contenant des affirmations scientifiquement erronées, sans les corriger par la suite, remet en question :

- La rigueur scientifique des membres de la Commission de l'IA.
- Leur capacité à fournir des recommandations éclairées et impartiales.

4.3.3 Arthur Mensch devant le Sénat

4.3.3.1 Contexte

Le 14 juin 2024, Arthur Mensch, cofondateur et directeur général de Mistral AI, a été auditionné par la Commission des affaires économiques du Sénat, présidée par Dominique Estrosi Sassone. Cette audition visait à mieux

comprendre les implications de l'IA pour l'écosystème français et les enjeux de régulation.

4.3.3.2 Déclarations controversées

Au cours de cette audition, Mensch a fait plusieurs déclarations qui soulèvent des questions quant à leur exactitude scientifique et leur honnêteté intellectuelle [199] :

a. « On parle de logiciel, y a pas de changement, c'est un langage de programmation, personne ne se fait contrôler par son langage de programmation ».
b. « Il y a rien d'autonome, c'est un logiciel. Quand on écrit ce genre de logiciel, on contrôle toujours ce qu'il va passer et l'ensemble des sorties que le logiciel peut avoir. Il n'y a pas de changement de paradigme ».
c. « On donne les outils pour que les développeurs contrôlent bien les applications qu'ils déploient ».

4.3.3.3 Analyse critique

Ces déclarations sont trompeuses pour plusieurs raisons (voir aussi section 3.3.1) :

a. Assimilation erronée à un langage de programmation : Les modèles d'IA, en particulier les grands modèles de langage (*LLM*), ne sont pas des langages de programmation et ne fonctionnent pas comme des logiciels traditionnels. Ils génèrent des réponses basées sur des probabilités calculées à partir de vastes ensembles de données, ce qui les rend fondamentalement différents des logiciels classiques. Par exemple, les *LLM* sont capables de persuader des humains [200]. Les langages de programmation ne le sont pas.
b. Illusion de contrôle total : Contrairement à ce qu'affirme Mensch, les créateurs de modèles d'IA ne contrôlent pas directement chaque sortie possible. La complexité rend impossible la prédiction exacte de toutes les sorties sans exécuter le modèle [201]. De plus, étant donné que les sorties de ces modèles sont du langage naturel humain, leur contrôle automatique est extrêmement complexe [192], [193], [194].
c. Surestimation des outils de contrôle : Les méthodes proposées par Mistral [202] pour contrôler les applications, basées sur quelques lignes d'instructions ajoutées au prompt, ne fournissent aucune garantie solide [203]. Même les laboratoires les plus avancés peinent à assurer l'intégrité éthique de leurs modèles [204], [205], [206]. Nous recommandons notamment la partie 3.2 de « *Foundational Challenges in Assuring Alignment and Safety of Large Language Models* » [207].

4.3.3.4 Implications

Ces déclarations minimisent la complexité et les défis posés par les avancées récentes en IA. Elles présentent une vision simpliste qui pourrait induire en erreur les législateurs sur la nature réelle des technologies d'IA et les risques associés. Cette approche pourrait compromettre l'élaboration de réglementations appropriées et efficaces.

Le fait qu'un membre influent de la Commission de l'IA tienne de tels propos devant une instance législative majeure soulève des inquiétudes quant à la qualité des conseils fournis aux décideurs politiques et remet en question la capacité de la Commission à guider efficacement la politique française en matière d'IA.

4.4 Conclusion

L'analyse approfondie de la composition de la Commission de l'IA a mis en lumière plusieurs problèmes préoccupants :

1. Un déséquilibre marqué dans la représentation, avec une prédominance du secteur privé et une absence d'experts en sécurité de l'IA.
2. Des conflits d'intérêts significatifs chez des membres clés, notamment liés à des entreprises comme Meta et Mistral AI, qui ont un intérêt direct dans une régulation minimale de l'IA.
3. Des actions et déclarations controversées de certains membres, soulevant des questions sur leur objectivité et leur rigueur scientifique.

Ces défaillances dans la composition de la Commission ont des implications profondes. Elles remettent en question la capacité de cet organe à produire des recommandations véritablement équilibrées et objectives, prenant en compte l'ensemble des enjeux liés à l'IA, y compris les risques nécessitant, si avérés, des réactions de la part de certaines entreprises.

La crédibilité du rapport final de la Commission s'en trouve sérieusement compromise. Les lacunes identifiées dans sa composition expliquent en grande partie les biais et les omissions observés dans le contenu du rapport, notamment la minimisation systématique des risques et l'absence de prise en compte des enjeux de sécurité de l'IA.

Cette situation souligne la nécessité et l'urgence de repenser la manière dont est constituée une Commission chargée de guider les politiques publiques sur des sujets aussi cruciaux et complexes que l'IA. Sans une réforme significative

de ce processus, il est à craindre que les recommandations produites ne servent pas pleinement l'intérêt général et ne préparent pas adéquatement la France aux défis futurs de l'IA.

5 Recommandations

Notre analyse a mis en évidence des lacunes graves dans le rapport de la Commission de l'IA, compromettant potentiellement la stratégie de la France en matière d'intelligence artificielle. Ces défaillances semblent découler d'un manque de diversité d'expertise et de conflits d'intérêts au sein de la Commission. Pour remédier à cette situation et assurer une approche plus équilibrée et rigoureuse, nous proposons les recommandations suivantes :

1. Remaniement de la Commission de l'IA

Éliminer les conflits d'intérêts et diversifier l'expertise au sein de la Commission. Écarter les membres ayant des conflits d'intérêts majeurs pour les repositionner comme consultants. Intégrer davantage de représentants de l'académie, de la société civile, de l'éthique et de la sécurité de l'IA.

2. Consultation d'experts en sécurité de l'IA

Solliciter l'avis de spécialistes reconnus en sécurité de l'IA pour évaluer rigoureusement les risques liés au développement de systèmes d'IA avancés. Leur expertise est cruciale pour identifier les vulnérabilités et proposer des mesures de protection robustes. Il existe de tels spécialistes en France. Nous recommandons le Centre pour la Sécurité de l'IA (CeSIA), que nous avons consulté pour corriger des parties de ce document, SaferAI, ou encore le lauréat du prix Turing québécois Yoshua Bengio.

3. Rédaction d'un addendum au rapport

Suite à la consultation d'experts, produire un addendum traitant spécifiquement des risques majeurs négligés dans le rapport initial. Cet ajout est essentiel pour offrir une vision complète des enjeux liés à l'IA et assurer la prise en compte de tous les aspects critiques dans l'élaboration des politiques futures.

4. Organisation d'un débat public sur l'avenir de l'IA, initié par cette contre-expertise

Lancer une convention citoyenne pour inaugurer la nécessaire délibération démocratique sur les bénéfices potentiels et les risques de l'IA. L'objectif est d'évaluer l'acceptabilité sociale et les modalités du

développement de l'IA et d'enrichir la réflexion collective en prenant en compte les préoccupations de la société civile.

La mise en œuvre de ces recommandations devrait permettre d'établir une stratégie nationale en matière d'IA qui soit à la fois ambitieuse et responsable, tenant compte des opportunités mais aussi des risques de cette technologie transformative, en remédiant aux manquements du rapport actuel.

6 Conclusion

L'analyse approfondie du rapport de la Commission de l'IA révèle des lacunes qui compromettent sa crédibilité et son utilité en tant que document d'orientation stratégique pour la France. Les omissions critiques concernant les risques existentiels et la sécurité de l'IA, la minimisation systématique des risques, et le manque d'anticipation des développements futurs témoignent d'une approche dangereusement biaisée et myope.

Ces défaillances trouvent leur origine dans la composition même de la Commission, marquée par des conflits d'intérêts majeurs et un manque criant de diversité d'expertise, notamment l'absence totale de spécialistes en sécurité de l'IA. Cette structure déséquilibrée a conduit à un rapport qui privilégie manifestement les intérêts de l'industrie au détriment d'une évaluation objective et complète des enjeux.

Les implications de ces manquements sont potentiellement catastrophiques pour la stratégie française en matière d'IA. En sous-estimant systématiquement les risques et en négligeant les scénarios futurs plausibles, le rapport ouvre la voie à des réglementations inadéquates et à une préparation insuffisante face aux défis imminents posés par l'IA avancée.

Il est impératif d'adopter une approche plus équilibrée et rigoureuse, intégrant une diversité de perspectives et accordant une attention particulière aux questions de sécurité et d'éthique. Seule une évaluation objective et exhaustive des risques et des opportunités permettra d'élaborer une stratégie nationale à la hauteur des enjeux.

Face à cette situation critique, nous appelons à :

1. Un remaniement immédiat de la Commission, éliminant les conflits d'intérêts et intégrant une véritable diversité d'expertise.
2. La consultation d'experts en sécurité de l'IA.
3. La rédaction d'un addendum au rapport, traitant spécifiquement des risques majeurs précédemment négligés.
4. L'organisation d'un débat public national sur l'avenir de l'IA en France.

Notre contre-expertise se veut le début d'un véritable débat public sur l'IA en France, répondant ainsi à l'appel de la Commission tout en allant au-delà de ses propositions initiales. Nous invitons les experts, les décideurs et les

citoyens à s'engager dans ce dialogue crucial pour façonner l'avenir de l'IA dans notre pays.

L'enjeu dépasse largement le cadre national : il s'agit de l'avenir de notre société et, potentiellement, de l'humanité elle-même. Le sommet sur l'IA prévu en 2025 offre à la France une opportunité unique de prendre une position de leadership sur ces questions de premier plan. Il est temps pour notre pays de prendre la pleine mesure des défis et des dangers que pose cette technologie révolutionnaire, et d'agir en conséquence avec lucidité, responsabilité et ambition.

Validation des experts

Cette section présente les experts qui ont validé notre analyse, ainsi que leur niveau de soutien. **Les recommandations n'engagent que les auteurs principaux et n'ont pas fait l'objet d'une validation par les experts.**

Définition d'expert et domaines d'expertise :

Dans le cadre de cette contre-expertise, nous considérons comme « expert » toute personne possédant une expertise reconnue dans un ou plusieurs domaines pertinents pour l'analyse de l'IA et de ses impacts. Ces domaines incluent, sans s'y limiter : l'intelligence artificielle, l'éthique de l'IA, la sécurité de l'IA, la cybersécurité, l'informatique, l'apprentissage automatique, la politique publique, l'économie de l'innovation, le droit du numérique, la sociologie du travail, la psychologie cognitive, la philosophie des sciences, la bioéthique, l'environnement, les relations internationales, la santé publique, les sciences de l'information et la gestion des risques.

Les experts ont eu la possibilité de valider spécifiquement les sections relevant de leur domaine d'expertise. Cette approche permet une validation ciblée et précise, reflétant la nature interdisciplinaire des enjeux liés à l'IA.

Niveaux de soutien :

1. **Validation :** L'expert confirme l'exactitude et la pertinence de l'analyse présentée pour l'ensemble du document ou pour des sections spécifiques.

2. **Soutien :** L'expert approuve les principales conclusions de l'analyse, tout en pouvant avoir des réserves sur certains points spécifiques, pour l'ensemble du document ou pour des sections spécifiques.

Liste des experts et leur niveau de soutien :

ALBERT Patrick
Fondateur de Hub France IA
Validation

« Il nous faut remercier Pause IA pour son analyse rigoureuse des conclusions de la Commission IA mandatée par E. Macron. Les conflits d'intérêts manifestes dans sa composition sont confirmés par les parti pris qui minent la crédibilité des recommandations, et évacuent la plus nécessaire : la convocation urgente

d'un authentique débat démocratique sur une industrie qui ne propose rien de moins que de déclencher une mutation anthropologique majeure. »

COUILLET Romain
Professeur des Universités à l'Université Grenoble-Alpes (membre du Laboratoire d'Informatique de Grenoble)
Mathématiques appliquées, intelligence artificielle, analyse systémique du numérique
Validation

« En sus de l'analyse de la contre-expertise, il me paraît urgent de rappeler que l'IA ouvre à la possibilité déjà actée sur le terrain (!) de produire des armes de guerres incontrôlables (cibles filtrées par IA, essaims de drones autonomes) et que l'IA renforce le verrou de dépendance à un numérique en proie aux pénuries à venir (métaux, pétrole). »

PERRET Jérémy
Suboptimal IA
Docteur et ingénieur en Intelligence artificielle
Validation

« Très regrettable qu'un tel groupe d'experts néglige à ce point les enjeux de sécurité de l'IA. En remettant à un futur organisme d'évaluation la tâche d'« anticiper l'apparition de nouveaux risques », le rapport ignore l'ensemble des travaux déjà réalisés sur le sujet. Ceci trahit l'approche responsable à laquelle prétend la Commission. »

BILLION Arnaud
Docteur en droit, directeur de recherche chez Themis 5.0
Spécialiste en droit de l'informatique, propriété intellectuelle, et en informatique juridique
Validation

« Il ne faut pas sous-estimer le problème de la dette technique logicielle. IA signifie transformation et création de données non structurées ni signifiantes. Cette data devra être calculée, en intrant des nouveaux modèles d'IA, menant à des problèmes considérables de montée en charge, d'administration et de stockage. Il faut pouvoir absorber toute cette nouvelle dette logicielle (donc énergétique). »

AMARSY Stéphane
PDG The Next Mind

Intelligence artificielle
Soutien

« Notre avenir mérite mieux qu'une étude d'opportunité parcellaire. Les entreprises et chaque citoyen doivent pouvoir avoir accès à des IA servant le collectif. Je milite pour une Commission ouverte à tous les acteurs. »

GIBERT Martin
Centre de Recherche en Éthique, à l'Université de Montréal
Philosophe et chercheur en éthique de l'IA
Soutien

« Je n'ai vu dans le rapport de la Commission de l'IA publié le 13 mars 2024 qu'un reflet très partiel des préoccupations des chercheurs et des chercheuses en éthique de l'IA. Les risques existentiels et les enjeux de cybersécurité en particulier y sont traités de façon cavalière. Quant à la question fondamentale de l'éthique des systèmes de recommandations, elle est absente. Ce n'est pas comme ça qu'on assure l'avenir d'une communauté. »

DE LA HIGUERA Colin
Professeur en Informatique à Nantes Université et Titulaire de la Chaire Unesco en Ressources Educatives Libres et Intelligence Artificielle
Apprentissage Automatique, Machine Learning, IA et Education
Soutien + Validation de la section 3.2.2

« Je soutiens l'analyse qui me paraît à la fois juste et importante. Elle est juste parce qu'effectivement, les risques existentiels sont très sous-estimés, voire ignorés dans le rapport. Or ceux-ci donnent lieu à des mises en garde de personnalités éminentes. À titre d'exemple, deux des trois lauréats du prix Turing 2018 pour les travaux fondamentaux sur l'apprentissage profond ont mis en garde. Le troisième est membre de la Commission. Elle est importante parce qu'elle émane de la jeunesse qui aura à gérer les erreurs que nous pourrions commettre demain. »

BOULLIER Dominique
Professeur des universités émérite à Sciences Po Paris
Sociologue et linguiste, spécialiste des technologies cognitives et des stratégies des plateformes numériques (réseaux sociaux notamment)
Soutien

SEGERIE Charbel-Raphaël
Directeur Exécutif du Centre pour la Sécurité de l'IA, Enseignant au Master MVA de l'ENS Paris-Saclay
General Purpose AI System, AI Safety, interpretability, alignment techniques
Soutien

DERIAN Maxime
Fondateur de Heruka-AI Consulting, membre du collectif ANR CulturIA (CNRS, Université de la Sorbonne), collaborateur avec Everyone.AI (Californie, USA)
Innovation en IA, éthique et régulation, technoréalisme
Soutien

BRETTEL Alexandre
Doctorant contractuel en éthique de l'intelligence artificielle à l'Université Grenoble Alpes Responsabilité de l'IA, philosophie de la technique, éthique de l'innovation
Soutien section 3

« Par rapport à la responsabilité, celle-ci est quelques fois évoquée dans le rapport, mais elle n'est que très superficiellement traitée. Il conviendrait par exemple de leur conseiller de se référer aux travaux relatifs à la responsabilité en éthique et en droit de l'IA. L'éthique est une discipline distincte du droit et la réglementation ne saurait servir de seule méthode de régulation. »

DUMAS Clément
École Normale Supérieure de Paris-Saclay
Recherche en Mechanistic Interpretability
Soutien

« Sans nécessairement prôner une pause dans le développement de l'IA, j'estime qu'il est crucial de reconnaître et d'affronter les défis complexes liés à la création d'une IA puissante, éthique et sûre. La France ne peut se permettre d'ignorer ces enjeux fondamentaux pour son avenir technologique et sociétal. »

NGUYEN HOANG Lê
PDG de Calicarpa, Président de Tournesol et Membre du Conseil d'Éthique d'Orange
Chercheur en sécurité et en gouvernance des systèmes d'intelligence artificielle
Validation section 3.2.4

FOURQUET Jean-Lou

Co-auteur de « La dictature des algorithmes », membre de l'association Tournesol, fondateur de la chaîne ApresLaBiere, ancien chroniqueur pour Arrêt sur Images

Numérique, vulgarisation, IA, transition énergétique

Validation section 3.2.4

Soutien des associations

Cette section présente les associations qui soutiennent notre contre-expertise. Le soutien d'une association n'implique pas nécessairement l'accord de tous ses membres avec chaque détail du rapport, mais indique un alignement général avec ses conclusions principales.

Types de soutien :

- **Soutien global :** L'association approuve l'ensemble de l'analyse et des conclusions principales du rapport.
- **Soutien partiel :** L'association soutient certaines parties spécifiques du rapport, qui sont explicitement mentionnées.

Liste des associations soutenant le rapport :

Centre pour la Sécurité de l'IA - *Soutien global*

Technoréalisme - *Soutien global*

Technologos - *Soutien global*

Tournesol - *Soutien à la section 3.2.4*

A.R.T.S. - *Soutien Global*

À propos de Pause IA

Pause IA est une association dédiée à la promotion d'un développement responsable et éthique de l'intelligence artificielle. Branche française du mouvement international *Pause AI,* Pause IA réunit une communauté de volontaires visant à réduire les risques de dommages catastrophiques liés au développement de l'intelligence artificielle. Nous cherchons à convaincre les gouvernements et les citoyens d'intervenir pour instaurer une pause indéfinie dans le développement des modèles les plus dangereux jusqu'à ce que des solutions techniques et sociétales soient trouvées. Pause IA n'est pas pour autant anti-technologie, ni même anti-IA, en particulier lorsqu'il s'agit d'IA spécialisée et dans la mesure où l'on peut garantir sa sécurité.

Cette contre-expertise a été réalisée sous l'égide de Pause IA, reflétant l'engagement de l'association pour un débat public éclairé sur les enjeux de l'IA.

Plus d'informations sur Pause IA : https://www.pauseia.fr

À propos des auteurs

Cette contre-expertise est le fruit d'un travail collectif bénévole, réalisé par un groupe diversifié d'experts et de citoyens engagés. Nos contributeurs partagent une préoccupation commune pour les enjeux de l'IA et un désir de promouvoir un débat public éclairé sur ce sujet crucial.

Auteurs principaux :

Maxime Fournes
Président, Pause IA
Expert en Machine Learning et Deep Learning

Pierre Lamotte
Archéologue et forgeron

Éloïse Benito-Rodriguez
Chercheuse indépendante en sécurité de l'IA

Gilles Bréda
Cofondateur, Pause IA
Musicien professionnel

Contributeurs :

Amaury Lorin

Aurélia Jauffret

Mathieu Bourrier

Site web et mise en page :

Moïri Gamboni

Illustrations :

Elizabeth Richards

Direction du projet :

Maxime Fournes

Nous tenons à souligner que tous les auteurs et contributeurs ont participé à ce projet sur leur temps personnel, sans aucune rémunération. Ce travail représente un effort collectif d'environ 400 heures.

Conflits d'intérêts :
Les auteurs déclarent n'avoir aucun conflit d'intérêt en relation avec le sujet de cette contre-expertise. Aucun financement externe n'a été reçu pour la réalisation de ce travail.

Pour toute question ou commentaire concernant ce document, veuillez contacter : Maxime Fournes maxime@pauseia.fr ou Gilles Bréda gilles@pauseia.fr.

Annexes

A — Graphique Composition Commission

Se réfère au graphique section 4.1

Dans la catégorie « Industrie », nous avons inclus les entrepreneurs et les chercheurs dans le domaine privé. Cela comprend :

Gilles Babinet (entrepreneur), Anne Bouverot (entrepreneuse), Bernard Charlès (PDG de Dassault Systèmes), Cédric O (lobbyiste), Arthur Mensch (entrepreneur et chercheur en IA dans le privé), Yann Lecun (chercheur en IA dans le privé), Joëlle Barral (chercheuse en IA dans le privé), Luc Julia (chercheur en IA dans le privé). Nozha Boujemaa (chercheuse en IA dans le privé).

Dans la catégorie « Académie », nous avons inclus les chercheurs exclusivement dans le domaine public. Cela comprend : Isabelle Ryl (chercheuse en IA), Gaël Varoquaux (chercheur en IA), Philippe Aghion (économiste), Alexandra Bensamoun (chercheuse en droit).

Dans la catégorie « Société civile », nous avons inclus les représentants d'organisations non gouvernementales, de syndicats, d'organisations caritatives, et de fondations privées.. Cela comprend : Franca Salis-Madinier (secrétaire nationale CFDT Cadres), et Martin Tisné ((co)fondateur de nombreuses organisations, notamment sur la transparence fiscale et le droit des données).

L'Éthique de l'IA examine les implications morales et sociétales de l'IA, en établissant des principes pour son développement et son utilisation responsables. Nous avons inclus dans cette catégorie :
Martin Tisné (PDG de AI Collaborative), Anne Bouverot (cofondatrice de la Fondation Abeona), Franca Salis-Madinier (étudie l'impact de l'IA sur le travail).

La sécurité de l'IA vise à développer des systèmes d'intelligence artificielle sûrs et fiables, en minimisant les risques de dommages intentionnels ou non. Aucun membre de la Commission ne travaille dans ce domaine.

B — Postes des membres de la Commission

Anne Bouverot, co-présidente de la Commission

- Présidente du conseil d'administration de Cellnex Telecom, le premier opérateur européen d'infrastructures de télécommunications sans fil.
- Membre du conseil d'administration de Thomson Reuters, agence de presse canado-britannique et société d'édition professionnelle, financière et juridique.
- Conseillère principale à TowerBrook Capital Partners L.P., un fond d'investissement.
- Membre indépendante du conseil d'administration de Ledger, licorne à croissance rapide qui développe des solutions de sécurité et d'infrastructure pour les crypto-monnaies et les applications blockchain.
- Présidente du Conseil d'administration de Technicolor Creative Studios de juin 2019 à février 2024, un des principaux fournisseurs d'effets visuels et de services d'animation.
- Présidente du conseil d'administration de l'ENS.

Philippe Aghion, co-président de la Commission

- Professeur au collège de France.
- Professeur à l'INSEAD, école privée de management.
- Conseiller spécial transformation au Secrétariat général pour l'investissement.
- Membre du Conseil d'administration à l'Institut national du service public.

Luc Julia

- Directeur scientifique chez Renault.
- Cofondateur de ODIA, entreprise de synthèse vocale par intelligence artificielle.

Gilles Babinet

- Entrepreneur, propose des services de consulting aux entreprises souhaitant engager un processus de transformation numérique.
- Professeur à HEC.
- Co-président du Conseil National du Numérique.
- Digital Champion pour la France à la Commission Européenne.
- Professeur Associé à Sciences Po.

Bernard Charles

- PDG de Dassault Systèmes, éditeur de logiciels.
- Membre du conseil d'administration de Sanofi, entreprise pharmaceutique.

Franca Salis-Madinier

- Secrétaire nationale CFDT Cadres.
- Vice-présidente « workers group » au Comité Économique et Social Européen.
- Salariée d'Orange.

Christophe Ravier

- Directeur R&D BU Agro chez AKANEA, éditeur de logiciel.

Nozha Boujemaa

- Vice-présidente global — Innovation en matière d'IA et confiance chez Decathlon.
- Directrice des données industrielles à l'Adra — AI-Data-Robotics-Association, partenaire privé du Partenariat européen ADR, qui cherche à stimuler la compétitivité européenne et le développement des synergies entre l'IA, la data et la robotique.
- Coprésidente du groupe d'experts OECD.AI, le réseau d'experts de l'OCDE sur l'IA.

Gaël Varoquaux

- Cofondateur de Therapixel, éditeur de logiciels d'IA appliquée à l'imagerie médicale.
- Cofondateur et conseiller scientifique de « :probabl », développeur de logiciels.
- Directeur de recherche à Inria.
- Cofondateur de scikit-learn, logiciel de modèle d'IA.

Philippe Chantepie

- Inspecteur général au ministère de la Culture et de la Communication.
- Chercheur associé à la Chaire Innovation & Régulation de l'École Polytechnique / Telecom Paris Tech / Orange.

Isabelle Ryl

- Vice-présidente pour l'IA — Directrice du PRAIRIE (Institut de recherche en IA de Paris).
- Membre du Conseil d'administration de l'École des Ponts ParisTech.
- Membre du conseil d'administration — Trésorière a Agoranov, incubateur de startup.

Alexandra Bensamoun

- Chercheuse associée à l'université Laval.
- Professeure de droit à l'université Paris-Saclay.
- Experte internationale IP/IT sur la diversité des expressions culturelles à l'UNESCO.
- Membre du comité exécutif à l'Institut DATAIA, un institut de recherche pluridisciplinaire en IA.

Martin Tisné

- PDG de AI Collaborative, organisation cherchant à réglementer l'intelligence artificielle sur des valeurs démocratiques.
- Membre du conseil d'administration de Partnership on AI, coalition œuvrant à une utilisation responsable de l'IA.

Yann LeCun
Voir section 4.2.1

Cédric O
Voir section 4.2.3

Arthur Mensch
Voir section 4.2.2

Joëlle Barral
Voir section 4.2.4

C — L'apprentissage profond et ses origines

Afin de corriger l'image présentée par le rapport d'une technologie mature, dont le fonctionnement et les implications sont compris depuis longtemps, et

qui se place dans la stricte continuité des paradigmes précédents, rappelons l'histoire de l'IA moderne[1].

Un changement de paradigme commence à s'opérer dans l'histoire de l'IA au milieu des années 1990 avec l'adoption du paradigme connexionniste[2] [208, p. 25], [209, p. 424]. Les approches symboliques en IA, qui dominaient le champ jusqu'alors, sont remplacées par des méthodes permettant aux systèmes de s'adapter à la complexité des données issues du monde réel sans nécessiter la programmation de règles préétablies [208, p. 24], [210].

Ce paradigme comprend différentes architectures de réseaux neuronaux, les réseaux bayésiens, et d'autres méthodes statistiques comme les *HMM*[3]. Ils sont capables d'analyser de vastes ensembles de données pour identifier des régularités, mais dépendent encore de l'ajustement manuel de leurs paramètres et de la qualité des données. Souvent de taille modeste, ils sont réputés pour leur transparence et leur interprétabilité, permettant une compréhension claire des processus décisionnels qu'ils emploient [211].

Les réseaux neuronaux apparaissent encore jusqu'à la fin des années 2000 comme une voie parmi d'autres, même s'ils prennent une importance grandissante dans la recherche en IA [208], [209]. C'est véritablement avec l'apprentissage profond que ce paradigme prend son plein essor, montrant une efficacité remarquable. La démonstration d'AlexNet en particulier, un modèle de reconnaissance d'images qui, en 2012, remporte la compétition internationale ImageNet, fait date [208, p. 25], [210].

Les réussites de l'apprentissage profond se multiplient par la suite, avec la création de modèles spécialisés dans toutes sortes de domaines, notamment dans des jeux de plus en plus complexes, ce qui est devenu apparent avec la victoire écrasante d'AlphaGo, un modèle de Google Deepmind, contre Lee Sedol, un des meilleurs joueurs mondiaux de go en 2016 [212]. En 2017, il bat le champion du monde Ke Jie [213]. On parle désormais de « révolution du *deep learning* » [208, p. 28], [214].

Ce sont cependant les avancées dans la génération de texte par des grands modèles de langage (*LLM*) qui constituent le tournant majeur des modèles généralistes comme ChatGPT.

---

[1] Notons d'emblée que cette histoire n'est pas le fruit d'un travail d'historiens professionnels mais d'acteurs du domaine, ayant vécu ces développements « de l'intérieur », depuis leur position particulière. Ainsi par exemple, les auteurs placent parfois les articulations importantes de cette histoire technique à des moments différents.

[2] Les innovations techniques à l'origine du connexionnisme apparaissent dans les années 1970 (ex. *backpropagation*) et 1980 (ex. *CNN*s), mais on ne peut pas parler de changement de paradigme à proprement parler avant le milieu des années 1990 [208], [209].

[3] *Hidden Markov Models* ou Modèle de Markov Caché.

D — Les *LLM*

Les *LLM* (*Large Language Model* — Grand modèle de langage) sont entraînés à prédire l'élément suivant dans un texte en fonction d'un contexte, utilisant des estimations probabilistes pour générer du contenu cohérent [215]. Cet entraînement produit une matrice vectorielle de très grande taille, contenant toutes les relations observées entre les *tokens* (ou lexèmes, qui sont des séries de caractères) sur lesquels le modèle est entraîné. Ces *tokens*, générés par un autre modèle d'intelligence artificielle, peuvent être des mots, des portions de mots ou des phrases [216].

Les modèles d'apprentissage profond sont d'autant plus efficaces qu'ils sont massifs, ce qui augmente également leur complexité et leur opacité. Les algorithmes, s'ils existent, menant d'une entrée (un *prompt*) à un résultat (un texte) sont invisibles aux « programmeurs » de l'IA, tout comme ils le sont pour l'IA elle-même. En réalité, un modèle d'apprentissage profond n'est pas « programmé » au sens traditionnel en informatique : ce que l'on programme, c'est sa structure et son algorithme d'apprentissage. Le modèle est ensuite entraîné et découvre des moyens de traiter l'information reçue par lui-même, sous l'effet des contraintes et des données fournies. En ce sens, ces modèles sont moins « programmés » que « cultivés »[4] [217], [218].

Cette opacité est telle qu'elle a conduit à la création d'un nouveau champ des sciences informatiques, l'interprétabilité, destiné à sonder et comprendre le contenu des modèles [219].

Les *LLM* sont particuliers car ils sont entraînés sur le langage humain, qui contient de nombreuses connaissances et modélisations du monde. Cela leur permet de développer des capacités de compréhension générale bien plus étendues que les modèles spécialisés en classification d'images ou reconnaissance d'objets [215]. Cette compréhension du monde, et le fait qu'il soit possible de spécialiser des *LLM* par « réglage fin » (*fine tuning*) sur des données plus restreintes, conduit à les utiliser comme « modèles de fondation » à la base de toutes sortes d'applications spécialisées [220].

Les *LLM* ont montré des capacités émergentes remarquables au fil des générations, comme la programmation de code informatique, la compréhension et la production de textes longs, la traduction multilingue, et l'explication de concepts complexes dans divers domaines. Cependant, certains

[4] On parle généralement « d'entraînement » pour désigner ce processus. Toutefois, la métaphore organique de modèles « cultivés » a l'avantage de mettre en exergue le peu de contrôle qu'ont les développeurs sur le contenu et les capacités des modèles qu'ils entraînent.

comportements potentiellement dangereux ont été observés, comme le mensonge, la manipulation, et le piratage informatique [215].

Ces capacités sont souvent découvertes après la mise à disposition des modèles au grand public. En plus de leurs capacités intrinsèques, il est possible d'améliorer l'efficacité des modèles de fondation grâce au *scaffolding* ou « enrobage », qui consiste à les intégrer dans des environnements informatiques permettant l'exécution de fonctions, comme faire appel à un autre modèle, ou à les faire réfléchir en suivant des étapes précises [221]. Les améliorations possibles sont difficiles à prédire.

Les IA actuelles sont en rupture totale avec les paradigmes précédents : opaques, « cultivées » plus que programmées, étonnantes dans leurs réussites comme dans leurs erreurs, elles se rapprochent plus que jamais de l'objectif du champ de recherche en IA, qui est de créer des machines qui pensent. Le paradigme, et les capacités qui l'accompagnent, sont tout jeunes – à peine plus d'une décennie pour l'apprentissage profond, la moitié d'une décennie pour les *LLM*. Les capacités de ces modèles de langage s'étendent rapidement, et si les expériences du passé éclairent peu sur les capacités futures, rien pour le moment ne vient contredire cette tendance. La trajectoire du développement de ces modèles nous mène vers des risques d'augmentation brusque et incontrôlée de leurs capacités.

E — La trajectoire actuelle de l'IA

Le premier facteur influençant la trajectoire actuelle de l'IA est la possibilité d'anticiper l'amélioration des modèles grâce aux « *scaling laws* », ou lois d'échelle, qui stipulent que l'efficacité des modèles de langage augmente logarithmiquement en fonction de leur taille, c'est-à-dire de la quantité de données et de calcul utilisée pour leur entraînement [222]. Depuis leur découverte récente, ces lois de scalabilité ont fait l'objet de nombreuses critiques, plusieurs chercheurs prédisant leur obsolescence à courte échéance. Malgré ces mises à l'épreuve, l'augmentation spectaculaire de la taille des modèles, qui s'est étendue sur plusieurs ordres de grandeur, a plusieurs fois réaffirmé leur validité et leur pertinence. Les entreprises développant ces modèles parient en partie sur leur continuité [223], cherchant à obtenir plus de données et de capacités de calcul, par exemple en générant artificiellement des données avec d'autres IA [224] et en construisant de nombreux « *GPU clusters* », à l'instar de Microsoft, qui projette même la construction d'une centrale nucléaire pour leur alimentation en électricité [225].

Des efforts considérables sont consacrés à la création de nouvelles puces et au développement de processeurs plus puissants. Par exemple, Sam Altman,

PDG d'OpenAI, a récemment annoncé un projet de plusieurs milliers de milliards de dollars avec les Émirats Arabes Unis pour la construction de nouvelles puces destinées à l'entraînement des modèles [226].

Cette course à la puissance de calcul est désormais internationale et prend une tournure géopolitique (p. ex. [227]). Les capacités de calcul actuelles sont déjà potentiellement supérieures à celles strictement nécessaires à la création de modèles aussi capables, voire plus, qu'un être humain dans la majorité des tâches. Ce phénomène, appelé « *compute overhang* », désigne une situation où une amélioration des modèles pourrait survenir de manière brusque avec la découverte d'algorithmes plus efficaces, multipliant soudainement la valeur des capacités de calcul déjà existantes.

Les jeux de données utilisés pour l'entraînement connaissent également des transformations profondes. La qualité des données joue un rôle crucial dans la performance du modèle, et des efforts importants sont faits pour leur collecte [228] et leur curation. Une innovation particulièrement notable des années 2023 et 2024 est la multimodalité des modèles, entraînés sur des images, des vidéos et du son en plus du texte [229], [230]. Les premiers modèles nativement multimodaux sont en développement, et ils pourraient combiner l'efficacité des modèles spécialisés avec la généralité des *LLM*.

Le *scaffolding* (ou « enrobage », voir *annexe D*) des modèles multimodaux sera probablement un élément clé des systèmes destinés à améliorer leurs compétences. Enfin, les paradigmes précédents de l'histoire l'IA ne sont pas oubliés et un effort important est fourni pour tenter de les appliquer aux modèles actuels, produisant des phénomènes de « pollinisation croisée » dont on peut craindre qu'ils débouchent rapidement sur la création de superintelligences (p. ex. [231]).

Une prochaine étape de l'histoire de l'IA aujourd'hui largement discutée est celle des agents autonomes (p. ex. [232]). On peut s'attendre à ce que l'essentiel de la valeur ajoutée de l'intelligence artificielle tienne dans sa capacité à exécuter des tâches complexes à la place des humains, simultanément, efficacement, rapidement et à grande échelle. Pour ce faire, les IA doivent pouvoir choisir un plan d'action, définir des étapes et sous-tâches, et mettre en œuvre ce plan de manière autonome. Plus les modèles sont capables et autonomes, plus ils peuvent agir dans le monde. Plus ils peuvent agir dans le monde, plus ils sont utiles.

Bibliographie

[1] « Comité de l'intelligence artificielle générative », info.gouv.fr. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.info.gouv.fr/communique/comite-de-lintelligence-artificielle

[2] « Le rapport IA : notre ambition pour la France ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.bercynumerique.finances.gouv.fr/le-rapport-ia-notre-ambition-pour-la-france

[3] « Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya », Amnesty International. Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://www.amnesty.org/en/documents/ASA16/5933/2022/en/

[4] Amnesty International UK, « Ethiopia: Facebook algorithms contributed to human rights abuses against Tigrayans during conflict - New Report », Amnesty International UK. Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://www.amnesty.org.uk/press-releases/ethiopia-facebook-algorithms-contributed-human-rights-abuses-against-tigrayans

[5] R. Manuvie, I. Mony, A. Kahle, et M. N. Khan, « Preachers of Hate: Documenting Hate Speech on Facebook India ». Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://research.rug.nl/en/publications/preachers-of-hate-documenting-hate-speech-on-facebook-india

[6] Center for AI Safety, « Statement on AI risk », CAIS | Center for AI Safety. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.safe.ai/work/statement-on-ai-risk

[7] « Pause giant AI experiments: An open letter », Future of Life Institute. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[8] K. Bryant, « How AI Is Impacting Society And Shaping The Future », *Forbes Magazine*, 13 décembre 2023. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.forbes.com/sites/kalinabryant/2023/12/13/how-ai-is-impacting-society-and-shaping-the-future/

[9] A. Jindal et A. Sharma, « THE IMPACT OF ARTIFICIAL INTELLIGENCE ON SOCIETY », 18 mars 2024. doi: 10.13140/RG.2.2.18522.96968.

[10] « Mission d'appui du CGE à la Commission de l'intelligence artificielle ». Consulté le: 3 septembre 2024. [En ligne]. Disponible sur: https://www.economie.gouv.fr/cge/commission-ia

[11] « Le rapport IA : notre ambition pour la France ». Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://www.bercynumerique.finances.gouv.fr/le-rapport-ia-notre-ambition-pour-la-france

[12] Commission de l'intelligence artificielle, « IA : Notre ambition pour la France », mars 2024. Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://www.info.gouv.fr/upload/media/content/0001/09/4d3cc456dd2f5b9d79ee75feea63b47f10d75158.pdf&sa=D&source=docs&ust=1725897993856594&usg=AOvVaw3e6zrU8MYf2OzNepwSzbUH

[13] C. Blanchot, « « Il est inconcevable que le déploiement de l'intelligence

artificielle se fasse sans débat public et sans évaluation de son impact sur notre travail » », *Le Monde*, Le Monde, 16 mai 2024. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.lemonde.fr/idees/article/2024/05/16/il-est-inconcevable-que-le-deploiement-de-l-intelligence-artificielle-se-fasse-sans-debat-public-et-sans-evaluation-de-son-impact-sur-notre-travail_6233566_3232.html

[14] « Pour une IA française tournée vers l'avenir ». Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.securite-ia.fr/post/pour-une-ia-francaise-tournee-vers-lavenir

[15] P. Aghion et A. Bouverot, *IA : notre ambition pour la France*. Odile Jacob, 2024.

[16] Wikipedia contributors, « AI safety », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=AI_safety&oldid=1239791378

[17] S. Russell, *Human Compatible: AI and the Problem of Control*. Penguin UK, 2019.

[18] J. Mecklin, « 'AI Godfather' Yoshua Bengio: We need a humanity defense organization », Bulletin of the Atomic Scientists. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://thebulletin.org/2023/10/ai-godfather-yoshua-bengio-we-need-a-humanity-defense-organization/

[19] « Reasoning through arguments against taking AI safety seriously », Yoshua Bengio. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/

[20] « Video: Geoffrey Hinton talks about the "existential threat" of AI », *MIT Technology Review*, 3 mai 2023. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.technologyreview.com/2023/05/03/1072589/video-geoffrey-hinton-google-ai-risk-ethics/

[21] S. Cave, « Risks from Artificial Intelligence ». Consulté le: 19 août 2024. [En ligne]. Disponible sur: https://www.cser.ac.uk/research/risks-from-artificial-intelligence

[22] « Center for AI safety (CAIS) ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.safe.ai/

[23] « Center for Human-Compatible Artificial Intelligence – Center for Human-Compatible AI is building exceptional AI for humanity ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://humancompatible.ai/

[24] F. Gustafsson, « Automotive safety systems », *IEEE Signal Process. Mag.*, vol. 26, nº 4, p. 32-47, juill. 2009.

[25] « Introducing Superalignment ». Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://openai.com/index/introducing-superalignment/

[26] K. Wiggers, « Google DeepMind forms a new org focused on AI safety », *TechCrunch*, 21 février 2024. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://techcrunch.com/2024/02/21/google-deepmind-forms-a-new-org-focused-on-ai-safety/

[27] « Anca Dragan named Head of AI Safety and Alignment at Google DeepMind », EECS at Berkeley. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://eecs.berkeley.edu/news/anca-dragan-named-head-of-ai-safety-and-alignment-at-google-deepmind

[28] « Responsibility & safety », Google DeepMind. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://deepmind.google/about/responsibility-safety/

[29] K. Wiggers, « Anthropic hires former OpenAI safety lead to head up new team », *TechCrunch*, 28 mai 2024. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://techcrunch.com/2024/05/28/anthropic-hires-former-openai-safety-lead-to-head-up-new-team/

[30] « Research ». Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://www.anthropic.com/research

[31] « Stanford AI safety ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://aisafety.stanford.edu/

[32] « MIT AI alignment », MIT AI Alignment. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.mitalignment.org/

[33] « DeCal ». Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://decal.studentorg.berkeley.edu/courses/7142

[34] « SafeAI: Safe Artificial Intelligence ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://safeai.ethz.ch/

[35] « Turing seminar – an introduction to AGI safety – master MVA ». Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.master-mva.com/cours/seminaire-turing/

[36] « Inauguration de la Chaire « Intelligence artificielle de confiance et responsable » », École polytechnique, école d'ingénieur. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.polytechnique.edu/actualites/inauguration-de-la-chaire-intelligence-artificielle-de-confiance-et-responsable

[37] D. Hendrycks, M. Mazeika, et T. Woodside, « An Overview of Catastrophic AI Risks », *arXiv [cs.CY]*, 21 juin 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2306.12001

[38] L. Weidinger *et al.*, « Ethical and social risks of harm from Language Models », *arXiv [cs.CL]*, 8 décembre 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2112.04359

[39] « Panorama », *Kardiologe*, vol. 3, n° 3, p. 192-193, juin 2009.

[40] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, et J. Stainer, « Machine learning with adversaries: Byzantine tolerant gradient descent », *Adv. Neural Inf. Process. Syst.*, p. 119-129, déc. 2017.

[41] E.-M. El-Mhamdi *et al.*, « On the Impossible Safety of Large AI Models », *arXiv [cs.LG]*, 30 septembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2209.15259

[42] B. Biggio, B. Nelson, et P. Laskov, « Poisoning Attacks against Support Vector Machines », *arXiv [cs.LG]*, 27 juin 2012. [En ligne]. Disponible sur: http://arxiv.org/abs/1206.6389

[43] F. Tramèr *et al.*, « Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets », in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, in CCS '22. New York, NY, USA: Association for Computing Machinery, nov. 2022, p. 2779-2792.

[44] N. Carlini *et al.*, « Are aligned neural networks adversarially aligned? », *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html

[45] E. Debenedetti *et al.*, « Privacy Side Channels in Machine Learning Systems », *arXiv [cs.CR]*, 11 septembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2309.05610

[46] Y. Huang, S. Gupta, M. Xia, K. Li, et D. Chen, « Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation », *arXiv [cs.CL]*, 10 octobre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2310.06987

[47] X. Liu, N. Xu, M. Chen, et C. Xiao, « AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models », *arXiv [cs.CL]*, 3 octobre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2310.04451

[48] J. Rando et F. Tramèr, « Universal Jailbreak Backdoors from Poisoned Human Feedback », *arXiv [cs.AI]*, 24 novembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2311.14455

[49] « Specification gaming: the flip side of AI ingenuity », Google DeepMind. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

[50] Z. Tufekci, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, 2017.

[51] S. Aral, *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy and Our Health – and How We Must Adapt*. HarperCollins, 2020.

[52] L. N. Hoang, « Science Communication Desperately Needs More Aligned Recommendation Algorithms », *Frontiers in Communication*, vol. 5, 2020, doi: 10.3389/fcomm.2020.598454.

[53] M. K. Lee *et al.*, « WeBuildAI: Participatory Framework for Algorithmic Governance », *Proc. ACM Hum.-Comput. Interact.*, vol. 3, nº CSCW, p. 1-35, nov. 2019.

[54] R. Noothigattu *et al.*, « A Voting-Based System for Ethical Decision Making », *AAAI*, vol. 32, nº 1, avr. 2018, doi: 10.1609/aaai.v32i1.11512.

[55] L. N. Hoang *et al.*, « Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees », *arXiv [cs.SI]*, 30 octobre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2211.01179

[56] R. A. Bradley et M. E. Terry, « Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons », *Biometrika*, vol. 39, nº 3/4, p. 324-345, 1952.

[57] J. Fageot, S. Farhadkhani, L.-N. Hoang, et O. Villemaud, « Generalized Bradley-Terry Models for Score Estimation from Paired Comparisons », *AAAI*, vol. 38, nº 18, p. 20379-20386, mars 2024.

[58] H. T. Pham, K. N. Nguyen, V. H. Phun, et T. K. Dang, « Secure Recommender System based on Neural Collaborative Filtering and Federated Learning », in *2022 International Conference on Advanced Computing and Analytics (ACOMPA)*, IEEE, nov. 2022, p. 1-11.

[59] J. Ding, Y. Quan, Q. Yao, Y. Li, et D. Jin, « Simplify and robustify negative sampling for implicit collaborative filtering », *Adv. Neural Inf. Process. Syst.*, vol. abs/2009.03376, sept. 2020, [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper/2020/hash/0c7119e3a6a2209da6a5b90e5b5b75bd-Abstract.html

[60] Wikipedia contributors, « Reinforcement learning from human feedback », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Reinforcement_learning_from_human_feedback&oldid=1234302797

[61] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, et D. Amodei, « Deep reinforcement learning from human preferences », *arXiv [stat.ML]*, 12 juin 2017. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.03741

[62] N. Nanda, « A comprehensive mechanistic interpretability explainer & glossary », Neel Nanda. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.neelnanda.io/mechanistic-interpretability/glossary

[63] L. Bereska et E. Gavves, « Mechanistic Interpretability for AI Safety -- A Review », *arXiv [cs.AI]*, 22 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.14082

[64] D. Rai, Y. Zhou, S. Feng, A. Saparov, et Z. Yao, « A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models », *arXiv [cs.AI]*, 2 juillet 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2407.02646

[65] « Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations », NIST. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://csrc.nist.gov/pubs/ai/100/2/e2023/final

[66] « [No title] ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

[67] « Centre pour la Sécurité de l'IA ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.securite-ia.fr

[68] « 2022 expert survey on progress in AI », AI Impacts. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

[69] « Yoshua Bengio ». Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://amturing.acm.org/award_winners/bengio_3406375.cfm

[70] « Geoffrey E. Hinton », in *Talking Nets*, The MIT Press, 2000.

[71] « 4 Charts That Show Why AI Progress Is Unlikely to Slow Down », *Time*. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://time.com/6300942/ai-progress-charts/

[72] « AI index report 2024 – artificial intelligence index ». Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://aiindex.stanford.edu/report/

[73] « International scientific report on the safety of advanced AI », GOV.UK. Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

[74] K. Grace, H. Stewart, J. F. Sandkühler, S. Thomas, B. Weinstein-Raun, et J. Brauner, « Thousands of AI Authors on the Future of AI », *arXiv [cs.CY]*, 5 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2401.02843

[75] K. Grace, J. F. Sandkühler, H. Stewart, S. Thomas, B. Weinstein-Raun, et J. Brauner, « 2022 expert survey on progress in AI », AI Impacts Wiki. Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://wiki.aiimpacts.org/doku.php?id=ai_timelines:predictions_of_human-level_ai_timelines:ai_timeline_surveys:2022_expert_survey_on_progress_in_ai.

[76] M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, et S. Russell, « Regulating advanced artificial agents », *Science*, vol. 384, n° 6691, p. 36-38, avr. 2024.

[77] *AutoGPT: AutoGPT is the vision of accessible AI for everyone, to use and to build on. Our mission is to provide the tools, so that you can focus on what matters*. Github. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://github.com/Significant-Gravitas/AutoGPT

[78] « Cognition ». Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.cognition.ai/blog/introducing-devin

[79] « Genie: SOTA Software engineering model ». Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://cosine.sh/genie

[80] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, et D. Ha, « The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery », *arXiv [cs.AI]*, 12 août 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2408.06292

[81] « Capital industriel moderne, demande de travail et dynamique des marchés de produits : le cas de la France ». Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://www.insee.fr/fr/statistiques/7613948

[82] L. Odot, « Are online recommendation algorithms polarising users' views? », Polytechnique Insights. Consulté le: 25 août 2024. [En ligne]. Disponible sur: https://www.polytechnique-insights.com/en/columns/digital/are-recommendation-algorithms-a-source-of-polarization/

[83] Y. Shi, « How Social Media and Embedded Recommender Algorithm Fostered Political Issues », Clark University, 2021. Consulté le: 25 août 2024. [En ligne]. Disponible sur: https://commons.clarku.edu/sps_masters_papers/91/

[84] P. Törnberg, « How digital media drive affective polarization through partisan sorting », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 119, n° 42, p. e2207159119, oct. 2022.

[85] K. Duskin, J. S. Schafer, J. D. West, et E. S. Spiro, « Echo Chambers in the Age of Algorithms: An Audit of Twitter's Friend Recommender System », *arXiv [cs.SI]*, 9 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.06422

[86] Y. Gao, F. Liu, et L. Gao, « Echo chamber effects on short video platforms », *Sci. Rep.*, vol. 13, n° 1, p. 6282, avr. 2023.

[87] « Myanmar: Time for Meta to pay reparations to Rohingya for role in ethnic cleansing », Amnesty International. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.amnesty.org/en/latest/news/2023/08/myanmar-time-for-meta-to-pay-reparations-to-rohingya-for-role-in-ethnic-cleansing/

[88] J. Whittaker, S. Looney, A. Reed, et F. Votta, « Recommender systems and the amplification of extremist content », *Internet Pol. Rev.*, vol. 10, n° 2, juin 2021, doi: 10.14763/2021.2.1565.

[89] R. Bellanova, K. Irion, K. Lindskov Jacobsen, F. Ragazzi, R. Saugmann, et L. Suchman, « Toward a Critique of Algorithmic Violence », *Int Polit Sociol*, vol. 15, n° 1, p. 121-150, mars 2021.

[90] T. T. Nguyen *et al.*, « Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures », *arXiv [cs.CR]*, 23 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.14942

[91] Z. Zhu *et al.*, « Understanding or Manipulation: Rethinking Online Performance Gains of Modern Recommender Systems », *ACM Trans. Inf. Syst. Secur.*, vol. 42, n° 4, p. 1-32, févr. 2024.

[92] Wikipedia contributors, « Russian interference in the 2016 United States elections », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Russian_interference_in_the_2016_United_States_elections&oldid=1240980413

[93] K. Dilanian, « Russian trolls who interfered in 2016 U.S. election also made ad money, report says », NBC News. Consulté le: 25 août 2024. [En ligne]. Disponible sur: https://www.nbcnews.com/politics/national-security/russian-trolls-who-interfered

-2016-u-s-election-also-made-n1013811

[94] « Here's What We Know So Far About Russia's 2016 Meddling », *Time*. Consulté le: 25 août 2024. [En ligne]. Disponible sur: https://time.com/5565991/russia-influence-2016-election/

[95] A. M. Khalaf, A. A. Alubied, A. M. Khalaf, et A. A. Rifaey, « The Impact of Social Media on the Mental Health of Adolescents and Young Adults: A Systematic Review », *Cureus*, vol. 15, nº 8, p. e42990, août 2023.

[96] « Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show », archive.ph. Consulté le: 25 août 2024. [En ligne]. Disponible sur: https://archive.ph/uX95E

[97] M. Chan et J. Yi, « Social Media Use and Political Engagement in Polarized Times. Examining the Contextual Roles of Issue and Affective Polarization in Developed Democracies », *Political Communication*, p. 1-20.

[98] A. P. Sundar, F. Li, X. Zou, T. Gao, et E. D. Russomanno, « Understanding shilling attacks and their detection traits: A comprehensive survey », *IEEE Access*, vol. 8, p. 171703-171715, 2020.

[99] I. Gunes, C. Kaleli, A. Bilge, et H. Polat, « Shilling attacks against recommender systems: a comprehensive survey », *Artificial Intelligence Review*, vol. 42, nº 4, p. 767-799, déc. 2014.

[100] « Facebook fake account deletion per quarter 2023 », Statista. Consulté le: 27 août 2024. [En ligne]. Disponible sur: https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/

[101] « La Dictature des Algorithmes », Éditions Tallandier. Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://www.tallandier.com/livre/la-dictature-des-algorithmes/

[102] Wikipedia contributors, « Intelligence artificielle », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Intelligence_artificielle&oldid=217518957

[103] Wikipedia contributors, « Deep learning », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1239902356

[104] Wikipedia contributors, « Large language model », Wikipedia, The Free Encyclopedia. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Large_language_model&oldid=1239909148

[105] J. Wei *et al.*, « Emergent Abilities of Large Language Models », *arXiv [cs.CL]*, 15 juin 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2206.07682

[106] S. Bubeck *et al.*, « Sparks of Artificial General Intelligence: Early experiments with GPT-4 », *arXiv [cs.CL]*, 22 mars 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2303.12712

[107] R. Ngo, « Visualizing the deep learning revolution - Richard Ngo », Medium. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5

[108] J. E. Dobson, « On reading and interpreting black box deep neural networks », *International Journal of Digital Humanities*, vol. 5, nº 2, p. 431-449, nov. 2023.

[109] T. Hagendorff, « Deception abilities emerged in large language models », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 121, nº 24, p. e2317967121, juin 2024.

[110] J. Vincent, « Microsoft's Bing is an emotionally manipulative liar, and people love it », The Verge. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

[111] R. Fang, R. Bindu, A. Gupta, Q. Zhan, et D. Kang, « LLM Agents can Autonomously Hack Websites », *arXiv [cs.CR]*, 6 février 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2402.06664

[112] R. Fang, R. Bindu, A. Gupta, Q. Zhan, et D. Kang, « Teams of LLM Agents can Exploit Zero-Day Vulnerabilities », *arXiv [cs.MA]*, 2 juin 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2406.01637

[113] R. Fang, R. Bindu, A. Gupta, et D. Kang, « LLM Agents can Autonomously Exploit One-day Vulnerabilities », *arXiv [cs.CR]*, 11 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.08144

[114] M. Shao *et al.*, « An Empirical Evaluation of LLMs for Solving Offensive Security Challenges », *arXiv [cs.CR]*, 19 février 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2402.11814

[115] D. Kiela, « Plotting progress in AI », Contextual AI. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://contextual.ai/news/plotting-progress-in-ai/

[116] J. Kaplan *et al.*, « Scaling Laws for Neural Language Models », *arXiv [cs.LG]*, 23 janvier 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/2001.08361

[117] « Large-scale AI models », Epoch AI. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://epochai.org/data/large-scale-ai-models

[118] « Exclusive: OpenAI working on new reasoning technology under code name 'Strawberry' », *Reuters*, 12 juillet 2024. Consulté le: 8 septembre 2024. [En ligne]. Disponible sur: https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12/

[119] « OpenAI Charter ». Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://openai.com/charter/

[120] « SITUATIONAL AWARENESS: The decade ahead », SITUATIONAL AWARENESS - The Decade Ahead. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://situational-awareness.ai/

[121] « When Might AI Outsmart Us? It Depends Who You Ask », *Time*. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://time.com/6556168/when-ai-outsmart-humans/

[122] W. Yu *et al.*, « Language to Rewards for Robotic Skill Synthesis », *arXiv [cs.RO]*, 14 juin 2023. doi: 10.48550/ARXIV.2306.08647.

[123] Figure, « Figure Status Update - OpenAI Speech-to-Speech Reasoning ». Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://www.youtube.com/watch?v=Sq1QZB5baNw

[124] P. Gmyrek, J. Berg, et D. Bescond, « Generative AI and jobs: A global analysis of potential effects on job quantity and quality », *SSRN Electron. J.*, août 2023, doi: 10.2139/ssrn.4584219.

[125] C. Pizzinelli, A. J. Panton, M. M. Tavares, M. Cazzaniga, et L. Li, *Labor Market Exposure to AI: Cross-country Differences and Distributional Implications*. International Monetary Fund, 2023.

[126] « The state of phishing 2024 », SlashNext | Complete Generative AI Security for Email, Mobile, and Browser. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://slashnext.com/the-state-of-phishing-2024/

[127] J. Hazell, « Spear Phishing With Large Language Models », *arXiv [cs.CY]*, 11 mai 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2305.06972

[128] D. Milmo, « Company worker in Hong Kong pays out £20m in deepfake video call scam », *The Guardian*, The Guardian, 5 février 2024. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam

[129] G. Boesch, « What Is Adversarial Machine Learning? Attack Methods in 2024 », viso.ai. Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://viso.ai/deep-learning/adversarial-machine-learning/

[130] « Training data extraction attacks ». Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://www.nightfall.ai/ai-security-101/training-data-extraction-attacks

[131] « The near-term impact of AI on the cyber threat ». Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

[132] « World Economic Forum », World Economic Forum. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.weforum.org/publications/global-cybersecurity-outlook-2023/

[133] « Voice of SecOps reports », Deep Instinct. Consulté le: 12 août 2024. [En ligne]. Disponible sur: https://www.deepinstinct.com/voice-of-secops-reports

[134] « Open-sourcing highly capable foundation models ». Consulté le: 13 août 2024. [En ligne]. Disponible sur: https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[135] R. Slayton, « Conceptions, Causes, and Assessment », *Int. Secur.*, vol. 41, n° 3, p. 72-109, 2016.

[136] *Defending a New Domain: The Pentagon's Cyberstrategy*. Defense Technical Information Center, 2010.

[137] P. Venables et C. Snyder, « Cloud CISO Perspectives: Building better cyber defenses with AI », Google Cloud Blog. Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://cloud.google.com/blog/products/identity-security/cloud-ciso-perspectives-building-better-cyber-defenses-with-ai/

[138] P. Gade, S. Lermen, C. Rogers-Smith, et J. Ladish, « BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B », *arXiv [cs.CL]*, 31 octobre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2311.00117

[139] « Disrupting deceptive uses of AI by covert influence operations ». Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/

[140] C. Mouton, C. Lucas, et E. Guest, *The Operational Risks of AI in Large-Scale Biological Attacks*. RAND Corporation, 2024.

[141] « Reddit - Dive into anything ». Consulté le: 13 août 2024. [En ligne]. Disponible sur: https://www.reddit.com/r/ArtistHate/

[142] J. Wolters, « The animation guild: Future unscripted: The Impact of Generative AI on entertainment industry jobs », INDAC. Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://indac.org/blog/the-animation-guild-future-unscripted-the-impact-of-gener

ative-ai-on-entertainment-industry-jobs/

[143] « The potentially large effects of artificial intelligence on economic growth (Briggs/kodnani) ». Consulté le: 13 août 2024. [En ligne]. Disponible sur: https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

[144] « Sondage Ifop - Talan : les Français et les IA génératives ». Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://www.talan.com/actualites/detail-actualites/news/sondage-ifop-talan-les-francais-et-les-ia-generatives/

[145] « OpinionWay pour Universcience - Barometre de l'esprit critique 2024 - Mars 2024 ». Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/opinion-societe/opinionway-pour-universcience-barometre-de-l-esprit-critique-2024-mars-2024.html

[146] « Intelligence artificielle : un Français sur deux inquiet », BCG Global. Consulté le: 14 août 2024. [En ligne]. Disponible sur: https://www.bcg.com/press/26april2024-intelligence-artificielle-un-francais-sur-deux-inquiet

[147] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, et Y. Zhang, « A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly », *arXiv [cs.CR]*, 4 décembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2312.02003

[148] « AI Safety Summit », AISS 2023. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.aisafetysummit.gov.uk/

[149] « The bletchley declaration by countries attending the AI safety summit, 1-2 November 2023 », GOV.UK. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[150] « Action Plan to increase the safety and security of advanced AI ». Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.gladstone.ai/action-plan

[151] « International scientific report on the safety of advanced AI », GOV.UK. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

[152] R. J. Neuwirth, *The EU artificial intelligence act: Regulating subliminal AI systems*. in Routledge Research in the Law of Emerging Technologies. London, England: Routledge, 2022. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://artificialintelligenceact.eu/

[153] The White House, « Executive order on the safe, secure, and trustworthy development and use of artificial intelligence », The White House. Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[154] D. Patterson *et al.*, « Carbon Emissions and Large Neural Network Training », *arXiv [cs.LG]*, 21 avril 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2104.10350

[155] J. Saul et D. Bass, « Artificial Intelligence Is Booming—So Is Its Carbon Footprint

», *Bloomberg News*, 9 mars 2023. Consulté le: 2 septembre 2024. [En ligne]. Disponible sur: https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure

[156] G. Livingstone, « 'It's pillage': thirsty Uruguayans decry Google's plan to exploit water supply », *The Guardian*, The Guardian, 11 juillet 2023. Consulté le: 2 septembre 2024. [En ligne]. Disponible sur: https://www.theguardian.com/world/2023/jul/11/uruguay-drought-water-google-data-center

[157] S. Kim *et al.*, « Chemical use in the semiconductor manufacturing industry », *Int. J. Occup. Environ. Health*, vol. 24, nº 3-4, p. 109-118, oct. 2018.

[158] C. Simpson, « American Chipmakers Had a Toxic Problem. Then They Outsourced It », *Bloomberg News*, 15 juin 2017. Consulté le: 2 septembre 2024. [En ligne]. Disponible sur: https://www.bloomberg.com/news/features/2017-06-15/american-chipmakers-had-a-toxic-problem-so-they-outsourced-it

[159] T. Tech, « Anne Bouverot : La France, capitale européenne de l'iA ? » Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://www.youtube.com/watch?v=oQJhqxnDsjo

[160] F. Brewster, « Big Tech is lobbying hard to keep copyright law favorable to AI ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://jacobin.com/2023/11/artificial-intelligence-big-tech-lobbying-copyright-infringement-regulation/

[161] « TTP - funding the fight against antitrust: How Facebook's antiregulatory attack dog spends its millions ». Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://www.techtransparencyproject.org/articles/funding-fight-against-antitrust-how-facebooks-antiregulatory-attack-dog-spends-its-millions

[162] S. Hashim, « Meta-funded group floods Facebook with anti-AI regulation ads », Transformer. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.transformernews.ai/p/american-edge-meta-ai-regulation-lobbying

[163] S. Hashim, « Tech companies are trying to kill California's AI regulation bill », Transformer. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.transformernews.ai/p/a16z-y-combinator-big-tech-sb1047-lobbying

[164] « Trojan horses: how European startups teamed up with Big Tech to gut the AI Act ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://corporateeurope.org/en/2024/03/trojan-horses-how-european-startups-teamed-big-tech-gut-ai-act

[165] S. User, « Yann Le Cun ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.academie-sciences.fr/fr/Liste-des-membres-de-l-Academie-des-sciences-/-L/yann-le-cun.html

[166] O. LinkedIn, « About linkedin », *LinkedIn Corporation*, 2022, [En ligne]. Disponible sur: https://www.linkedin.com/posts/yann-lecun_today-i-was-made-a-chevalier-de-la-l%C3%A9gion-activity-7138326352776056832-GQFp/

[167] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/ylecun

[168] « LinkedIn ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.linkedin.com/in/yann-lecun/

[169] A. R. Chow, « Yann LeCun », *Time*, 7 septembre 2023. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://time.com/collection/time100-ai/6309052/yann-lecun/

[170] « LinkedIn ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.linkedin.com/posts/yann-lecun_current-and-former-colleagues-of-the-french-activity-7200524405267988480-NyaA/

[171] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/ylecun/status/1747714991292399763

[172] « Yann LeCun - Our CTO, Mike Schroepfer, gives the details of the ». Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.facebook.com/yann.lecun/posts/10155043225132143

[173] « [No title] », X (formerly Twitter). Consulté le: 18 août 2024. [En ligne]. Disponible sur: https://x.com/ylecun/status/1688140941079498752

[174] A. I. Mistral, « Mistral AI ». Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://mistral.ai/

[175] K.D, « IA: la start-up française Mistral AI valorisée 6 milliards d'euros après une nouvelle levée de fonds », Tech&Co. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.bfmtv.com/tech/intelligence-artificielle/ia-la-startup-francaise-mistral-ai-valorisee-6-milliards-d-euros-apres-une-nouvelle-levee-de-fonds_AD-202406110709.html

[176] S. Pommier, « L'ex-ministre Cédric 0 pourrait empocher 23 millions d'euros après avoir investi 176 euros dans Mistral AI », Capital.fr. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.capital.fr/economie-politique/mistral-ai-la-bonne-affaire-du-conseiller-fondateur-cedric-o-1488506

[177] H. A. P. la T. de la Vie Publique, « Délibération n° 2022-189 du 14 juin 2022 relative au projet de reconversion professionnelle de Monsieur Cédric O », juin 2022. [En ligne]. Disponible sur: https://www.hatvp.fr/wordpress/wp-content/uploads/2022/07/2022-189-Cedric-O.pdf

[178] Z. Wanat, « 'EU's AI act could kill our company,' says Mistral's Cédric O », Sifted. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://sifted.eu/articles/eu-ai-act-kill-mistral-cedric-o

[179] C. Axiotes, « Lobbying for loopholes: The battle over foundation models in the EU AI act », EURACTIV. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.euractiv.com/section/digital/opinion/lobbying-for-loopholes-the-battle-over-foundation-models-in-the-eu-ai-act/

[180] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/tegmark/status/1728851164291059948

[181] I. Ramdani, « Intelligence artificielle : comment un ministre devenu lobbyiste a retourné le gouvernement », Mediapart. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.mediapart.fr/journal/france/011223/intelligence-artificielle-comment-un-ministre-devenu-lobbyiste-retourne-le-gouvernement

[182] S. Pommier et S. Barge, « Régulation de l'IA : l'ex-secrétaire d'Etat Cédric O jongle entre ses intérêts et ceux de la France », Capital.fr. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.capital.fr/economie-politique/regulation-de-lia-lex-secretaire-detat-c

edric-o-jongle-entre-ses-interets-et-ceux-de-la-france-1491749

[183] T. Hartmann, « Les coulisses de l'opposition de la France à la réglementation des modèles d'IA », EURACTIV. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.euractiv.fr/section/intelligence-artificielle/news/les-coulisses-de-lop position-de-la-fance-a-la-reglementation-des-modeles-dia/

[184] T. Hartmann, « AI Act : le gouvernement accusé d'avoir été influencé par un lobbyiste en situation de conflit d'intérêts », EURACTIV. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.euractiv.fr/section/intelligence-artificielle/news/ai-act-le-gouvernem ent-accuse-davoir-ete-influence-par-un-lobbyiste-en-situation-de-conflit-dinteret s/

[185] « Open letter EU AI act and signatories.Pdf », Google Docs. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://drive.google.com/file/d/1wrtxfvcD9FwfNfWGDL37Q6Nd8wBKXCkn/view

[186] L. Bertuzzi, « AI Act : négociations bloquées à cause de divergences sur les modèles de fondation », EURACTIV. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.euractiv.fr/section/intelligence-artificielle/news/ai-act-negociations-b loquees-a-cause-de-divergences-sur-les-modeles-de-fondation/?_ga=2.5019709 2.31701998.1701074244-1669050714.1683052757

[187] Propos recueillis par Philippe Mabille et Sylvain Rolland au Forum AIM Marseille, « Intelligence artificielle : « Les Gafam et la startup Mistral ne défendent pas l'intérêt général » (Thierry Breton) », La Tribune. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://www.latribune.fr/technos-medias/informatique/intelligence-artificielle-les-gafam-et-la-startup-mistral-ne-defendent-pas-l-interet-general-thierry-breton-98 4046.html

[188] O. Cédric, « This is worth some clarification. Before I do, Mr @tegmark, I shall say how pathetic I find your original tweet. You can obviously deeply disagree with my point of view, but the way you phrased it on a personal attack on my probity is to me epitomizes the decay of public debate.... », Twitter. Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://twitter.com/cedric_o/status/1728724005459235052

[189] Marc Andreessen, Ben Horowitz, Yann LeCun, Ben Fielding, Ion Stoica, Naveen Rao, Arthur Mensch, Garry Tan, Amjad Masad, Bill Gurley, Herman Narula,Tobi Lütke, Suhail Doshi, Clem Delangue, Aravind Srinivasan, Soumith Chintala, Tyler Cowen, John Carmack., « Letter to President Biden regarding the AI Executive Order », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/a16z/status/1720524920596128012

[190] Andreessen-Horowitz, 1 décembre 2023. [En ligne]. Disponible sur: https://committees.parliament.uk/writtenevidence/127070/pdf

[191] M. Casado, 6 décembre 2023. Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://www.schumer.senate.gov/imo/media/doc/Martin%20Casado%20-%20Stat ement.pdf

[192] V. Hassija *et al.*, « Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence », *Cognit. Comput.*, vol. 16, nº 1, p. 45-74, janv. 2024.

[193] A. Chaszczewicz, « Is Task-Agnostic Explainable AI a Myth? », *arXiv [cs.AI]*, 13 juillet 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2307.06963

[194] A. Barredo Arrieta *et al.*, « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Inf. Fusion*, vol. 58, p. 82-115, juin 2020.

[195] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/edardaman/status/1744453999695176046

[196] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/NeelNanda5/status/1799203292066558403

[197] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/m_bourgon/status/1798863250173657312

[198] « [No title] », X (formerly Twitter). Consulté le: 15 août 2024. [En ligne]. Disponible sur: https://x.com/ID_AA_Carmack/status/1799147185793348006

[199] Sénat, « Intelligence artificielle : la France en pointe ? » Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://www.youtube.com/watch?v=tWXuGuOHMhk

[200] T. Petersen, « AI's new power of persuasion: it can change your mind », avr. 2024, Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://actu.epfl.ch/news/ai-s-new-power-of-persuasion-it-can-change-your-mi/

[201] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, et J. Zhu, « Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges », in *Natural Language Processing and Chinese Computing*, Springer International Publishing, 2019, p. 563-574.

[202] « Guardrailing ». Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://docs.mistral.ai/capabilities/guardrailing/

[203] Y. Wolf, N. Wies, O. Avnery, Y. Levine, et A. Shashua, « Fundamental Limitations of Alignment in Large Language Models », *arXiv [cs.CL]*, 19 avril 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2304.11082

[204] « Aligning language models to follow instructions ». Consulté le: 16 août 2024. [En ligne]. Disponible sur: https://openai.com/index/instruction-following/

[205] J. Ji *et al.*, « Language Models Resist Alignment », *arXiv [cs.CL]*, 10 juin 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2406.06144

[206] Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, et G. Irving, « Alignment of Language Agents », *arXiv [cs.AI]*, 26 mars 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2103.14659

[207] U. Anwar *et al.*, « Foundational Challenges in Assuring Alignment and Safety of Large Language Models », *arXiv [cs.LG]*, 15 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.09932

[208] « Artificial Intelligence: A Modern Approach, 4th US ed ». Consulté le: 2 septembre 2024. [En ligne]. Disponible sur: https://aima.cs.berkeley.edu

[209] N. J. Nilsson, *The Quest for Artificial Intelligence*. Cambridge University Press, 2009.

[210] « NEW SAVANNA ». Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://new-savanna.blogspot.com/2023/11/a-dialectical-view-of-history-of-ai.html

[211] « A brief history: 70 years of machine learning ». Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://www.inveniam.fr/blog-brief-history-70-years-of-machine-learning

[212] « AlphaGo », Google DeepMind. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://deepmind.google/technologies/alphago/

[213] « AlphaGo versus Ke Jie ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://en.wikipedia.org/wiki/AlphaGo_versus_Ke_Jie

[214] « Terrence J. sejnowski », MIT Press. Consulté le: 5 septembre 2024. [En ligne]. Disponible sur: https://mitpress.mit.edu/author/terrence-j-sejnowski-2310

[215] « Grand modèle de langage ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/Grand_mod%C3%A8le_de_langage

[216] « Analyse lexicale ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/Analyse_lexicale

[217] F. C. Oettl, J. F. Oeding, et K. Samuelsson, « Explainable artificial intelligence in orthopedic surgery », *J Exp Orthop*, vol. 11, n° 3, juill. 2024, doi: 10.1002/jeo2.12103.

[218] « Explainable artificial intelligence ». Consulté le: 9 septembre 2024. [En ligne]. Disponible sur: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

[219] A. Jones, « Introduction to mechanistic Interpretability », BlueDot Impact. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://aisafetyfundamentals.com/blog/introduction-to-mechanistic-interpretability/

[220] « Modèle de fondation ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_fondation

[221] « Deployment ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://ailabwatch.org/categories/deployment/

[222] P. Villalobos, « Scaling laws literature review », Epoch AI. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://epochai.org/blog/scaling-laws-literature-review

[223] S. Capital, « Microsoft CTO Kevin Scott on How Far Scaling Laws Will Extend | Training Data ». Consulté le: 8 septembre 2024. [En ligne]. Disponible sur: https://www.youtube.com/watch?v=aTQWymHp0n0

[224] A. Patel, « NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models », NVIDIA Blog. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/

[225] J. Calma, « Microsoft is going nuclear to power its AI ambitions », The Verge. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring

[226] K. Hagey et A. Fitch, « Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI », *The Wall Street Journal*, The Wall Street Journal, 9 février 2024. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0

[227] D. Leprince-Ringuet, « AI startups need more data centres. France wants to build them », Sifted. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://sifted.eu/articles/ai-startups-infrastructure-europe

[228] « OpenAI Data Partnerships ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://openai.com/index/data-partnerships/

[229] « Introducing GPT-4o and more tools to ChatGPT free users ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free

[230] « Claude 3.5 Sonnet ». Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.anthropic.com/news/claude-3-5-sonnet

[231] E. Woo, S. Palazzolo, et A. Efrati, « OpenAI Races to Launch 'Strawberry' Reasoning AI to Boost Chatbot Business », The Information. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.theinformation.com/articles/openai-races-to-launch-strawberry-reasoning-ai-to-boost-chatbot-business

[232] J. O'Donnell, « Sam Altman says helpful agents are poised to become AI's killer function », *MIT Technology Review*, 1 mai 2024. Consulté le: 1 septembre 2024. [En ligne]. Disponible sur: https://www.technologyreview.com/2024/05/01/1091979/sam-altman-says-helpful-agents-are-poised-to-become-ais-killer-function/