Note de synthèse

À propos de ce rapport

- Ce document constitue la publication intermédiaire du premier « rapport scientifique international sur la sécurité de l'IA avancée ». Un groupe hétérogène de 75 experts de l'intelligence artificielle (IA) a contribué à ce rapport, dont un groupe consultatif d'experts internationaux désignés par 30 pays, l'Union européenne et les Nations Unies (ONU).
- Placés sous la direction du Président de ce rapport, les experts indépendants chargés de sa rédaction étaient collectivement libres d'en déterminer le contenu.
- À une époque de progrès sans précédent dans le développement de l'IA, cette première publication limite son champ d'analyse à un type d'IA dont l'avancée est particulièrement rapide depuis quelques années : l'IA à usage général, l'intelligence artificielle capable d'exécuter une vaste gamme de tâches. Parmi les avancées rapides, la recherche sur l'IA à usage général traverse actuellement une phase de découverte scientifique. Il ne s'agit pas encore d'une science établie.
- Le monde ne pourra tirer parti en toute sécurité des nombreux avantages potentiels de l'IA à usage général que si ses risques sont correctement gérés. Ce rapport se concentre sur l'identification de ces risques et sur l'évaluation des méthodes techniques qui permettront de les évaluer et de les atténuer. Son but n'est pas de fournir une évaluation exhaustive de tous les impacts sociétaux possibles liés à l'IA à usage général ou de ses nombreux avantages potentiels.
- Pour la première fois de notre histoire, ce rapport intermédiaire a rassemblé des experts désignés par 30 pays, l'UE, l'ONU et d'autres experts de référence mondiale, chargés de fournir une base commune fondée sur des données probantes susceptibles de servir de matière à discussion et à décision sur la sécurité de l'IA à usage général. Nous ne nous sommes toujours pas mis d'accord sur plusieurs questions, certaines mineures et d'autres majeures, relatives aux capacités, aux risques et à l'atténuation des risques de l'IA à usage général. Nous n'en considérons pas moins ce projet comme étant essentiel pour améliorer notre compréhension collective de cette technologie et de ses risques, mais aussi pour nous rapprocher d'un consensus et de mécanismes efficaces d'atténuation des risques, dans le but de faire en sorte que tout le monde puisse profiter en toute sécurité des avantages de l'IA à usage général. Les enjeux sont considérables. Nous avons hâte de poursuivre cet effort.

Grandes lignes de la note de synthèse

- Bien gouvernée, l'IA à usage général peut être appliquée au profit de l'intérêt général et pourrait aboutir à une amélioration du bien-être, de la prospérité et à de nouvelles découvertes scientifiques. Toutefois, utilisée en état de dérèglement ou avec malveillance, l'IA à usage général peut avoir des effets néfastes, notamment en cas de décisions subjectives, dans les scénarios de gros enjeux ou d'arnaque, de fake news ou de violation de la vie privée.
- Au fur et à mesure de l'avancée continue des capacités de l'IA à usage général, les risques tels que les impacts de grande envergure sur le marché du travail, le piratage ou

les attaques biologiques utilisant l'IA et la possibilité que la société perde le contrôle de l'IA à usage général pourraient apparaître. Notons toutefois que les chercheurs sont loin d'être unanimes quant à la probabilité de la survenance de ces scénarios. Les différentes opinions sur ces risques proviennent souvent d'attentes différentes sur les mesures que la société sera prête à prendre pour les limiter, sur l'efficacité de ces mesures et sur la rapidité avec laquelle les capacités de l'IA à usage général évolueront.

- L'incertitude quant au rythme des futurs progrès des capacités de l'IA à usage général est considérable. Certains experts pensent qu'un ralentissement du progrès est largement plus probable. D'autres en revanche, pensent qu'une progression extrêmement rapide est possible, voire probable.
- Les développeurs et autorités règlementaires peuvent respectivement recourir à et stipuler diverses méthodes techniques pour évaluer et réduire les risques de l'IA à usage général, mais toutes ont leurs limitations. À titre d'exemple, les techniques actuelles pour expliquer pourquoi les modèles d'IA à usage général produisent un résultat donné sont très limitées.
- L'avenir de la technologie de l'IA à usage général est incertain. Une vaste gamme de trajectoires semblent possibles même dans un avenir proche, avec à la fois des résultats très positifs et très négatifs. Mais aucun aspect de l'avenir de l'IA n'est inévitable. Les décisions des sociétés et gouvernements détermineront l'avenir de l'IA. Le but de ce rapport intermédiaire est d'alimenter une discussion constructive à propos de ces décisions.

Ce rapport synthétise l'état de la compréhension scientifique de l'IA à usage général – une IA capable d'exécuter toutes sortes de tâches – en se penchant tout particulièrement sur la compréhension et la gestion de ses risques.

Les capacités des systèmes qui utilisent l'IA progressent rapidement. Cette progression a mis en évidence les nombreuses opportunités que crée l'IA pour les entreprises, la recherche, les gouvernements et la vie privée. Elle a aussi mené à une plus grande prise de conscience des effets nuisibles actuels et des risques futurs associés à l'IA avancée.

Le rapport scientifique international sur la sécurité de l'IA avancée a pour but de faire un pas en avant vers une compréhension internationale commune des risques de l'IA et des moyens susceptibles de permettre de les atténuer. Cette première publication intermédiaire du rapport limite son champ d'analyse à un type d'IA dont les capacités ont évolué particulièrement rapidement, nommément l'IA à usage général, l'intelligence artificielle capable d'exécuter une vaste gamme de tâches.

Dans le contexte des avancées rapides, la recherche sur l'IA à usage général traverse actuellement une phase de découverte scientifique. Il ne s'agit pas encore d'une science établie. Le rapport fournit un instantané de la compréhension scientifique actuelle de l'IA à usage général et de ses risques. Il porte notamment sur l'identification des domaines de consensus scientifique et de ceux qui inspirent différentes opinions ou questions de recherche ouvertes.

Le monde ne pourra tirer parti en toute sécurité des nombreux avantages potentiels de l'IA à usage général que si ses risques sont correctement gérés. Ce rapport se concentre sur l'identification des risques associés à l'IA à usage général et sur l'évaluation des méthodes techniques susceptibles de permettre de les évaluer et de les atténuer, notamment en exploitant les avantages bénéfiques de l'IA à usage général elle-même pour en atténuer les risques. Son but n'est pas d'évaluer de manière exhaustive tous les impacts sociétaux possibles de l'IA à usage général, voire de ses avantages potentiels.

Les capacités de l'IA à usage général augmentent rapidement depuis quelques années selon de nombreuses métriques. En outre, l'opinion quant à la manière d'en prédire la progression future n'est pas unanime. Dans un tel contexte, toutes sortes de scénarios peuvent paraître possibles.

Un grand nombre de métriques montrent que les capacités de l'IA à usage général augmentent rapidement. Il y a cinq ans, les modèles de langage de l'IA à usage général de référence parvenaient à peine à produire un paragraphe cohérent de texte. Aujourd'hui, certains modèles d'IA à usage général sont capables de participer à des conversations multitours sur une vaste gamme de sujets, d'écrire de petits programmes informatiques ou de générer des vidéos à partir d'une description. Les capacités de l'IA à usage général n'en sont pas moins difficiles à estimer avec fiabilité ou à définir précisément.

Le rythme de l'avancée de l'IA à usage général dépend d'une part de celui des avancées technologiques et de l'autre, de l'environnement de réglementation. Ce rapport se concentre sur les aspects technologiques de la question. Il ne fournit pas matière à discussion sur la manière dont les efforts de règlementation pourraient avoir une incidence sur la vitesse de développement et de déploiement de l'IA à usage général.

Les développeurs d'IA font rapidement progresser les capacités de l'intelligence artificielle à usage général depuis quelques années, surtout en augmentant continuellement les ressources d'entraînement de nouveaux modèles (tendance baptisée « scaling ») et en peaufinant les algorithmes existants. À titre d'exemple, les modèles d'IA de pointe ont fait l'objet annuellement d'une multiplication quasiment par quatre des ressources computationnelles (« puissance de calcul ») accordées à l'entraînement, par 2,5 de la taille des ensembles de données d'entraînement et par 1,5 de l'efficacité des algorithmes (performance par rapport à la puissance de calcul). Le « scaling » a-t-il abouti à un progrès par rapport aux enjeux fondamentaux comme le raisonnement causal ? La question continue de faire couler beaucoup d'encre parmi les chercheurs.

Le rythme de la progression future des capacités de l'IA à usage général joue un rôle substantiel dans la gestion des risques émergents, mais les experts ne s'entendent pas dans leurs prévisions quant à ce qu'elles pourraient nous apporter dans un avenir proche. Les experts soutiennent diversement la possibilité que les capacités de l'IA à usage général progresseront lentement, rapidement ou extrêmement rapidement. Cette divergence part d'une

question clé : le « scaling » continu des ressources et le peaufinage des techniques existantes suffiront-ils pour produire des progrès rapides et résoudre les problématiques relatives notamment à la fiabilité et à l'exactitude factuelle ou encore, aura-t-on besoin de nouvelles percées de la recherche pour faire suffisamment avancer les capacités de l'IA à usage général ?

Plusieurs entreprises de référence spécialisées dans le développement de l'IA à usage général parient que le « scaling » continuera de mener à de meilleures performances. Si les tendances récentes se confirment, d'ici à la fin de 2026, certains modèles d'IA à usage général seront entraînés avec 40 à 100 fois plus de puissance de calcul que les modèles les plus intensifs en calcul publiés en 2023, dans un contexte où les méthodes d'entraînement qui les exploitent seront entre 3 et 20 fois plus efficaces. Toutefois, des goulets d'étranglement pourraient freiner l'augmentation des données et des capacités de calcul, notamment en termes de disponibilité des données, de puces d'IA, de dépenses d'investissement et de capacité locale en énergie. Les entreprises qui développent l'IA à usage général œuvrent pour élargir la sortie de ces goulets d'étranglement potentiels.

Plusieurs initiatives de recherche tentent de comprendre et d'évaluer l'IA à usage général de manière plus fiable, mais notre compréhension globale du fonctionnement des modèles et systèmes d'IA à usage général est limitée.

Les approches de la gestion des risques de l'IA à usage général sont souvent fondées sur la supposition que les développeurs d'IA et décideurs ont les moyens d'évaluer les capacités et impacts potentiels des modèles et systèmes d'IA à usage général. Mais si les méthodes techniques peuvent faciliter les évaluations, les méthodes existantes présentent toutes des limitations et ne peuvent fournir aucune garantie solide contre la plupart des effets nuisibles liés à l'IA à usage général. Globalement, la compréhension scientifique des rouages internes, capacités et impacts sociétaux de l'IA à usage général est très limitée ; les experts s'entendent largement pour admettre que l'amélioration de notre compréhension de cette technologie doit figurer parmi nos priorités. Les difficultés suivantes figurent parmi les principaux défis à relever :

- Les développeurs n'ont encore qu'une compréhension limitée du fonctionnement de leurs modèles d'IA à usage général. En effet, les modèles d'IA à usage général ne sont pas programmés au sens traditionnel du terme. Ils sont au contraire « entraînés » : les développeurs d'IA définissent un processus d'entraînement à base de grands volumes de données. Le résultat de ce processus d'entraînement est le modèle d'IA à usage général. Ces modèles peuvent être composés de plusieurs billions de composantes appelées paramètres et la plupart de leurs rouages internes sont impénétrables, même pour le développeur de modèles. Les techniques d'explication et d'interprétabilité des modèles peuvent aider les chercheurs et développeurs à mieux comprendre la manière dont fonctionnent les modèles d'IA à usage général, mais cette recherche n'en est encore qu'à ses balbutiements.
- L'évaluation de l'IA à usage général s'effectue surtout en testant le modèle ou le système par rapport à diverses données « input ». Ces contrôles ponctuels sont utiles

pour évaluer les points forts et faibles, vulnérabilités et capacités potentiellement nuisibles comprises, mais ne fournissent toutefois aucune garantie de sécurité quantitative. Les tests passent souvent à côté de dangers, surestiment ou sous-estiment les capacités parce que le comportement des systèmes d'IA à usage général peut changer en fonction des circonstances, des utilisateurs ou parce que leurs composants ont fait l'objet de nouveaux réglages.

- Des acteurs indépendants peuvent, en principe, contrôler les modèles ou systèmes d'IA à usage général développés par une société. Toutefois souvent, les sociétés ne fournissent pas aux contrôleurs indépendants le niveau nécessaire d'accès direct aux modèles ou aux informations sur les données et méthodes utilisées et dont dépend une évaluation rigoureuse. Plusieurs gouvernements commencent à renforcer leurs capacités d'exécution d'évaluations et de contrôles techniques.
- L'impact sociétal en aval d'un système d'IA à usage général est difficile à évaluer. En effet, l'évaluation du risque n'a pas encore fait l'objet d'une recherche suffisante pour produire des méthodologies d'évaluation rigoureuses et complètes. De plus, le fait que les cas d'utilisation de l'IA à usage général soient très variés, souvent non prédéfinis et seulement légèrement limités complique d'autant l'évaluation du risque. Comprendre les impacts sociétaux potentiels en aval des modèles et systèmes d'IA à usage général requiert une analyse nuancée et multidisciplinaire. Augmenter la représentation de perspectives diverses dans les processus de développement et d'évaluation de l'IA à usage général est un défi technique et institutionnel continu.

L'IA à usage général peut gravement menacer la sécurité et le bien-être individuels et publics.

Ce rapport classe les risques associés à l'IA à usage général dans trois catégories : risques d'utilisation malveillante, risques consécutifs à un dysfonctionnement et risques systémiques. Il aborde également plusieurs facteurs transversaux qui contribuent à de nombreux risques.

Utilisation malveillante. Comme toutes les technologies puissantes, les systèmes d'IA à usage général peuvent être utilisés de manière malveillante pour causer du tort. Les types d'utilisation malveillante possible vont des usages relativement bien documentés, comme les arnaques basées sur l'IA à usage général, aux usages qui, selon certains experts pourraient se produire au cours des prochaines années, comme l'utilisation malveillante des capacités scientifiques de l'IA à usage général.

- Porter préjudice à des personnes en recourant à de fausses informations générées par l'IA à usage général est une classe d'utilisation malveillante de cette technologie relativement bien documentée. L'IA à usage général peut être exploitée pour augmenter l'envergure et la sophistication des arnaques et fraudes, comme c'est notamment le cas des tentatives d'hameçonnage enrichies par l'IA à usage général. L'IA à usage général peut aussi servir pour générer de fausses informations compromettantes mettant en scène des personnes à leur insu. C'est le cas de la pornographie non-consensuelle par contrefaçons numériques ou « deepfakes ».

- L'utilisation malveillante de l'IA à usage général à des fins de désinformation et de manipulation de l'opinion public compte aussi parmi les domaines préoccupants de cette technologie. L'IA à usage général et d'autres technologies modernes facilitent la génération et la dissémination de la désinformation, y compris en vue de troubler les processus politiques. Bien qu'utiles, les contre-mesures techniques comme le tatouage numérique de contenus peuvent habituellement être contournées par des acteurs modérément sophistiqués.
- L'IA à usage général peut aussi être utilisée de manière malveillante à des fins de cyberinfraction, augmentant la cyberexpertise des individus et facilitant l'exécution par des utilisateurs malveillants de cyberattaques efficaces. Les systèmes d'IA à usage général peuvent servir à amplifier et partiellement automatiser certains types de cyberopérations comme les attaques d'ingénierie sociale, par exemple. Toutefois l'IA à usage général peut aussi être exploitée à des fins de cyberdéfense. Globalement, aucune preuve évidente ne permet pour l'instant de suggérer que l'IA à usage général est capable d'automatiser des tâches sophistiquées de cybersécurité.
- Certains experts ont également exprimé leur crainte que l'IA à usage général puisse être utilisée pour soutenir le développement et l'utilisation malveillante des armes et notamment, des armes biologiques. Aucune preuve solide ne permet d'affirmer que les systèmes d'IA à usage général actuels posent ce risque. Par exemple, même si ces systèmes d'IA à usage général manifestent des capacités croissantes dans le domaine de la biologie, les études limitées disponibles ne prouvent pas clairement qu'ils puissent renforcer les capacités des acteurs malveillants au point de leur permettre d'obtenir des pathogènes biologiques plus facilement que sur l'Internet. Notons toutefois que les menaces de grande envergure futures ont peu été évaluées et sont difficiles à exclure.

Risques consécutifs aux dysfonctionnements. Même lorsque les utilisateurs n'ont aucune intention de causer du tort, des risques graves peuvent émerger consécutivement à un dysfonctionnement de l'IA à usage général. Ces dysfonctionnements peuvent avoir plusieurs causes et conséquences.

- Peut-être que les utilisateurs comprennent mal la fonctionnalité des produits basés sur les modèles et systèmes d'IA à usage général, consécutivement à une mauvaise communication ou à la publicité trompeuse, par exemple. Cette situation peut être préjudiciable si les utilisateurs déploient alors les systèmes de manière inappropriée ou à des fins inadaptées.
- Le biais des systèmes d'IA est un problème généralement bien documenté et qui reste tout aussi irrésolu pour l'IA à usage général. Les données « output » ou résultats de l'IA à usage général peuvent être subjectifs par rapport à des caractéristiques protégées comme la race, le sexe, la culture, l'âge et le handicap. Cette subjectivité peut être génératrice de risques, notamment dans des domaines à forts enjeux comme les soins de santé, le recrutement et les prêts financiers. En outre, un grand nombre de modèles d'IA à usage général largement utilisés sont surtout entraînés sur des données qui représentent de manière disproportionnée les cultures occidentales, d'où le risque

d'intensifier le potentiel de tort subi par les personnes insuffisamment représentées par ces données.
- Les scénarios de « perte de contrôle » sont des scénarios potentiels futurs dans lesquels la société ne parviendrait plus à restreindre suffisamment les systèmes d'IA à usage général, quand bien même leur capacité de causer du tort ne ferait plus aucun doute. L'opinion générale estime que l'IA à usage général actuelle n'a pas les capacités nécessaires pour poser ce risque. Pour certains experts, les efforts faits actuellement pour développer les systèmes d'IA à usage général *autonomes* – systèmes capables d'agir, de planifier et de poursuivre des objectifs – pourraient, s'ils donnent leurs fruits, mener à la perte de contrôle. Les experts sont en désaccord quant au degré de plausibilité des scénarios de perte de contrôle, pour définir quand ils pourraient se produire et à quel point ils seraient difficiles à atténuer.

Risques systémiques. Le développement et l'adoption généralisés de la technologie de l'IA à usage général posent plusieurs risques systémiques, des impacts potentiels sur le marché du travail aux risques pour la vie privée et aux effets sur l'environnement :

- L'IA à usage général, surtout si elle maintient son rythme d'avancée rapide, pourrait automatiser un très large éventail de tâches capables d'avoir un effet considérable sur le marché du travail. Un grand nombre de personnes pourrait perdre leur travail actuel. Cependant un grand nombre d'économistes prévoient que les pertes potentielles de travail pourraient être en partie, voire entièrement, compensées par la création de nouveaux emplois et par l'accroissement de la demande dans les secteurs non-automatisés.
- La recherche et le développement de l'IA à usage général sont actuellement concentrés dans quelques pays occidentaux et en Chine. Bien que multicausal, ce « fossé de l'IA » découle en partie de différents niveaux d'accès à la puissance de calcul nécessaire pour développer l'IA à usage général. Étant moins bien placés que les pays à haut revenu et les sociétés de technologie en termes d'accès à la puissance de calcul, les pays à faible revenu et les institutions académiques sont défavorisés.
- La concentration du marché en découlant par rapport au développement de l'IA à usage général accroît la vulnérabilité des sociétés face à plusieurs risques systémiques. C'est ainsi que, par exemple, l'utilisation généralisée d'un petit nombre de systèmes d'IA à usage général dans des secteurs critiques comme la finance ou les soins de santé, pourrait provoquer des pannes et perturbations simultanées à grande échelle au niveau de ses secteurs interdépendants, à cause de bogues ou de vulnérabilités par exemple.
- L'exploitation croissante des calculs dans le développement et le déploiement de l'IA à usage général a rapidement augmenté la consommation d'énergie associée à l'IA à usage général. Cette tendance, qui ne donne aucun signe de déclin, pourrait encore augmenter les émissions de CO_2 et la consommation d'eau.
- Les modèles et systèmes d'IA à usage général peuvent présenter un risque pour la vie privée. La recherche montre par exemple que l'utilisation de données « input » antagonistes permet aux utilisateurs d'extraire d'un modèle des données d'entraînement contenant des informations sur des personnes. Pour les prochains modèles entraînés à

partir de données à caractère personnel sensibles comme les données sur la santé ou financières, cette possibilité pourrait mener à des fuites particulièrement graves de données privées.
- S'agissant du développement de l'IA à usage général, les violations de droits d'auteur potentielles posent un défi aux lois sur la propriété intellectuelle traditionnelles, ainsi qu'aux systèmes de consentement, de compensation financière et de contrôle applicables aux données. Les régimes de droits d'auteur flous dissuadent les développeurs d'IA à usage général de déclarer quelles données ils utilisent ; ils ne rassurent pas les créateurs quant aux protections dont ils bénéficieraient en cas d'utilisation sans consentement de leurs œuvres pour entraîner les modèles d'IA à usage général.

Facteurs de risque transversaux. Plusieurs facteurs de risque transversaux caractéristiques de l'IA à usage général qui augmentent la probabilité ou la gravité non seulement d'un, mais de plusieurs risques, sous-tendent les risques associés à cette technologie.

- Citons parmi les facteurs de risque transversaux techniques, la difficulté de veiller à ce que les systèmes d'IA à usage général se comportent systématiquement comme il se doit, notre manque de compréhension de leurs rouages internes et le développement continu d'« agents » de l'IA à usage général capables d'agir de manière autonome dans des conditions de supervision limitée.
- La disparité potentielle entre le rythme des progrès technologiques et celui de la réponse réglementaire ou encore, les incitations concurrentielles susceptibles de pousser les développeurs de l'IA à usage général à publier leurs produits rapidement, potentiellement au détriment d'une gestion du risque méticuleuse, comptent parmi les facteurs de risque transversaux sociétaux.

Plusieurs approches techniques peuvent contribuer aux efforts visant à atténuer les risques, mais aucune méthode connue ne procure actuellement des assurances ou garanties solides contre les torts associés à l'IA à usage général.

Bien que ce rapport n'aborde pas les interventions de politiques d'atténuation des risques liés à l'IA à usage général, il se penche sur les méthodes d'atténuation des risques techniques dans lesquelles les chercheurs progressent. Malgré ces progrès, les méthodes actuelles n'ont pas prouvé leurs capacités de servir de garde-fou fiable, même lorsque les données résultats de l'IA à usage général sont manifestement préjudiciables dans le monde réel. Plusieurs approches techniques sont utilisées pour évaluer et atténuer les risques :

- L'entraînement pour améliorer la sécurité des modèles d'IA à usage général fait quelques progrès. D'autre part, les développeurs entraînent les modèles d'IA à usage général pour les rendre plus résistants aux données « input » conçues pour les faire échouer (« entraînement antagoniste »). Ceci n'empêche néanmoins pas les adversaires de trouver normalement des données « input » de substitution capables de réduire l'efficacité des mécanismes de protection moyennant des efforts faibles à

modérés. Limiter les capacités d'un système d'IA à usage général à un cas d'utilisation spécifique peut aider à réduire les risques découlant de pannes imprévues ou d'usages malveillants.

- Plusieurs techniques permettent d'identifier les risques, d'inspecter les actions d'un système et d'évaluer la performance post-déploiement d'un système d'IA à usage général. Le terme « surveillance » est souvent employé pour qualifier ces pratiques.
- L'atténuation des biais des systèmes d'IA à usage général peut être traitée du début à la fin du cycle de vie du système et notamment aux phases de conception, d'entraînement, de déploiement et d'utilisation. Néanmoins, la prévention totale des biais des systèmes d'IA à usage général n'est pas chose facile, dans la mesure où elle requiert un la collecte systématique des données d'entraînement, une évaluation continue et l'identification efficace de ces biais. Elle peut également obliger à abandonner l'équité en faveur d'autres objectifs comme la précision et la confidentialité, à décider ce qui constitue des connaissances utiles ou des biais indésirables à tenir à l'écart des données résultats.
- La protection de la vie privée compte parmi les domaines actifs de la recherche et du développement. Le simple fait de réduire au minimum l'utilisation des données à caractère personnel sensibles dans l'entraînement fait partie des approches susceptibles de réduire considérablement les risques d'atteintes à la vie privée. Toutefois lorsque les données sensibles, intentionnellement ou involontairement, sont utilisées, les outils techniques existants pour réduire les risques d'atteintes à la vie privée ont du mal à s'adapter aux vastes modèles d'IA à usage général. Ils n'ont pas nécessairement les moyens de fournir aux utilisateurs des mécanismes de contrôle adéquats.

Conclusion : les trajectoires possibles de l'IA à usage général sont légion. Leur choix dépendra largement de la manière dont agiront les sociétés et gouvernements dans ce domaine.

L'avenir de l'IA à usage général est incertain, dans un contexte où de nombreuses trajectoires paraissent possibles, même dans un avenir proche, avec des résultats parfois très positifs, parfois très négatifs. Mais aucun aspect de l'avenir de l'IA à usage général n'est inévitable. Comment l'IA à usage général sera développée et par qui, quels problèmes sera-t-elle conçue pour résoudre, les sociétés pourront-elles en extraire pleinement le potentiel économique, qui en bénéficiera, à quels types de risques nous exposons-nous et quelles sommes investirons-nous dans la recherche pour en atténuer les risques ? Ces questions, comme un tas d'autres dépendent des choix que font les sociétés et gouvernements aujourd'hui et qu'ils feront demain pour façonner le développement de l'IA à usage général.

Soucieux de faciliter les discussions constructives sur ces décisions, ce rapport donne un aperçu de l'état actuel de la recherche scientifique et des débats sur la gestion des risques de l'IA à usage général. Les enjeux sont considérables. Nous avons hâte de poursuivre cet effort.