AI ACTION SUMMIT

Rapport international sur la sûreté de l'IA

Rapport scientifique international sur la sûreté de l'IA avancée

Janvier 2025

Collaborateurs

PRÉSIDENT

Prof. Yoshua Bengio, Université de Montréal / Mila, Institut québécois d'intelligence artificielle

GROUPE CONSULTATIF D'EXPERTS

Ce panel international a été nommé par les gouvernements des 30 pays énumérés ci-dessous, l'ONU, l'UE et l'OCDE.

Australie : Bronwyn Fox, Université de Nouvelle-Galles du Sud

André Carlos Ponce de Leon Ferreira de Carvalho, Institut de mathématiques et d'informatique, Université de São Paulo

Canada : Mona Nemer, conseillère scientifique en chef du Canada

Chili : Raquel Pezoa Rivera, Universidad Técnica Federico Santa Maria

Chine : Yi Zeng, Académie chinoise des sciences

Union européenne : Juha Heikkilä, Bureau européen de l'IA

France : Guillaume Avrin, Coordination nationale pour l'intelligence artificielle

Allemagne : Antonio Krüger, Centre allemand de recherche sur l'intelligence artificielle

Inde : Balaraman Ravindran, École Wadhwani de science des données et d'IA, Institut indien de technologie de Madras

Indonésie : Hammam Riza, Recherche Collaborative et Innovation Industrielle en Intelligence Artificielle (KORIKA)

Irlande : Ciarán Seoighe, Research Ireland

Israël : Ziv Katzir, Autorité israélienne de l'innovation

Italie : Andrea Monti, juriste auprès du sous-secrétaire d'État à la transformation numérique, présidence italienne du Conseil des ministres

Japon : Hiroaki Kitano, Sony Group Corporation

Kenya : Nusu Mwamanzi, ministère des TIC et de l'économie numérique

Royaume d'Arabie saoudite : Fahad Albalawi, Autorité saoudienne pour les données et l'intelligence artificielle

Mexique : José Ramón López Portillo, LobsterTel

Pays-Bas : Haroon Sheikh, Conseil scientifique néerlandais pour la politique gouvernementale

Nouvelle-Zélande : Gill Jolly, ministère des Entreprises, de l'Innovation et de l'Emploi

Nigéria : Olubunmi Ajala, ministère des Communications, de l'Innovation et de l'Économie numérique

OCDE : Jerry Sheehan, directeur de la Direction de la science, de la technologie et de l'innovation

Philippines : Dominic Vincent Ligot, CirroLytix

République de Corée : Kyoung Mu Lee, Département du génie électrique et informatique, Université nationale de Séoul

Rwanda : Crystal Rugege, Centre pour la quatrième Révolution industrielle

Singapour : Denise Wong, Groupe d'innovation et de protection des données, Infocomm Media Development Authority

Espagne : Nuria Oliver, ELLIS Alicante

Suisse : Christian Busch, Département fédéral de l'économie, de la formation et de la recherche

Turquie : Ahmet Halit Hatip, ministère turc de l'Industrie et de la Technologie

Ukraine : Oleksii Molchanovskyi, Comité d'experts sur le développement de l'intelligence artificielle en Ukraine

Émirats arabes unis : Marwan Alserkal, Ministère des Affaires du Cabinet, Cabinet du Premier ministre

Royaume-Uni : Chris Johnson, conseiller scientifique en chef au ministère de la Science, de l'Innovation et de la Technologie

Nations Unies : Amandeep Singh Gill, Secrétaire général adjoint aux technologies numériques et émergentes et Envoyé du Secrétaire général pour la technologie

États-Unis : Saïf M. Khan, États-Unis Département du Commerce

DIRECTEUR SCIENTIFIQUE

Sören Mindermann, Mila, Institut québécois d'intelligence artificielle

RÉDACTEUR PRINCIPAL

Daniel Privitera, KIRA Center

GROUPE D'ÉCRITURE

Tamay Besiroglu, Epoch AI

Rishi Bommasani, Université de Stanford

Stephen Casper, Institut de technologie du Massachusetts

Yejin Choi, Université de Stanford

Philip Fox, KIRA Center

Ben Garfinkel, Université d'Oxford

Danielle Goldfarb, Mila, Institut québécois d'intelligence artificielle

Hoda Heidari, Université Carnegie Mellon

Anson Ho, Epoch AI

Sayash Kapoor, Université de Princeton

Leila Khalatbari, Université des sciences et technologies de Hong Kong

Shayne Longpre, Institut de technologie du Massachusetts

Sam Manning, Centre pour la gouvernance de l'IA

Vasilios Mavroudis, Institut Alan Turing

Mantas Mazeika, Université de l'Illinois à Urbana-Champaign

Julian Michael, Université de New York

Jessica Newman, Université de Californie, Berkeley

Kwan Yee Ng, Concordia AI

Chinesa T. Okolo, Brookings Institution

Deborah Raji, Université de Californie, Berkeley

Girish Sastry, Indépendant

Elizabeth Seger (écrivain généraliste), Demos

Theodora Skeadas, Humane Intelligence

Tobin South, Institut de technologie du Massachusetts

Emma Strubell, Université Carnegie Mellon

Florian Tramèr, ETH Zurich

Lucia Velasco, Université de Maastricht

Nicole Wheeler, Université de Birmingham

REMERCIEMENTS

Examinateurs de la société civile et de l'industrie

Société civile : Institut Ada Lovelace, AI Forum New Zealand / Te Kāhui Atamai Iahiko o Aotearoa, Groupe temporaire d'experts en IA d'Australie, Carnegie Endowment for International Peace, Center for Law and Innovation / Certa Foundation, Centre pour la gouvernance de l'IA, Chief Justice Meir Shamgar Center for Digital Law and Innovation, Institut Eon, Gradient Institute, Israel Democracy Institute, Fondation Mozilla, Old Ways New, RAND, SaferAI, Centre pour la résilience à long terme, The Future Society, Institut Alan Turing, Royal Society, Association turque des politiques d'intelligence artificielle.

Industrie : Advai, Anthropic, Cohere, Deloitte Consulting USA et Deloitte LLM UK, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, Zhipu.ai.

Remerciements spéciaux

Le Secrétariat apprécie le soutien, les commentaires et les retours d'information d'Angie Abdilla, Nitarshan Rajkumar, Geoffrey Irving, Shannon Vallor, Rebecca Finlay et Andrew Strait.

Toute demande de renseignements concernant cette publication doit être envoyée à : secretariat.AIStateofScience@dsit.gov.uk.

Les demandes de renseignements concernant le contenu du rapport doivent également être adressées au directeur scientifique.

Avertissement

Le rapport ne représente pas les points de vue du Président, d'un quelconque individu au sein des groupes de rédaction ou de conseil, ni d'aucun des gouvernements qui ont soutenu son élaboration. Ce rapport est une synthèse des recherches existantes sur les capacités et les risques de l'IA avancée. Le Président du rapport en est l'ultime responsable et a supervisé son élaboration du début à la fin.

Numéro de série de recherche : DSIT 2025/001

À propos de ce rapport

- **Il s'agit du premier Rapport international sur la sûreté de l'IA.** Suite à une publication intermédiaire en mai 2024, un groupe diversifié de 96 experts en intelligence artificielle (IA) a contribué à ce premier rapport complet, dont un Groupe consultatif d'experts internationaux nommés par 30 pays, l'Organisation de coopération et de développement économiques (OCDE), l'Union européenne (UE) et les Nations unies (ONU). Le rapport vise à fournir des informations scientifiques qui soutiendront l'élaboration de politiques éclairées. Il ne recommande pas de politiques spécifiques.
- **Le rapport est le travail d'experts indépendants.** Sous la direction du Président, les experts indépendants ayant rédigé ce rapport ont bénéficié d'une totale liberté pour en définir le contenu.
- **Bien que ce rapport porte sur les risques liés à l'IA et sa sûreté, l'IA offre également de nombreux bénéfices potentiels pour les personnes, les entreprises et la société.** Il existe de nombreux types d'IA, chacun présentant des bénéfices et des risques différents. La plupart du temps et dans la plupart des applications, l'IA aide les individus et les organisations à être plus efficaces. Mais les citoyens du monde entier ne pourront profiter pleinement et en toute sécurité des nombreux bénéfices potentiels de l'IA que si les risques qu'elle comporte sont gérés de manière appropriée. Ce rapport vise à identifier ces risques et à évaluer les méthodes permettant de les atténuer. Il ne vise pas à évaluer de manière exhaustive tous les impacts sociétaux possibles de l'IA, y compris ses nombreux bénéfices potentiels.
- **Le rapport se concentre sur l'IA à usage général.** Le périmètre du rapport se limite à un type d'IA qui a progressé particulièrement rapidement ces dernières années et dont les risques associés ont été moins étudiés et compris : l'IA à usage général, ou IA capable d'effectuer une grande variété de tâches. L'analyse présentée dans ce rapport se concentre sur les systèmes d'IA à usage général les plus avancés au moment de la rédaction, ainsi que sur les systèmes futurs qui pourraient être encore plus performants.
- **Le rapport résume les éléments scientifiques sur trois questions fondamentales :** Que peut faire l'IA à usage général ? Quels sont les risques associés à l'IA à usage général ? Et quelles techniques d'atténuation existent contre ces risques ?
- **Les enjeux sont importants.** Nous, les experts contribuant à ce rapport, continuons d'être en désaccord sur plusieurs questions, mineures et majeures, concernant les capacités de l'IA à usage général, les risques et les mesures d'atténuation de ces risques. Mais nous considérons que ce rapport est essentiel pour améliorer notre compréhension collective de cette technologie et de ses risques potentiels. Nous espérons que le rapport aidera la communauté internationale à progresser vers un plus grand consensus sur l'IA à usage général et à atténuer ses risques plus efficacement, afin que chacun puisse bénéficier en toute sécurité de ses nombreux bénéfices potentiels. Les enjeux sont importants. Nous sommes impatients de poursuivre cet effort.

Point sur les dernières avancées de l'IA après la rédaction de ce rapport : Note du président

Entre la fin de la période de rédaction de ce rapport (5 décembre 2024) et la publication de ce rapport en janvier 2025, une évolution importante a eu lieu. La société d'IA OpenAI a partagé les premiers résultats des tests d'un nouveau modèle d'IA, *o3*. Ces résultats indiquent des performances significativement plus fortes que tout autre modèle précédent sur un certain nombre des tests les plus difficiles du domaine de la programmation, du raisonnement abstrait et du raisonnement scientifique. Dans certains de ces tests, o3 surpasse de nombreux (mais pas tous) experts humains. De plus, il représente une avancée décisive dans un test clé de raisonnement abstrait que de nombreux experts, dont moi-même, pensaient hors de portée jusqu'à récemment. Cependant, au moment de la rédaction du présent document, il n'existe aucune information publique sur ses capacités dans le monde réel, en particulier pour résoudre des tâches plus ouvertes.

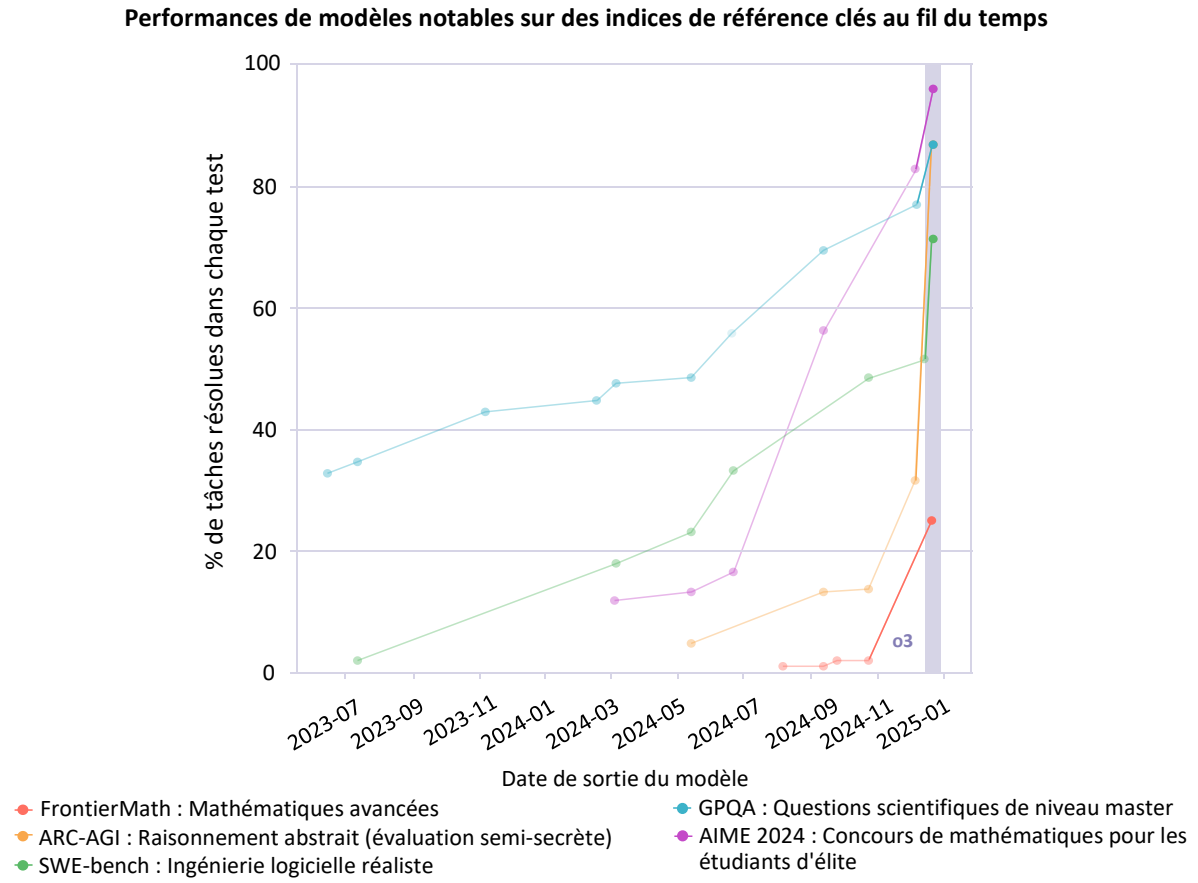

Figure 0.1 : *Résultats de modèles d'IA à usage général notables sur les principaux tests de référence de juin 2023 à décembre 2024. o3 a montré des performances considérablement améliorées par rapport au précédent état de l'art (zone ombrée). Ces tests de référence comptent parmi les plus difficiles du domaine de la programmation, du raisonnement abstrait et du raisonnement scientifique. Pour le o3 non publié, la date d'annonce est indiquée ; pour les autres modèles, la date de sortie est indiquée. Certains des modèles d'IA les plus récents, comme o3, ont été optimisés grâce à une structure renforcée et une puissance de calcul accrue lors des phases de test.*

Sources : Anthropic, 2024; Chollet, 2024; *Chollet et al., 2025*; *Epoch AI, 2024*; *Glazer et al. 2024*; *OpenAI, 2024a*; *OpenAI, 2024b*; *Jimenez et al., 2024*; *Jimenez et al., 2025.*

Les résultats d'o3 démontrent que le rythme des avancées en matière de capacités de l'IA pourrait rester élevé, voire s'accélérer. Plus précisément, ils suggèrent que donner aux modèles plus de puissance de calcul pour résoudre un problème donné (« mise à l'échelle de l'inférence ») peut aider à surmonter les limitations précédentes. D'une manière générale, la mise à l'échelle de l'inférence rend les modèles plus coûteux à utiliser. Mais comme un autre modèle notable récent, *R1*, lancée par la société DeepSeek en janvier 2025, a montré que les chercheurs travaillent avec succès à réduire ces coûts. Dans l'ensemble, la mise à l'échelle de l'inférence peut permettre aux développeurs d'IA de réaliser de nouvelles avancées à l'avenir. Les résultats d'o3 soulignent également la nécessité de mieux comprendre comment l'utilisation croissante de l'IA par les développeurs d'IA peut affecter la vitesse du développement ultérieur de l'IA lui-même.

Les tendances mises en évidence par o3 pourraient avoir de profondes implications sur les risques liés à l'IA. Les progrès de la science et des capacités de programmation ont déjà produit davantage d'éléments probants sur les risques tels que les attaques cybernétiques et biologiques. Les résultats d'o3 sont également pertinents en ce qui concerne les impacts potentiels sur le marché du travail, le risque de perte de contrôle et la consommation d'énergie, entre autres. Mais les capacités d'o3 pourraient également être utilisées pour aider à protéger contre les dysfonctionnements et les utilisations malveillantes. Dans l'ensemble, les évaluations des risques présentées dans ce rapport doivent être lues en gardant à l'esprit que l'IA a acquis de nouvelles capacités depuis la rédaction du rapport. Cependant, en l'état, il n'existe encore aucune preuve des impacts réels d'o3, ni aucune information permettant de confirmer ou d'exclure de nouveaux risques majeurs et/ou immédiats.

L'amélioration des capacités suggérée par les résultats d'o3 et notre compréhension limitée des implications pour les risques liés à l'IA soulignent un défi majeur pour les décideurs politiques que ce rapport identifie : ils devront souvent évaluer les bénéfices et risques potentiels des avancées imminentes de l'IA sans disposer d'un large corpus de preuves scientifiques. Néanmoins, la production de preuves sur les implications en matière de sûreté et de sécurité des tendances suggérées par o3 sera une priorité urgente pour la recherche en IA dans les semaines et les mois à venir.

Principales conclusions du rapport

- **Les capacités de l'IA à usage général, le type d'IA sur lequel se concentre ce rapport, ont augmenté rapidement ces dernières années et se sont encore améliorées ces derniers mois[1].** Il y a quelques années, les meilleurs grands modèles de langage (LLM) pouvaient rarement produire un paragraphe de texte cohérent. Aujourd'hui, l'IA à usage général peut écrire des programmes informatiques, générer des images photoréalistes personnalisées et s'engager dans des conversations ouvertes prolongées. Depuis la publication du rapport intermédiaire (mai 2024), de nouveaux modèles ont montré des performances nettement meilleures aux tests de raisonnement scientifique et de programmation.
- **De nombreuses entreprises investissent désormais dans le développement d'agents d'IA à usage général comme voie potentielle de progrès futurs.** Les agents d'IA sont des systèmes d'IA à usage général qui peuvent agir, planifier et déléguer de manière autonome pour atteindre des objectifs avec peu ou pas de supervision humaine. Des agents d'IA sophistiqués pourraient, par exemple, utiliser des ordinateurs pour mener à bien des projets plus longs que les systèmes actuels. Ceci débloquerait à la fois des bénéfices et des risques supplémentaires.
- **Les progrès en matière de capacités au cours des mois et des années à venir pourraient être lents ou extrêmement rapides.** Les progrès dépendront de la capacité des entreprises à déployer rapidement encore plus de données et de puissance de calcul pour entraîner de nouveaux modèles, et de l'efficacité de ce passage à l'échelle pour dépasser leurs limites actuelles. Selon des recherches récentes, il serait encore physiquement possible d'augmenter rapidement la taille des modèles pendant au moins plusieurs années. Mais les avancées majeures en matière de capacités des modèles peuvent aussi nécessiter d'autres facteurs : par exemple, des avancées scientifiques imprévisibles ou le succès d'une nouvelle méthode d'expansion récemment adoptée par les entreprises.
- **Plusieurs effets néfastes de l'IA à usage général sont déjà bien établis.** Il s'agit notamment d'escroqueries, d'images intimes sans consentement et de matériel pédopornographique, de sorties de modèles biaisées envers certains groupes de personnes ou certaines opinions, de problèmes de fiabilité et de violations de la vie privée. Les chercheurs ont développé des techniques d'atténuation pour ces problèmes, mais à ce jour, aucune combinaison de techniques ne peut les résoudre complètement. Depuis la publication du rapport intermédiaire, de nouveaux signes de discrimination liée aux systèmes d'IA à usage général ont révélé des formes de biais plus subtiles.
- **À mesure que l'IA à usage général devient plus performante, des signes de risques supplémentaires apparaissent progressivement.** Il s'agit notamment de risques tels que des répercussions à grande échelle sur le marché du travail, le piratage informatique par l'IA ou des attaques biologiques, et que la société perde le contrôle sur l'IA à usage général. Les experts interprètent différemment les signes existants sur ces risques : certains pensent que de tels risques ne se produiront que dans des décennies, tandis que d'autres pensent que l'IA à usage général pourrait entraîner des dommages à l'échelle de la société dans les prochaines années.

[1] Veuillez vous référer à la « Note du Président » sur les dernières avancées de l'IA suite à la rédaction de ce rapport.

Les avancées récentes sur les capacités d'IA à usage général, notamment dans les tests de raisonnement scientifique et de programmation, ont mis en évidence de nouveaux risques potentiels tels que le piratage informatique et les attaques biologiques, ce qui a conduit une grande entreprise d'IA à augmenter son évaluation du risque biologique de son meilleur modèle de « faible » à « moyen ».

- **Les techniques de gestion des risques sont naissantes, mais des progrès sont possibles.** Il existe différentes méthodes techniques permettant d'évaluer et de réduire les risques liés à l'IA à usage général que les développeurs peuvent utiliser et que les régulateurs peuvent exiger, mais elles ont toutes des limites. Par exemple, les techniques d'interprétabilité actuelles permettant d'expliquer pourquoi un modèle d'IA à usage général a produit un résultat donné restent très limitées. Toutefois, les chercheurs font des progrès pour remédier à ces limitations. Les chercheurs et les décideurs politiques tentent aussi, et de plus en plus, de normaliser les approches de gestion des risques et de les coordonner à l'échelle internationale.
- **Le rythme et l'imprévisibilité des progrès de l'IA à usage général posent un « dilemme de la preuve » aux décideurs politiques[2].** Compte tenu des avancées parfois rapides et inattendues, les décideurs politiques devront souvent évaluer les bénéfices et risques potentiels des avancées imminentes de l'IA sans disposer d'un large corpus de preuves scientifiques. Ce faisant, ils se trouvent confrontés à un dilemme. D'une part, les mesures préventives d'atténuation des risques fondées sur des preuves limitées pourraient s'avérer inefficaces ou inutiles. D'autre part, attendre des preuves plus solides d'un risque imminent pourrait laisser la société démunie, voire rendre toute atténuation impossible, par exemple si des bonds soudains dans les capacités de l'IA, et les risques qui y sont associés, se produisent. Les entreprises et les gouvernements développent des systèmes d'alerte précoce et des cadres de gestion des risques susceptibles de réduire ce dilemme. Certains d'entre eux déclenchent des mesures d'atténuation spécifiques lorsqu'il existe de nouvelles preuves de risques, tandis que d'autres exigent que les développeurs fournissent des preuves de sûreté avant de publier un nouveau modèle.
- **Il existe un large consensus parmi les chercheurs sur le fait que des avancées concernant les questions suivantes seraient utiles :** À quelle vitesse les capacités de l'IA à usage général progresseront-elles dans les années à venir, et comment les chercheurs peuvent-ils mesurer ces progrès de manière fiable ? Quels sont les seuils de risque raisonnables pour déclencher des mesures d'atténuation ? Comment les décideurs politiques peuvent-ils accéder au mieux aux informations sur l'IA à usage général qui sont pertinentes pour la sécurité publique ? Comment les chercheurs, les entreprises technologiques et les gouvernements peuvent-ils évaluer de manière fiable les risques liés au développement et au déploiement de l'IA à usage général ? Comment fonctionnent en interne les modèles d'IA à usage général ? Comment de l'IA à usage général peut-elle être conçue pour se comporter de manière fiable ?
- **Nous ne subissons pas l'IA : les choix faits par les gens déterminent son avenir.** L'avenir de la technologie d'IA à usage général est incertain, avec un large éventail de trajectoires semblant possibles même dans un avenir proche, allant d'impacts très positifs à des conséquences très

[2] Veuillez vous référer à la « Note du Président » sur les dernières avancées de l'IA suite à la rédaction de ce rapport.

négatives. Cette incertitude peut inciter au fatalisme et faire apparaître l'IA comme quelque chose que nous subissons. Mais ce seront les décisions des sociétés et des gouvernements sur la manière de gérer cette incertitude qui détermineront le chemin que nous emprunterons. Ce rapport vise à faciliter sur ces décisions une discussion constructive et fondée sur des données probantes.

Résumé exécutif

Le but de ce rapport

Ce rapport synthétise l'état des connaissances scientifiques sur l'IA à usage général, l'IA capable d'effectuer une grande variété de tâches, en mettant l'accent sur la compréhension et la gestion de ses risques.

Ce rapport résume les preuves scientifiques sur la sûreté de l'IA à usage général. Le but de ce rapport est de contribuer à créer une compréhension internationale et commune des risques suscités par l'IA avancée et de la manière dont ils peuvent être atténués. Pour y parvenir, ce rapport se concentre sur l'IA à usage général (ou l'IA capable d'effectuer une grande variété de tâches) car ce type d'IA a progressé particulièrement rapidement ces dernières années et a été largement déployé par les entreprises technologiques à des fins diverses, tant pour les consommateurs que pour les entreprises. Le rapport synthétise l'état de la compréhension scientifique de l'IA à usage général, en mettant l'accent sur la compréhension et la gestion de ses risques.

Dans un contexte de progrès rapides, la recherche sur l'IA à usage général se trouve actuellement dans une période de découverte scientifique et, dans de nombreux cas, n'est pas encore une science établie. Le rapport fournit un instantané de la compréhension scientifique actuelle de l'IA à usage général et de ses risques. Il s'agit notamment d'identifier les domaines de consensus scientifique et ceux dans lesquels il existe des points de vue divergents ou des lacunes dans la compréhension scientifique actuelle.

Les citoyens du monde entier ne pourront profiter pleinement et en toute sécurité des bénéfices potentiels de l'IA à usage général que si les risques qu'elle présente sont gérés de manière appropriée. Ce rapport se concentre sur l'identification de ces risques et l'évaluation des méthodes techniques pour les analyser et les atténuer, y compris l'utilisation de l'IA à usage général elle-même comme outil d'atténuation. Son objectif n'est pas d'évaluer de manière exhaustive tous les impacts sociétaux possibles de l'IA à usage général. Plus particulièrement, les bénéfices actuels et les potentiels futurs bénéfices de l'IA à usage général, bien que vastes, dépassent le cadre de ce rapport.

L'élaboration d'une politique globale nécessite de prendre en compte à la fois les bénéfices potentiels de l'IA à usage général et les risques abordés dans ce rapport. Il convient également de prendre en compte le fait que d'autres types d'IA présentent des profils de risque/bénéfice différents de ceux de l'IA à usage général actuelle.

Les trois principales sections du rapport résument les éléments scientifiques sur trois questions fondamentales : Que peut faire l'IA à usage général ? Quels sont les risques associés à l'IA à usage général ? Et quelles techniques d'atténuation existent contre ces risques ?

Section 1 – Capacités de l'IA à usage général : Que peut faire l'IA à usage général aujourd'hui et à l'avenir ?

Les capacités de l'IA à usage général se sont rapidement améliorées ces dernières années, et les avancées futures pourraient aussi bien être lentes qu'extrêmement rapides.

Les capacités de l'IA contribuent largement à bon nombre des risques qu'elle pose et, selon de nombreux indicateurs, les capacités de l'IA à usage général progressent rapidement. Il y a cinq ans, les principaux modèles de langue d'IA à usage général pouvaient rarement produire un paragraphe de texte cohérent. Aujourd'hui, certains modèles d'IA à usage général peuvent engager des conversations sur un large éventail de sujets, écrire des programmes informatiques ou générer de courtes vidéos réalistes à partir d'une description. Cependant, il est techniquement difficile d'estimer et de décrire de manière fiable les capacités de l'IA à usage général.

Les développeurs ont continuellement augmenté les ressources utilisées pour entraîner de nouveaux modèles, principalement du fait de leur « mise à l'échelle »[3]. Ils ont constamment accru les ressources utilisées pour entraîner de nouveaux modèles (un processus souvent appelé « mise à l'échelle ») tout en perfectionnant les méthodes existantes pour optimiser l'utilisation de ces ressources. Par exemple, d'après des estimations récentes, les modèles d'IA de l'état de l'art ont vu environ quadrupler chaque année leurs ressources de calcul (« compute ») pour l'apprentissage, tandis que la taille des jeux de données d'entraînement a été multipliée par 2,5.

Le rythme des progrès futurs des capacités de l'IA à usage général a des implications substantielles pour la gestion des risques émergents, mais les experts ne s'accordent pas sur ce à quoi il faut s'attendre, même dans les mois et années à venir. Les experts soutiennent diversement la possibilité que les capacités de l'IA à usage général progressent lentement, rapidement ou extrêmement rapidement.

Les experts ne s'accordent pas sur le rythme des progrès à venir, en raison de points de vue divergents sur la promesse apportée par davantage de « mise à l'échelle ». Par ailleurs, les entreprises explorent une nouvelle forme de mise à l'échelle qui pourrait encore accélérer les capacités. Bien que la mise à l'échelle ait souvent permis de surmonter les limites des systèmes précédents, les experts sont divisés sur sa capacité à résoudre les faiblesses résiduelles actuelles, comme le manque de fiabilité dans les interactions avec le monde réel ou dans la réalisation de tâches longues et complexes sur un ordinateur. Ces derniers mois, un nouveau type de mise à l'échelle a montré du potentiel pour améliorer encore les capacités : plutôt que de simplement

[3] Veuillez vous référer à la « Note du Président » sur les dernières avancées de l'IA suite à la rédaction de ce rapport.

augmenter les ressources utilisées pour l'apprentissage des modèles, les entreprises d'IA s'intéressent de plus en plus à la « mise à l'échelle de l'inférence » : permettre à un modèle déjà entraîné d'utiliser davantage de calculs pour résoudre un problème donné, par exemple pour améliorer sa propre solution, ou pour écrire ce que l'on appelle des « chaînes de pensée » qui décomposent le problème en étapes plus simples.

Plusieurs entreprises leaders qui développent de l'IA à usage général misent sur la « mise à l'échelle » pour continuer à améliorer les performances. Si les tendances récentes se poursuivent, d'ici la fin de 2026, certains modèles d'IA à usage général seront entraînés en utilisant environ 100 fois plus de ressources de calcul d'apprentissage que les modèles les plus gourmands en calcul de 2023, pour atteindre un coût d'apprentissage 10 000 fois plus élevé d'ici 2030, combiné à des algorithmes qui atteignent de plus grandes capacités pour une quantité donnée de calcul disponible. Outre cette mise à l'échelle potentielle des ressources d'apprentissage, les tendances récentes telles que la mise à l'échelle de l'inférence et l'utilisation de modèles pour générer des données d'entraînement pourraient signifier qu'encore plus de ressources de calcul seront utilisées dans l'ensemble. Il existe toutefois des obstacles potentiels à l'augmentation rapide des données et du calcul, notamment la disponibilité des données, des puces IA, du capital et de la capacité énergétique locale. Les entreprises qui développent de l'IA à usage général s'efforcent de surmonter ces obstacles potentiels.

Depuis la publication du Rapport Intermédiaire (mai 2024), l'IA à usage général a atteint des performances de niveau expert dans certains tests et compétitions de raisonnement scientifique et de programmation, et les entreprises ont déployé d'importants efforts pour développer des agents autonomes d'IA. Les progrès en science et en programmation ont été stimulés par des techniques de mise à l'échelle de l'inférence telles que l'écriture de longues « chaînes de pensée ». De nouvelles études suggèrent que poursuivre la mise à l'échelle de ces approches, par exemple en permettant aux modèles de résoudre des problèmes en élaborant des chaînes de pensées encore plus longues que les modèles d'aujourd'hui, pourrait favoriser des avancées supplémentaires dans des domaines où le raisonnement est crucial, comme les sciences, l'ingénierie logicielle et la planification. En plus de cette tendance, les entreprises déploient de gros efforts pour développer des agents d'IA à usage général plus avancés, capables de planifier et d'agir de manière autonome pour atteindre un objectif donné. Enfin, le prix du marché de l'utilisation d'IA à usage général avec un niveau de capacité donné a fortement chuté, ce qui rend cette technologie plus largement accessible et largement utilisée.

Ce rapport se concentre principalement sur les aspects techniques des progrès de l'IA, mais la vitesse à laquelle l'IA à usage général progressera n'est pas une question purement technique. Le rythme des avancées futures dépendra également de facteurs non techniques, notamment potentiellement des approches adoptées par les gouvernements pour réglementer l'IA. Ce rapport n'aborde pas la manière dont les différentes approches réglementaires pourraient affecter la vitesse de développement et d'adoption de l'IA à usage général.

Section 2 – Risques : Quels sont les risques associés à l'IA à usage général ?

Plusieurs effets néfastes de l'IA à usage général sont déjà bien établis. À mesure que l'IA à usage général devient plus performante, des signes de risques supplémentaires apparaissent progressivement.

Ce rapport classe les risques liés à l'IA à usage général en trois catégories : risques d'utilisation malveillante, risques de dysfonctionnement et risques systémiques. Chacune de ces catégories contient des risques qui se sont déjà matérialisés ainsi que des risques qui pourraient se matérialiser dans les quelques prochaines années.

***Risques d'utilisation malveillante :* les acteurs malveillants peuvent utiliser l'IA à usage général pour nuire aux individus, aux organisations ou à la société.** Les formes d'utilisation malveillante incluent :

- **Atteinte aux individus sur la base de contenus falsifiés :** Les acteurs malveillants peuvent actuellement utiliser l'IA à usage général pour générer du faux contenu qui nuit aux individus de manière ciblée. Ces utilisations malveillantes incluent la création de « deepfakes » pornographiques sans consentement, la génération de matériel pédopornographique par IA, la fraude financière par usurpation de voix, le chantage pour extorsion, le sabotage de réputations personnelles et professionnelles, ainsi que les abus psychologiques. Cependant, bien que les signalements de préjudices dus à des contenus fictifs générés par IA soient fréquents, il manque encore des statistiques fiables sur la fréquence de ces incidents.
- **Manipulation de l'opinion publique :** L'IA à usage général facilite la génération de contenu persuasif à grande échelle. Cela peut aider les acteurs qui cherchent à manipuler l'opinion publique, par exemple pour influencer les résultats politiques. Toutefois, les données probantes sur la prévalence et l'efficacité de ces efforts restent limitées. Les contre-mesures techniques, comme l'ajout de filigranes sur le contenu, bien qu'utiles, peuvent généralement être contournées par des acteurs moyennement sophistiqués.
- **Cybercriminalité :** L'IA à usage général peut permettre à des acteurs malveillants de différents niveaux de compétence de mener des cyberattaques plus facilement ou plus rapidement. Les systèmes actuels ont démontré leurs capacités dans les tâches de cybersécurité de faible et moyenne complexités, et des acteurs para-étatiques explorent activement l'IA pour examiner les systèmes cibles. De nouvelles recherches confirment que les capacités de l'IA à usage général en matière de cyberattaques progressent considérablement, mais on ignore encore si cela modifiera l'équilibre entre les cybermenaces et les mesures de protection.

- **Attaques biologiques et chimiques :** Les récents systèmes d'IA à usage général ont montré une certaine aptitude à fournir des instructions et des astuces pour reproduire des armes biologiques et chimiques connues et à faciliter la conception de nouveaux composés toxiques. Dans de nouvelles expériences visant à tester l'aptitude à générer des plans de production d'armes biologiques, un système d'IA à usage général a parfois obtenu de meilleurs résultats que des experts humains ayant accès à Internet. En réaction, une entreprise d'IA a augmenté l'évaluation du risque biologique de son meilleur modèle de « faible » à « moyen ». Cependant, les tentatives concrètes de développement de ces armes nécessitent tout de même des ressources et une expertise supplémentaires et importantes. Une évaluation complète des risques biologiques et chimiques est difficile, car une grande partie des recherches pertinentes sont classifiées.

Depuis la publication du rapport intermédiaire, l'IA à usage général est devenue plus performante dans les domaines pertinents pour des usages malveillants. Par exemple, des chercheurs ont récemment créé des systèmes d'IA à usage général capables de trouver et d'exploiter par eux-mêmes certaines vulnérabilités en matière de cybersécurité et, avec une aide humaine, de découvrir des vulnérabilités jusque-là inconnues dans des logiciels largement utilisés. Les capacités de l'IA à usage général en matière de raisonnement et d'intégration de différents types de données, qui peuvent contribuer à la recherche sur les pathogènes ou dans d'autres domaines à usage dual, se sont également améliorées.

***Risques dus à des dysfonctionnements :* L'IA à usage général peut également causer des dommages involontaires**. Même lorsque les utilisateurs n'ont pas l'intention de causer du tort, de graves risques peuvent survenir en raison du dysfonctionnement de l'IA à usage général. Ces dysfonctionnements incluent :

- **Problèmes de fiabilité :** L'IA à usage général actuelle n'est pas toujours fiable, ce qui peut porter préjudice. Par exemple, si les utilisateurs consultent un système d'IA à usage général pour obtenir des conseils médicaux ou juridiques, le système peut générer une réponse contenant des données erronées. Les utilisateurs ne sont souvent pas conscients des limites d'un produit d'IA, par exemple en raison d'une « maîtrise de l'IA » limitée, d'une publicité trompeuse ou d'une mauvaise communication. Il existe un certain nombre de cas connus de torts causés par des problèmes de fiabilité, mais les informations factuelles sur l'étendue exacte des différentes formes de ce problème sont encore limitées.
- **Biais :** Les systèmes d'IA à usage général peuvent amplifier les biais sociaux et politiques, causant ainsi des torts concrets. Ils présentent fréquemment des biais liés à la race, au genre, à la culture, à l'âge, au handicap, à l'opinion politique ou à d'autres aspects de l'identité humaine. Cela peut conduire à des résultats discriminatoires, notamment une répartition inégale des ressources, le renforcement des stéréotypes et la négligence systématique des groupes ou points de vue sous-représentés. Les approches techniques visant à atténuer les biais et la discrimination dans les systèmes d'IA à usage général progressent, mais se

heurtent à des compromis entre l'atténuation des biais et des objectifs concurrents tels que l'exactitude et la vie privée, ainsi qu'à d'autres défis.

- **Perte de contrôle :** Les scénarios de « perte de contrôle » sont des scénarios futurs hypothétiques dans lesquels un ou plusieurs systèmes d'IA à usage général se mettent à fonctionner en dehors du contrôle de quiconque, sans moyen clair pour reprendre le contrôle. Il existe un large consensus sur le fait que l'IA à usage général actuelle n'a pas les capacités nécessaires pour présenter ce risque. Cependant, l'opinion des experts sur la probabilité d'une perte de contrôle au cours des prochaines années varie considérablement : certains considèrent cela comme peu plausible, d'autres comme probable, et d'autres encore comme un risque de probabilité modeste qui mérite une attention particulière en raison de sa gravité potentielle élevée. Les recherches empiriques et mathématiques en cours font progressivement avancer ces débats.

Depuis la publication du rapport intermédiaire, de nouvelles recherches ont permis d'apporter de nouvelles connaissances sur les risques de biais et de perte de contrôle. Les indices de biais dans les systèmes d'IA à usage général se sont multipliés et des travaux récents ont détecté des formes supplémentaires de biais de l'IA. Des chercheurs ont observé de modestes avancées supplémentaires vers les capacités de l'IA qui sont probablement nécessaires pour que les scénarios de perte de contrôle couramment évoqués se produisent. Il s'agit notamment de capacités permettant d'utiliser des ordinateurs de manière autonome, de programmer, d'obtenir un accès non autorisé à des systèmes numériques et d'identifier des moyens d'échapper à la supervision humaine.

***Risques systémiques :* au-delà des risques directement posés par les capacités des modèles individuels, le déploiement généralisé de l'IA à usage général est associé à plusieurs risques systémiques plus larges.** Les exemples de risques systémiques vont des impacts potentiels sur le marché du travail aux risques pour la vie privée et aux effets environnementaux :

- **Risques pour le marché du travail :** L'IA à usage général, surtout si elle continue de progresser rapidement, a le potentiel d'automatiser un très large éventail de tâches, ce qui pourrait avoir un effet significatif sur le marché du travail. Cela signifie que de nombreuses personnes pourraient perdre leur emploi actuel. Toutefois, de nombreux économistes s'attendent à ce que les pertes d'emplois potentielles puissent être compensées, en partie, voire potentiellement totalement, par la création de nouveaux emplois et par une demande accrue dans les secteurs non automatisés.
- **Fossé mondial en R&D de l'IA :** La recherche et le développement (R&D) en matière d'IA à usage général sont actuellement concentrés dans quelques pays occidentaux et en Chine. Ce « fossé de l'IA » risque d'accroître considérablement la dépendance du monde à l'égard de ce petit groupe de pays. Certains experts s'attendent également à ce que cela contribue aux inégalités mondiales. Ce fossé a de nombreuses causes, dont un certain nombre ne sont

pas propres à l'IA. Cependant, cela résulte en grande partie des différents niveaux d'accès aux ressources de calcul très coûteuses qui sont nécessaires au développement d'IA à usage général : la plupart des pays à revenu faible et intermédiaire (LMIC) ont un accès aux ressources de calcul nettement inférieur à celui des pays à revenu élevé (HIC).

- **Concentration du marché et points de défaillance uniques :** Un petit nombre d'entreprises dominent actuellement le marché de l'IA à usage général. Cette concentration du marché pourrait rendre les sociétés plus vulnérables à plusieurs risques systémiques. Par exemple, si les organisations de secteurs critiques, tels que la finance ou la santé, s'appuient toutes sur un petit nombre de systèmes d'IA à usage général, une erreur ou une vulnérabilité dans un tel système pourrait provoquer des pannes et des perturbations simultanées à grande échelle.
- **Risques environnementaux :** L'utilisation croissante de ressources de calcul dans le développement et le déploiement d'IA à usage général a rapidement augmenté les quantités d'énergie, d'eau et de matières premières consommées dans la construction et l'exploitation de l'infrastructure de calcul nécessaire. Cette tendance ne montre aucun signe clair de ralentissement, malgré les progrès des techniques qui permettent d'utiliser les ressources de calcul plus efficacement. L'IA à usage général possède également un certain nombre d'applications qui peuvent soit bénéficier, soit nuire aux efforts de développement durable.
- **Risques pour la vie privée :** L'IA à usage général peut provoquer ou contribuer à des violations de la vie privée des utilisateurs. Par exemple, des informations sensibles contenues dans les données d'apprentissage peuvent être divulguées involontairement lorsqu'un utilisateur interagit avec le système. De plus, lorsque des utilisateurs partagent des informations sensibles avec le système, ces informations peuvent également être divulguées. Mais l'IA à usage général peut également faciliter des violations délibérées de la vie privée, par exemple si des acteurs malveillants utilisent l'IA pour déduire des informations sensibles sur des individus spécifiques à partir de grandes quantités de données. Cependant, jusqu'à présent, les chercheurs n'ont pas trouvé de preuve de violations généralisées de la vie privée associées à l'IA à usage général.
- **Violations des droits d'auteur :** L'IA à usage général apprend à partir des œuvres d'expression créative tout en en créant. Elle remet en question les systèmes traditionnels de consentement, de rémunération et de contrôle des données. La collecte de données et la génération de contenu peuvent impliquer une variété de lois sur les droits des données, qui varient selon les juridictions et peuvent faire l'objet de litiges actifs. Compte tenu de l'incertitude juridique entourant les pratiques de collecte de données, les entreprises d'IA partagent moins d'informations sur les données qu'elles utilisent. Cette opacité rend les recherches tierces sur la sûreté de l'IA plus difficiles.

Depuis la publication du rapport intermédiaire, des éléments supplémentaires sur les impacts de l'IA à usage général sur le marché du travail sont apparus, tandis que de nouveaux développements ont accru les préoccupations en matière de vie privée et de copyright. De nouvelles analyses des données du marché du travail suggèrent que les individus adoptent l'IA à usage général très rapidement par rapport aux technologies précédentes. Le rythme d'adoption par les entreprises varie considérablement selon les secteurs. De plus, les

avancées récentes en matière de capacités ont conduit à un déploiement croissant de l'IA à usage général dans des contextes sensibles tels que les services de santé ou la surveillance du lieu de travail, ce qui crée de nouveaux risques pour la vie privée. Enfin, alors que les conflits liés aux droits d'auteur s'intensifient et que les mesures techniques d'atténuation des violations du droit d'auteur restent peu fiables, les titulaires des droits des données ont rapidement restreint l'accès à leurs données.

***Modèles à poids ouverts* : un facteur important lors de l'évaluation de nombreux risques qu'un modèle d'IA à usage général pourrait présenter est la manière dont il est mis à disposition du public.** Les « modèles à poids ouverts » sont des modèles d'IA dont les composants centraux, appelés « poids », sont mis à disposition publiquement pour téléchargement. L'accès avec poids ouverts facilite la recherche et l'innovation, notamment en matière de sûreté de l'IA, tout en augmentant la transparence et en facilitant la détection des failles dans les modèles par la communauté scientifique. Toutefois, les modèles à poids ouverts peuvent aussi présenter des risques, par exemple en facilitant une utilisation malveillante ou malavisée qu'il est difficile, voire impossible, pour le développeur du modèle de surveiller ou d'atténuer. Une fois que les poids du modèle sont disponibles au téléchargement public, il n'existe aucun moyen de mettre en œuvre un retrait complet de toutes les copies existantes ou de garantir que toutes les copies existantes reçoivent des mises à jour de sûreté. Depuis le rapport intermédiaire, un consensus de haut niveau s'est dégagé selon lequel les risques posés par une plus grande ouverture de l'IA devraient être évalués en termes de risque « marginal » : la mesure dans laquelle la publication d'un modèle à poids ouverts augmenterait ou diminuerait un risque donné, par rapport aux risques posés par les alternatives existantes telles que les modèles fermés ou d'autres technologies.

Section 3 – Gestion des risques : Quelles techniques existent pour gérer les risques dus à l'IA à usage général ?

Plusieurs approches techniques peuvent aider à gérer les risques, mais dans de nombreux cas, les meilleures approches disponibles présentent encore des limites très importantes et ne disposent d'aucune estimation quantitative des risques ni de garanties qui sont disponibles dans d'autres domaines critiques pour la sécurité.

La gestion des risques (identification et évaluation des risques, puis atténuation et surveillance de ces derniers) est difficile dans le contexte de l'IA à usage général. Bien que la gestion des risques soit également très difficile dans de nombreux autres domaines, certaines caractéristiques de l'IA à usage général semblent créer des difficultés particulières.

Plusieurs caractéristiques techniques de l'IA à usage général rendent la gestion des risques dans ce domaine particulièrement difficile. Il s'agit, entre autres, de :

- **La gamme des usages et des contextes d'utilisation possibles des systèmes d'IA à usage général est inhabituellement large.** Par exemple, le même système peut être utilisé pour fournir des conseils médicaux, analyser les vulnérabilités d'un code informatique et générer des photos. Cela augmente la difficulté à anticiper de manière exhaustive les cas d'usage pertinents, d'identifier les risques ou de tester le comportement des systèmes dans des circonstances réelles pertinentes.
- **Les développeurs comprennent encore peu le fonctionnement de leurs modèles d'IA à usage général.** Ce manque de compréhension rend plus difficile la prévision des problèmes de comportement ainsi que l'explication et la résolution des problèmes connus une fois ceux-ci observés. La compréhension reste difficile à atteindre, principalement parce que les modèles d'IA à usage général ne sont pas programmés au sens traditionnel du terme. Au lieu de cela, ils sont entraînés : les développeurs d'IA mettent en place un processus d'entraînement qui implique un grand volume de données, et le résultat de ce processus d'entraînement est le modèle d'IA à usage général. Les rouages internes de ces modèles sont en grande partie impénétrables, y compris pour les développeurs de modèles. Les techniques d'explication des modèles et d'« interprétabilité » peuvent améliorer la compréhension par les chercheurs et les développeurs du fonctionnement des modèles d'IA à usage général, mais, malgré les progrès récents, cette recherche reste naissante.
- **Des agents d'IA de plus en plus performants, des systèmes d'IA à usage général capables d'agir, de planifier et de déléguer de manière autonome pour atteindre des objectifs, présenteront probablement de nouveaux défis importants pour la gestion des risques.** Les agents d'IA poursuivent généralement des objectifs de manière autonome en utilisant des logiciels généraux tels que des navigateurs Web et des outils de programmation. Actuellement, la plupart d'entre eux ne sont pas encore suffisamment fiables pour une utilisation généralisée, mais les entreprises déploient de gros efforts pour créer des agents d'IA plus performants et plus fiables, et ont fait des progrès ces derniers mois. Les agents d'IA deviendront probablement de plus en plus utiles, mais ils pourraient également exacerber un certain nombre de risques évoqués dans ce rapport et introduire des difficultés supplémentaires en matière de gestion des risques. Parmi ces nouveaux défis potentiels, on peut citer la possibilité que les utilisateurs ne sachent pas toujours ce que font leurs propres agents d'IA, la possibilité que les agents d'IA opèrent en dehors du contrôle de quiconque, la possibilité que des attaquants « détournent » des agents, et la possibilité que les interactions entre systèmes d'IA créent de nouveaux risques complexes. Les approches de gestion des risques associés aux agents commencent seulement à être développées.

En plus des facteurs techniques, plusieurs facteurs économiques, politiques et autres facteurs sociétaux rendent particulièrement difficile la gestion des risques dans le domaine de l'IA à usage général.

Le rythme des progrès de l'IA à usage général crée un « dilemme de la preuve » pour les décideurs[4]. La progression rapide des capacités permet à certains risques d'apparaître soudainement. Par exemple, le risque de tricherie académique grâce à l'IA à usage général est passé

d'un phénomène négligeable à une pratique répandue en l'espace d'un an. Plus un risque apparaît rapidement, plus il est difficile de le gérer de manière réactive et plus la préparation devient précieuse. Toutefois, tant que les preuves d'un risque demeurent incomplètes, les décideurs ne peuvent pas savoir avec certitude si le risque apparaîtra ou s'il est peut-être déjà apparu. Cela crée un dilemme : la mise en œuvre de mesures préventives ou d'atténuation précoces peut s'avérer inutile, mais attendre des preuves concluantes pourrait rendre la société vulnérable aux risques qui émergent rapidement. Les entreprises et les gouvernements développent des systèmes d'alerte précoce et des cadres de gestion des risques susceptibles de réduire ce dilemme. Certains d'entre eux déclenchent des mesures d'atténuation spécifiques lorsqu'il existe de nouvelles preuves de risques, tandis que d'autres exigent que les développeurs fournissent des preuves de sûreté avant de publier un nouveau modèle.

- **Il existe un écart d'information entre ce que les entreprises d'IA savent de leurs systèmes d'IA et ce que savent les gouvernements et les chercheurs en dehors de l'industrie.** Les entreprises ne partagent souvent que des informations limitées sur leurs systèmes d'IA à usage général, en particulier avant leur diffusion à grande échelle. Les entreprises invoquent un mélange de préoccupations commerciales et de sûreté comme raisons pour limiter le partage d'informations. Toutefois, ce manque d'information rend également plus difficile pour d'autres acteurs de participer efficacement à la gestion des risques, en particulier des risques émergents.
- **Autant les entreprises d'IA que les gouvernements sont souvent confrontés à une forte pression concurrentielle, ce qui peut les amener à considérer la gestion des risques moins prioritaire.** Dans certaines circonstances, la pression concurrentielle peut inciter les entreprises à investir moins de temps ou d'autres ressources dans la gestion des risques qu'elles ne le feraient autrement. De même, les gouvernements peuvent investir moins dans les politiques de soutien à la gestion des risques lorsqu'ils perçoivent des compromis à faire entre la concurrence internationale et la réduction des risques.

Néanmoins, il existe diverses techniques et cadres de travail pour gérer les risques dus à l'IA à usage général que les entreprises peuvent adopter et que les régulateurs peuvent exiger. Il s'agit notamment de méthodes d'identification et d'évaluation des risques, ainsi que de méthodes d'atténuation et de surveillance de ces risques.

- **L'évaluation des systèmes d'IA à usage général en termes de risques fait partie intégrante de la gestion des risques, mais les évaluations de risques actuelles sont extrêmement limitées.** Les évaluations existantes des risques de l'IA à usage général reposent principalement sur des « contrôles ponctuels », c'est-à-dire sur des tests du comportement de l'IA à usage général dans un ensemble de situations spécifiques. Cela peut aider à identifier les dangers potentiels avant de déployer un modèle. Cependant, les tests

[4] Veuillez vous référer à la « Note du Président » sur les dernières avancées de l'IA à la fin de la rédaction de ce rapport

existants passent souvent à côté des dangers et sur- ou sous-estiment les capacités et les risques de l'IA à usage général, car les conditions de test diffèrent du monde réel.

- **Pour que l'identification et l'évaluation des risques soient effectives, les évaluateurs doivent disposer d'une expertise considérable, de ressources et d'un accès suffisant aux informations pertinentes.** Une évaluation rigoureuse des risques dans le contexte de l'IA à usage général nécessite de combiner plusieurs approches d'évaluation. Ces analyses vont des analyses techniques des modèles et des systèmes eux-mêmes aux évaluations des risques possibles dus à certains modèles d'utilisation. Les évaluateurs ont besoin d'une expertise considérable pour mener à bien de telles évaluations. Pour réaliser des évaluations des risques complètes, ils ont souvent aussi besoin de plus de temps, d'un accès plus direct aux modèles et à leurs données d'entraînement, et de plus d'informations sur les méthodologies techniques utilisées que celles que fournissent généralement les entreprises développant de l'IA à usage général.
- **Des progrès ont été réalisés dans l'entraînement des modèles d'IA à usage général pour qu'ils fonctionnent de manière plus sûre, mais aucune méthode actuelle ne peut prévenir de manière fiable même des résultats ouvertement dangereux.** Par exemple, une technique appelée « apprentissage antagoniste » consiste à exposer délibérément des modèles d'IA à des exemples conçus pour les faire échouer ou mal se comporter pendant l'entraînement, dans le but de créer une résistance à de tels cas. Toutefois, les acteurs malveillants peuvent toujours trouver de nouveaux moyens (« attaques ») pour contourner ces mesures de protection avec un effort faible à modéré. Par ailleurs, des études récentes montrent que les méthodes d'entraînement actuelles, basées en grande partie sur des retours humains imparfaits, pourraient inciter involontairement les modèles à induire les humains en erreur sur des questions complexes, en rendant leurs erreurs plus difficiles à repérer. Améliorer la quantité et la qualité de ces retours constitue une piste de progrès, bien que de nouvelles techniques d'entraînement, utilisant l'IA pour détecter les comportements trompeurs, montrent également un potentiel prometteur.
- **La surveillance, qui consiste à identifier les risques et à évaluer les performances une fois qu'un modèle est déjà en utilisation, ainsi que diverses interventions visant à prévenir des actions nuisibles, peuvent améliorer la sûreté d'une IA à usage général après son déploiement auprès des utilisateurs.** Les outils actuels peuvent détecter le contenu généré par IA, suivre les performances du système et identifier les entrées/sorties potentiellement dangereuses, bien que des utilisateurs moyennement qualifiés puissent souvent contourner ces mesures de protection. Plusieurs couches de défense combinant des capacités de surveillance technique et d'intervention avec une supervision humaine améliorent la sûreté, mais elles peuvent entraîner des coûts et des retards. À l'avenir, des mécanismes intégrés au matériel pourraient aider les clients et les régulateurs à mieux surveiller les systèmes d'IA à usage général lors de leur déploiement et potentiellement à vérifier les accords internationaux. Cependant, de telles solutions fiables n'existent pas encore.
- **Plusieurs méthodes existent tout au long du cycle de vie de l'IA pour protéger la vie privée.** Cela comprend la suppression des informations sensibles des données

d'entraînement, des méthodes d'entraînement qui contrôlent la quantité d'informations apprises (comme la « confidentialité différentielle »), et des techniques qui permettent d'utiliser l'IA avec des données sensibles tout en rendant leur récupération quasiment impossible (comme le « calcul confidentiel » et d'autres technologies visant à renforcer la confidentialité). De nombreuses méthodes d'amélioration de la vie privée issues d'autres domaines de recherche ne sont pas encore applicables aux systèmes d'IA à usage général en raison des besoins en calcul des systèmes d'IA. Ces derniers mois, les méthodes de protection de la vie privée se sont développées pour répondre à l'utilisation croissante de l'IA dans des domaines sensibles, notamment les assistants pour smartphones, les agents d'IA, les assistants vocaux toujours à l'écoute et l'utilisation dans les soins de santé ou la pratique juridique.

Depuis la publication du rapport intermédiaire, les chercheurs ont fait de nouveaux progrès pour pouvoir expliquer pourquoi un modèle d'IA à usage général a produit un résultat donné. Être capable d'expliquer les décisions de l'IA pourrait aider à gérer les risques de dysfonctionnement allant des biais et inexactitudes factuelles à la perte de contrôle. De plus, des efforts croissants ont été déployés pour normaliser les approches d'évaluation et d'atténuation dans le monde entier.

Conclusion : Un large éventail de trajectoires sont possibles pour l'avenir de l'IA à usage général, et beaucoup dépendra de la manière dont les sociétés et les gouvernements agiront.

L'avenir de l'IA à usage général est incertain, avec un large éventail de trajectoires semblant possibles même dans un avenir proche, allant d'impacts très positifs à des conséquences très négatives. Mais rien dans l'avenir de l'IA à usage général n'est inévitable. Le développement de l'IA à usage général, les acteurs impliqués, les problèmes qu'elle est conçue pour résoudre, l'aptitude des sociétés à exploiter pleinement son potentiel économique, les bénéficiaires de ses avancées, les types de risques que nous encourons et les investissements que nous consacrons à la recherche sur leur gestion : ces questions et bien d'autres dépendent des choix que les sociétés et les gouvernements font aujourd'hui et feront à l'avenir pour façonner le développement de l'IA à usage général.

Afin de faciliter une discussion constructive sur ces décisions, ce rapport fournit un aperçu de l'état actuel de la recherche scientifique et des discussions sur la gestion des risques de l'IA à usage général. Les enjeux sont élevés. Nous sommes impatients de poursuivre cet effort.