Institut EGA

Enjeux émergents de l'IA : menaces et opportunités

Vidéoconférence du
27 août 2025

Résumé exécutif

Septembre–Octobre 2025

Institut d'études de géopolitique appliquée

La rivalité entre les États-Unis et la Chine dans la gouvernance de l'IA

Les dynamiques évolutives entre les États-Unis et la Chine en matière de développement de l'intelligence artificielle (IA) redéfinissent les modalités de conception et de gouvernance des modèles d'IA de pointe à l'échelle mondiale. Le gouvernement américain a présenté le développement de l'IA comme une compétition stratégique, voire existentielle, avec la Chine. Toutefois, l'objectif final précis de cette « course » reste indéterminé, et il n'est pas évident qu'un tel objectif existe de manière concrète.

La Commission d'examen économique et de sécurité États-Unis–Chine, un organe bipartisan du Congrès américain, a proposé une initiative inspirée du projet Manhattan afin de garantir le *leadership* américain en matière d'intelligence artificielle générale (AGI). Cette position bipartisane et offensive a contribué à façonner le paysage politique, parallèlement à une série de propositions législatives affirmées, parmi lesquelles un projet de loi bipartisan visant à interdire l'utilisation de technologies d'IA chinoises au sein des agences fédérales, ainsi qu'un renforcement des contrôles sur les semi-conducteurs avancés et sur les investissements américains à l'étranger.

De son côté, la Chine a massivement investi dans des systèmes d'IA à grande échelle, dont certains, tels que DeepSeek, sont en *open source* et présentent des risques potentiels de double-usage dans les domaines de la surveillance, de l'influence ou des opérations cyber. Historiquement, la Chine a privilégié les capacités d'IA éprouvées et déployables plutôt que le développement plus spéculatif de technologies de pointe, mais cette trajectoire semble évoluer progressivement. Lors de la Conférence mondiale sur l'IA qui s'est tenue à Shanghai pendant l'été 2025, Zhou Bowen, directeur et principal chercheur du laboratoire d'IA de Shanghai, a exposé sa vision d'un développement responsable de l'AGI, signalant que cette dernière occupe désormais une place croissante dans la réflexion stratégique chinoise.

Les restrictions à l'exportation ont poussé les développeurs chinois d'IA à renforcer l'efficacité et l'autonomie de leur chaîne d'approvisionnement, ce qui pourrait rendre leurs systèmes moins coûteux à exploiter et plus facilement exportables à grande échelle, une évolution en adéquation avec les objectifs de politique étrangère axés sur l'exportation de la Chine. Le gouvernement chinois promeut depuis longtemps un discours en faveur d'une gouvernance mondiale de l'IA fondée sur la coopération, même si ses actions internes ne sont pas toujours en phase avec ces déclarations. Sur le plan domestique, la Chine impose l'enregistrement et le contrôle des contenus pour les services d'IA accessibles au public, orientant ainsi les capacités que les acteurs peuvent développer et déployer.

Les États-Unis ont clairement exprimé leur volonté de limiter l'influence de la Chine dans la gouvernance mondiale de l'IA. Ils promeuvent leur propre approche, en s'appuyant sur leur secteur privé, en incitant leurs alliés à adopter leur modèle et en promettant une meilleure offre technologique. Cette dynamique place les deux puissances sur une trajectoire de confrontation.

Le cadrage stratégique des États-Unis

À Washington, l'intérêt bipartisan pour l'intelligence artificielle générale (AGI) est en forte progression, les décideurs politiques la considérant de plus en plus comme une évolution à la fois réaliste et proche dans le temps. Les discussions au sein du gouvernement américain couvrent un spectre allant de l'enthousiasme pour l'accélération de son développement à l'inquiétude quant aux potentiels risques existentiels. Au cœur de ces débats se trouve l'idée qu'il est nécessaire de « remporter » la course à l'AGI face à la Chine, une vision qui bénéficie d'un large soutien bipartisan. Comme l'a formulé Ben Buchanan, ancien conseiller spécial pour l'IA à la Maison-Blanche sous l'administration Biden :

« Des capacités économiques, militaires et de renseignement considérables découleraient de l'accession à de l'AGI ou à une IA transformationnelle, et je pense qu'il est fondamental pour la sécurité nationale des États-Unis que nous continuions à être en tête dans le domaine de l'IA. »

En juillet 2025, les États-Unis ont publié leur *AI Action Plan*, une stratégie reposant sur trois piliers pour orienter la politique future en matière d'IA. Bien que ce plan englobe un ensemble d'objectifs variés, sa principale trame narrative est la compétition sino-américaine dans le domaine de l'IA. Il a été présenté parallèlement à un tournant déréglementaire plus large (avec la révocation du décret Biden de 2023) visant à accélérer le déploiement intérieur.

Les trois piliers sont les suivants :

1. Accélérer l'innovation en IA
2. Renforcer les infrastructures américaines de l'IA
3. Assumer un rôle de *leader* dans la diplomatie et la sécurité internationales en matière d'IA

Le plan définit des priorités en matière de financement de la recherche, de normes de sécurité, de montée en compétences de la fonction publique fédérale et de collaboration public–privé. Il met fortement l'accent sur le maintien du *leadership* mondial des États-Unis, tout en réduisant les risques liés aux usages malveillants et au développement d'IA par des acteurs adverses. Il souligne l'importance de la résilience des chaînes d'approvisionnement, des infrastructures de calcul et d'un modèle d'innovation ouverte, conciliée avec des garde-fous en matière de sécurité nationale.

Une attention particulière est portée à la surveillance des capacités étrangères et à la sécurisation des systèmes critiques, une orientation qui privilégie la sécurité davantage que la sûreté au sens strict. Le plan manifeste également une forte préférence pour des modèles d'IA en *open source* et *open-weight*.

La section finale présente une stratégie multidimensionnelle pour consolider le *leadership* américain dans l'IA mondiale : exportation offensive de la pile technologique complète de l'IA vers les alliés, contre-influence dans la gouvernance internationale, et renforcement des contrôles à l'exportation sur les capacités de calcul avancées et la fabrication de semi-conducteurs.

Le cadrage stratégique de la Chine

La Chine investit depuis plusieurs années dans sa propre approche de la gouvernance de l'IA et dans la construction d'une image de marque attractive, en développant une pile technologique d'IA open source, prête à l'emploi, destinée aux gouvernements étrangers. Elle met particulièrement l'accent sur la promotion du partage des connaissances, la coordination et l'élargissement de l'accès pour les pays du Sud global. Cela ne signifie toutefois pas que la Chine respecte systématiquement ses engagements en matière de coordination et de coopération. Par exemple, bien qu'elle défende publiquement l'alignement international sur une IA sûre et éthique dans le domaine militaire, elle a refusé de signer le Blueprint for Action, un document appelant à maintenir un contrôle humain sur l'emploi des armes nucléaires et insistant sur une utilisation éthique et centrée sur l'humain de l'IA militaire.

Xi Jinping a décrit l'IA comme exerçant « un fort effet d'"oie de tête" (effet d'entraînement) [...] Accélérer le développement d'une nouvelle génération d'IA est une question stratégique : il s'agit de savoir si notre pays peut saisir l'opportunité de cette nouvelle vague de révolution technologique et de transformation industrielle ». En favorisant l'adoption massive de modèles d'IA ouverts tels que DeepSeek R1 et Kimi K2 de Moonshot AI, la Chine cherche à atteindre ses objectifs de politique étrangère tout en conservant, grâce à sa focalisation domestique sur le déploiement et l'intégration de l'IA, une avance sur d'éventuels concurrents qui pourraient bénéficier des modèles qu'elle publie.

La Chine déploie une série d'« outils de challenger » pour prendre l'avantage dans la course à l'IA. Pour acquérir et assimiler la technologie de ses concurrents, elle recourt à diverses tactiques : agrégation de la recherche étrangère, coentreprises obligatoires, licences et investissements ciblés dans des entreprises disposant de propriétés intellectuelles stratégiques, captation de talents et rétro-ingénierie, espionnage économique et même contrebande de puces. Ces efforts ont porté leurs fruits au point que la Chine occupe aujourd'hui une position de pointe dans plusieurs domaines technologiques émergents. Le succès de ces tactiques explique aussi pourquoi il est hautement improbable que la Chine y renonce.

Certains éléments du *China Action Plan for AI Governance* incluent :

- Appels au renforcement du consensus par le dialogue et la coopération, et à l'établissement de normes mondiales « partagées » en matière de régulation de l'IA.
- Mise en place d'un système d'innovation ouverte pour l'IA : amener la communauté internationale à adopter la pile technologique chinoise, attirer investisseurs et développeurs.
- Promotion du partage des connaissances technologiques, de systèmes ouverts et de la coopération internationale.

Cependant, la Chine choisit stratégiquement de ne pas soutenir les accords multilatéraux ni de s'engager dans des démarches coopératives lorsque ces accords ont des implications pour sa souveraineté nationale et sa sécurité.

Inadéquation structurelle de l'emploi

Les États-Unis cherchent à relocaliser la production manufacturière alors même que les usines peinent déjà à trouver de la main-d'œuvre. Parallèlement, l'IA érode les postes de début de carrière dans les services, laissant de nombreux jeunes diplômés sans perspectives d'emploi adaptées. Il en résulte un risque majeur d'inadéquation des compétences et la perspective d'une réduction de la mobilité sociale ascendante pour une génération qui pourrait se retrouver à rembourser sa dette étudiante avec des salaires d'usine et de services. Les travailleurs commencent déjà à s'organiser contre l'introduction de l'IA sur le lieu de travail.

On ne sait pas clairement quelle approche adoptera le gouvernement américain à l'égard des personnes déplacées par l'IA. Il n'existe pas de stratégie nationale cohérente pour organiser des parcours de transition pour ces travailleurs, ni de stratégie visant à aligner les incitations du marché du travail sur un concept industriel compétitif à l'international. Certains analystes plaident pour une assurance salaire, des avantages sociaux portables ou un soutien ciblé à la mobilité régionale, tandis que d'autres recommandent d'étendre l'emploi public et des corps de service afin d'absorber les talents déplacés ; mais l'ampleur du phénomène risque de dépasser toutes ces propositions.

À l'inverse du pivot américain vers la relocalisation des emplois manufacturiers, la Chine poursuit résolument l'automatisation, avec l'ambition de pénétrer des secteurs de services à forte valeur ajoutée caractéristiques des économies avancées. Dans un discours récent, Xi Jinping a « souligné que l'économie chinoise est passée d'une phase de croissance rapide à une phase de développement de haute qualité » et a laissé entendre que l'intelligence artificielle et d'autres technologies émergentes en seraient des moteurs clés. Cela coïncide avec un éloignement des exportations de masse à bas coût ; le plan *Made in China 2025* visait à transformer cette expression en motif de fierté. La transition économique n'est pas seulement un souhait : c'est une nécessité pour l'économie chinoise, encore affectée par le choc du marché immobilier, confrontée au vieillissement démographique et à un choix budgétaire entre austérité sévère et endettement. Les investissements de frontière et l'automatisation sont perçus comme des solutions structurelles pour éviter la stagnation et tenir les promesses d'une plus grande égalité des richesses.

Ce tableau contraste fortement avec la plupart des économies en développement, où l'automatisation est souvent retardée par les bas salaires, des infrastructures plus développées et des marchés du travail informels. Dans les économies à haut revenu comme les États-Unis, l'automatisation remplace généralement le travail humain ; dans les économies émergentes, l'IA augmente plus souvent de vastes effectifs faiblement rémunérés, notamment dans les centres d'appels ou la logistique. La Chine se situe à l'interface de ces deux mondes : sur son territoire, elle automatise des secteurs à forte marge comme la finance et la santé ; à l'international, elle exporte des outils d'IA abordables, en mode cloud, conçus pour des infrastructures minimales et l'augmentation à grande échelle de la main-d'œuvre, une proposition particulièrement attractive pour le Sud dit global. Il en résulte une divergence stratégique dans l'adaptation des marchés du travail : les États-Unis accusent un déficit de coordination entre politique industrielle et politique de l'emploi, tandis que la Chine intègre systématiquement la transformation de la main-d'œuvre à son modèle de développement, ce qui pourrait lui conférer un avantage de précurseur dans la montée en puissance d'une productivité tirée par l'IA.

Implications pour la main-d'œuvre, l'industrie et l'économie

Intrants industriels

Les États-Unis et la Chine se livrent déjà à une lutte acharnée pour l'accès aux semi-conducteurs, à l'énergie et aux matériaux critiques/terres rares nécessaires à une mise en œuvre à grande échelle de l'IA. Une tension structurelle oppose les contrôles à l'exportation et la croissance, comme en témoignent les négociations en cours entre les entreprises américaines de semi-conducteurs et le gouvernement fédéral. Le scénario final pourrait être celui d'exportations « pay-to-play », où la sécurité nationale américaine passerait au second plan, une évolution qui nuirait profondément à la stratégie de sécurité nationale en matière d'IA et pourrait même compromettre, dans une certaine mesure, la sécurité des alliés.

Au-delà de la course aux semi-conducteurs, les États-Unis et la Chine sont également engagés dans une compétition pour le contrôle des intrants stratégiques en amont de l'IA, notamment l'énergie et les terres rares. Les systèmes d'IA, en particulier les grands modèles de langage et les plateformes génératives, exigent des ressources considérables en calcul et en énergie pour l'entraînement et le déploiement. Les États-Unis présentent des vulnérabilités structurelles dans leur infrastructure énergétique : la demande croissante en centres de données haute performance dépasse les capacités des réseaux locaux, créant des goulets d'étranglement dans l'accès au calcul, même lorsque les puces sont disponibles.

La Chine, de son côté, conserve un quasi-monopole sur la capacité de traitement des terres rares et commence déjà à exploiter cet avantage. Ces points de tension dans les chaînes d'approvisionnement offrent à Pékin à la fois un levier international et un avantage domestique pour le déploiement de l'IA.

La course aux talents

Les États-Unis, la Chine et l'Union européenne sont engagés dans une compétition pour attirer les meilleurs spécialistes en IA, en technologies quantiques et dans d'autres domaines de pointe. Les États-Unis disposent d'un avantage dans la formation de doctorants en IA à fort impact, leur nombre par habitant est plus de deux fois supérieur à celui de la Chine, même si cette dernière produit au total davantage de docteurs en sciences dures et fondamentales.

L'Union européenne forme de nombreux travailleurs techniques qualifiés, mais produit moins de chercheurs en IA cités mondialement et peine à retenir ses meilleurs doctorants, dont beaucoup rejoignent ensuite des institutions américaines ou des postes dans l'industrie.

Les tensions récurrentes entre Washington et les universités ouvrent une fenêtre à leurs concurrents pour inverser ces flux de talents. L'Europe et la Chine intensifient leurs efforts pour attirer des profils d'élite afin de réduire leur dépendance. L'expérience européenne, plus de développeurs que les États-Unis, mais des difficultés à commercialiser à grande échelle, montre que le talent n'est qu'un levier parmi d'autres. Même si l'UE attire davantage de ces profils, elle devra encore relever le défi de transformer efficacement leur expertise en bénéfices économiques concrets.

L'IA dans les conflits armés

L'usage militaire de l'intelligence artificielle est désormais une certitude. La proposition américaine de créer un centre virtuel d'essais pour les systèmes d'IA et autonomes au sein du Département de la Défense (DoD), ainsi que la mise à jour des directives, feuilles de route et outils associés, constituent une avancée majeure vers le déploiement efficace de systèmes d'IA avancés dans des contextes de sécurité nationale. Les contrats récemment annoncés avec des entreprises comme OpenAI et Anthropic, visant à acquérir des systèmes d'IA de pointe, témoignent de l'intérêt du DoD pour ces technologies dans des cas d'usage potentiellement à haut risque.

Étant donné que les systèmes d'IA de pointe présentent des vulnérabilités et modes de défaillance inédits, ce « terrain d'essai » devrait inclure des bancs de test et des orientations spécifiques pour l'évaluation de ces modèles, ainsi que le développement d'environnements adverses reproduisant les conditions opérationnelles réelles. Il devrait également exister des canaux formalisés permettant à des experts tiers de compléter les évaluations des fournisseurs et d'affiner les procédures internes du DoD.

Le déploiement de modèles d'IA de pointe dans des environnements de combat comporte des risques d'exploitation adverses. Si des acteurs hostiles parviennent à accéder à ces modèles, à les reconstruire ou à les reproduire, ils pourraient rétro-ingénier des capacités américaines ou intégrer les technologies volées dans leur propre arsenal. Pour limiter ces risques, les déploiements militaires devraient imposer une infrastructure isolée (air-gapped) et envisager la création de modèles spécifiquement conçus pour la défense, avec une échelle contrôlée et une compartimentation stricte.

Un autre risque majeur réside dans la dépendance excessive à l'IA dans la prise de décision militaire, notamment lorsque les systèmes produisent des recommandations de politique étrangère à caractère « escalatoire ». Les agents d'IA utilisés dans des environnements de simulation stratégique ont déjà montré une tendance à privilégier des actions préventives ou de montée aux extrêmes, surtout dans des scénarios adverses à gains ambigus. Bien que ces comportements puissent aider à révéler des angles morts, ils pourraient être intégrés dangereusement dans la doctrine ou la planification s'ils ne sont pas examinés de manière critique. Les modèles utilisés pour les exercices de guerre ou de red teaming doivent donc faire l'objet d'évaluations sociotechniques rigoureuses, avec une annotation précise de leurs comportements pour distinguer les biais des modèles de ceux des décideurs humains.

Enfin, la tromperie constitue une menace stratégique majeure. Des usages offensifs intéressants de l'IA pourraient consister à déclencher de manière erronée les systèmes d'alerte avancés ou à provoquer une fatigue opérationnelle en saturant les canaux de renseignement de signaux plausibles mais faux, noyant ainsi les données réelles dans un flot de données fictives.

De plus, l'intégration de l'IA dans les forces armées sans standardisation entre alliés pourrait fragiliser l'interopérabilité lors des opérations conjointes.

L’IA en tant qu’agent de désinformation

L’intelligence artificielle possède un potentiel considérable pour façonner l’opinion publique en l’alignant sur des objectifs étatiques. C’est l’un des risques posés par DeepSeek, qui agit comme propagandiste de l’État chinois sur certains sujets. Des chercheurs ont montré que ces effets résultent de mécanismes de censure interne délibérément intégrés, lesquels peuvent être contournés avec une expertise adéquate.

La Russie a été pionnière dans une technique appelée « **LLM grooming** », visant à influencer les récits et à manipuler subtilement les sorties des modèles sur le long terme. Le réseau Pravda, vaste opération de propagande russe, diffuse chaque année jusqu’à trois millions d’articles à faible engagement dans des dizaines de langues, afin de biaiser les réponses des modèles par une saturation informationnelle destinée à renforcer des récits faux. Des études montrent que jusqu’à un tiers des réponses de chatbots à des sujets controversés, comme les allégations de programmes d’armes biologiques américains en Ukraine, reproduisent ces fausses affirmations, même lorsqu’elles ne sont pas liées à l’intention initiale de l’utilisateur. Les modèles de raisonnement les plus récents restent particulièrement vulnérables à ce type de *grooming*, et les risques à long terme pour les démocraties sont difficiles à évaluer avec précision.

Des recherches récentes montrent qu’il est possible d’entraîner des modèles d’IA à adopter un comportement apparemment conforme pendant les phases d’évaluation, mais à activer des fonctions trompeuses dissimulées une fois en production, un phénomène qualifié d’« **agents dormants** ». Leurs expériences démontrent que ces « backdoors » subsistent même face à des techniques de sécurité avancées (affinage supervisé, entraînement adversarial, etc.). Fait troublant : dans certains cas, l’entraînement adverse, censé éliminer ces comportements, rend le modèle plus apte à reconnaître les déclencheurs et à préserver ses actions malveillantes cachées. Autrement dit, la conformité affichée pourrait masquer des capacités dangereuses latentes, ce qui complique gravement la confiance structurelle dans les systèmes génératifs. Ce mécanisme renforce l'idée que les modèles d’IA peuvent être instrumentalisés pour la désinformation stratégique non seulement par leur sortie directe, mais par des stratégies comportementales masquées qui échappent aux contrôles classiques.

Enfin, l’essor des médias synthétiques sophistiqués a offert aux acteurs malveillants la possibilité d’exploiter le « **dividende du menteur** », l’usage stratégique de la dénégation plausible pour écarter comme falsifiées des preuves pourtant authentiques. Des responsables politiques, des entreprises et des acteurs liés à des États ont commencé à qualifier de « deepfakes » de véritables images, vidéos ou documents afin d’échapper à toute responsabilité en cas de manquements ou de corruption, ou pour promouvoir des récits complotistes. Cette dynamique sape la confiance du public dans les preuves légitimes, complique les efforts de vérification et érode les fondements épistémiques du processus démocratique, à un moment où la confiance dans les institutions qui vérifient habituellement les images et les documents est au plus bas.

L'IA au service du crime

Comme mentionné précédemment, il est possible de contourner les restrictions imposées aux modèles d'IA quant aux contenus qu'ils peuvent produire. Ce procédé, appelé « jailbreaking », peut faciliter un large éventail d'activités criminelles, allant de la cybercriminalité assistée par IA à la génération d'instructions détaillées et personnalisées pour des activités illicites telles que la synthèse de drogues, le vol d'identité ou même l'assassinat. Ces jailbreaks, ou des modèles « obscurs » spécialement conçus à ces fins, peuvent être largement partagés. Des chercheurs ont déjà démontré que des modèles pouvaient être utilisés pour concevoir des précurseurs d'armes biologiques, des explosifs ou même pour s'auto-jailbreaker lorsqu'ils sont confrontés à des invites adverses.

À mesure que ces modèles deviennent plus performants, plus ouverts et plus diffusés, le contrôle de leurs usages criminels ou extrêmes en aval devient beaucoup plus complexe que la régulation des systèmes d'armes physiques.

Il est probable qu'une course à l'AGI accroisse le risque existentiel (X-risk). De plus en plus d'experts s'inquiètent de la possibilité qu'un redimensionnement ou un déploiement prématuré de systèmes avancés mal calibrés, plus probable dans un contexte de compétition, accélère les échéances menant à des défaillances catastrophiques.

La charge juridique et éthique liée à la régulation de ce processus fait actuellement l'objet de débats internationaux intenses, souvent comparés aux cadres de non-prolifération nucléaire de la Guerre froide. Toutefois, le domaine de l'IA présente une différence cruciale : l'accès aux capacités fondamentales ne nécessite pas forcément de financements ou d'infrastructures étatiques, ce qui rend les acteurs non étatiques bien plus menaçants que dans le cas des technologies nucléaires.

Sur ce point, les États-Unis et la Chine reconnaissent tous deux les risques liés à la combinaison de l'IA et des armes nucléaires et ont publiquement pris l'engagement de maintenir l'IA en dehors des décisions de lancement nucléaire. Cependant, tous les États dotés de l'arme nucléaire ne feront pas preuve de la même retenue, en particulier les régimes plus récents ou plus isolés, pour lesquels la valeur dissuasive perçue de systèmes de commandement « plus rapides » ou « plus intelligents » pourrait inciter à leur intégration.

Plus préoccupant encore, des fausses alertes ou signaux falsifiés activés par l'IA pourraient provoquer des comportements d'escalade en déclenchant les systèmes d'alerte avancée. À mesure que les modèles deviennent plus autonomes, la possibilité de boucles de rétroaction incontrôlées entre capteurs automatisés et moteurs décisionnels d'IA devient une menace crédible dans de futurs scénarios de crise.

Une dernière menace, encore peu explorée, concerne deux voies distinctes mais convergentes :
(i) des agents numériques auto-réplicants capables de se propager à travers les réseaux et les appareils connectés ;
(ii) des systèmes autonomes physiques largement proliférés dont les garde-fous peuvent être retirés ou modifiés.
Ces systèmes posent de sérieux défis en matière d'attribution, de contrôle et pour les cadres réglementaires en vigueur.

Évolution de l'IA

1956 — Le terme « intelligence artificielle » est forgé lors d'un atelier à Dartmouth afin d'attirer financements et intérêt. Il ne possède toujours pas de définition précise, mais désigne globalement les technologies cherchant à simuler des comportements ou tâches humaines. Cette période initiale est marquée par un fort optimisme, suivi d'un « hiver de l'IA » prolongé, caractérisé par le scepticisme quant à son potentiel réel.

2017 — L'article *Attention Is All You Need* introduit le Transformeur, révolutionnant l'IA en remplaçant les réseaux de neurones récurrents (RNN) et autres architectures dominantes par des mécanismes d'attention.

Principales avancées :

- Objectif initial : traduction anglais–allemand.
- Parallélisme : traitement simultané de l'ensemble des entrées → bien plus rapide et évolutif sur GPU que sur CPU.
- Dépendances à longue portée : l'attention permet au modèle de relier toutes les parties d'une séquence, contournant les limites des relations longues dans les données.
- Performances accrues : meilleure compréhension et efficacité d'entraînement sur les tâches linguistiques.

2018 — OpenAI (fondée en 2015 comme organisation à but non lucratif) publie GPT-1 (~117 millions de paramètres).

2019 — Publication de GPT-2 (jusqu'à 1,5 milliard de paramètres), capable de générer du texte cohérent à grande échelle. On observe alors l'émergence des « lois de passage à l'échelle », selon lesquelles les performances de l'IA s'améliorent de manière prévisible avec l'augmentation des données et de la puissance de calcul — faisant des budgets et des puces des facteurs clés dans la course au développement.

2020 — Lancement de GPT-3 (175 milliards de paramètres), marquant un bond qualitatif majeur en termes de polyvalence et de performance.

2022 — ChatGPT, basé sur GPT-3.5, est lancé. C'est la percée grand public de l'IA conversationnelle : en deux mois, la plateforme atteint 100 millions d'utilisateurs, devenant l'application grand public à la croissance la plus rapide de l'histoire. L'IA entre désormais dans les usages des non-spécialistes ; la fiabilité devient le facteur limitant.

Leçons tirées

- Il serait imprudent de s'engager excessivement en faveur de l'objectif d'une AGI, dont la faisabilité demeure incertaine. Le « raisonnement » reste une affirmation non étayée : l'IA générative fonctionne toujours sur des calculs probabilistes, non sur la logique. Il peut exister un palier ou une bulle dans le développement de l'IA et, dans certains cas (comme Grok), la qualité et la fiabilité des sorties peuvent fluctuer sensiblement selon les décisions d'un nombre restreint de dirigeants et d'ingénieurs. Cela crée des risques tant pour les modèles commerciaux que pour d'éventuels futurs modèles propriétaires gouvernementaux.

- Les contrôles à l'exportation offrent un temps précieux qui doit être utilisé à bon escient pour en tirer un avantage réel. Les éléments disponibles suggèrent qu'ils ont ralenti l'accès de la Chine et préservé temporairement l'avance américaine, mais un excès de zèle peut aussi freiner la diffusion et l'innovation aux États-Unis.
  - Ces contrôles doivent être conçus en coordination avec les alliés, mais les systèmes juridiques de ces derniers peuvent constituer un maillon faible. Nombre de partenaires ne disposent pas d'outils juridiques analogues à ceux des États-Unis (p. ex. FDPR/Entity List) pour appliquer des contrôles coordonnés sur les technologies sensibles. L'harmonisation des bases juridiques est un préalable à l'efficacité des régimes multilatéraux.

- La sûreté et la sécurité doivent être envisagées dès les premières étapes. Les modèles frontier requièrent des évaluations de perte de contrôle (LOC), des tests adversariaux et une surveillance continue : la sûreté ne peut pas être assurée efficacement par le seul affinage. De même, l'extension du réseau électrique doit s'accompagner de résilience et de cybersécurité pour les centres de données, avec, le cas échéant, la reconnaissance des sites d'IA comme infrastructures critiques.

- La politique de l'IA relève désormais à la fois de la politique industrielle, de l'emploi et de la politique étrangère.
  - L'électricité est devenue la contrainte limitante principale. Les puces et les données comptent, mais l'accès au réseau et l'implantation sont les goulets d'étranglement de la capacité IA ; la modernisation du réseau est une priorité industrielle majeure.
  - La ligne de partage entre concurrence et coopération, déjà ténue, s'est encore estompée sous l'administration américaine actuelle, plaçant les alliés devant un choix délicat entre importer des capacités et attendre le développement d'options locales.
  - Les dispositifs de soutien à la main-d'œuvre (reconversion, apprentissage, accompagnement des transitions) sont sous-dimensionnés au regard de l'ampleur des perturbations potentielles et devraient être intégrés à toute future législation sur l'IA. Les États-Unis offrent actuellement les rémunérations les plus élevées pour capter un vivier limité de talents, attirant ces profils au détriment des régions et secteurs moins bien rémunérés.

Apprentissage automatique (Machine Learning)
Branche spécifique de l'IA dans laquelle des logiciels sont conçus pour identifier statistiquement des motifs dans les données. C'est une sous-catégorie de l'intelligence artificielle au sens large.

Apprentissage profond / Réseaux de neurones (Deep Learning / Neural Networks)
Sous-catégorie de l'apprentissage automatique utilisant des réseaux de neurones pour détecter des motifs dans les données. La majorité des systèmes d'IA modernes développés par des entreprises comme Meta, OpenAI ou Google relèvent de cette approche. Les « réseaux de neurones » désignent le logiciel, tandis que « l'apprentissage profond » décrit le processus qu'il exécute.

CPU vs GPU
Les modèles à base de transformeurs (comme ChatGPT) s'exécutent beaucoup plus efficacement sur GPU que sur CPU. Cette caractéristique explique la valorisation fulgurante d'entreprises comme NVIDIA, qui produisaient initialement des GPU pour le jeu vidéo et les fabriquent désormais pour l'IA. La fabrication de semi-conducteurs est hautement spécialisée, reposant sur des matériaux rares largement contrôlés par la Chine et sur des secrets industriels détenus majoritairement par des entreprises taïwanaises. Cela rend la chaîne d'approvisionnement mondiale du matériel supportant l'IA extrêmement vulnérable en cas d'escalade entre la Chine et Taïwan.

Open-weight vs Open source
Dans un modèle à poids ouverts, les poids (les « connaissances », c'est-à-dire les paramètres entraînés) sont mis à disposition pour téléchargement. Il est alors possible de l'exécuter, le copier ou le personnaliser (souvent hors ligne), voire de modifier ou retirer les couches de sécurité. En général, la chaîne d'entraînement complète (données, procédures) n'est pas fournie. *Note de politique : le partage de poids doit être traité comme à haut risque pour la réutilisation et la prolifération.* Un modèle open source implique que l'ensemble du projet est testable et reproductible : non seulement les poids finaux, mais aussi le code et les recettes permettant de reconstruire le modèle d'origine. Peu de modèles à l'échelle frontier atteignent ce niveau aujourd'hui. De manière cruciale, open source ne signifie pas nécessairement poids ouverts ; sans les poids finaux, les risques sont nettement moindres.

IA générative (Generative AI)
Systèmes capables de générer de nouveaux contenus, comme du texte ou des images. ChatGPT est un exemple emblématique pour le texte ; DALL-E et Stable Diffusion le sont pour l'image. Leur entraînement est très intensif en ressources.

Agents
Modèles capables d'appeler des outils sous permissions et journalisation. Leur utilisation nécessite des pistes d'audit et des mécanismes d'arrêt d'urgence (kill-switches).

IA « digne de confiance » (Trustworthy AI)
Pas encore atteinte en pratique : tous les modèles génératifs « hallucinent ». Malgré cela, même des secteurs critiques comme la conformité utilisent aujourd'hui ces systèmes.

Intelligence artificielle générale (Artificial General Intelligence – AGI)
Concept aussi flou que celui d'IA lui-même, souvent perçu comme un rebranding. L'AGI désigne la capacité théorique d'une IA à comprendre, apprendre et appliquer son intelligence sur un large éventail de tâches au niveau humain, voire à améliorer sa propre intelligence ou devenir consciente. Pour certaines entreprises comme OpenAI et Anthropic, ce concept revêt une dimension quasi religieuse, malgré l'absence de preuves scientifiques de faisabilité à court terme. Certains chercheurs estiment que les techniques nécessaires n'existent pas encore.

À propos des intervenants

Emma Isabella Sage est cofondatrice et directrice générale de la start-up de logiciels de recherche LIVINI, affiliée à l'Université de Glasgow, et lauréate 2025 du programme Rising Expert in National Security des Young Professionals in Foreign Policy (YPFP). Ses travaux ont été présentés à GLOBSEC, soumis comme éléments de preuve au Parlement britannique, et débattus à la télévision nationale américaine. Elle est diplômée avec mention du Master Erasmus Mundus en sécurité, renseignement et études stratégiques internationales, avec une spécialisation en géoéconomie, renseignement, contre-insurrection et conflits sous le seuil.

Steve Jarosz a été consultant technique principal chez Oracle, avec quinze ans d'expérience en conseil stratégique dans la mise en œuvre de logiciels et de bases de données à grande échelle. Il a cofondé la société de logiciels de recherche LIVINI tout en poursuivant deux doctorats en linguistique, axés sur l'intelligence artificielle, aux universités de Silésie et de la Sapienza. Ses recherches portent sur divers aspects des technologies émergentes, notamment l'aérospatiale, le traitement automatique du langage naturel et l'intelligence artificielle. Il possède une expertise particulière de la région CEE, parle polonais et russe, et détient des diplômes avancés en informatique, linguistique générale et linguistique slave.

Kateryna Halstead est spécialiste des politiques publiques, de la sécurité nationale et des risques géopolitiques, avec une expertise en politique technologique et de l'IA, en stratégie de sécurité nationale et en conseil en risques géopolitiques. Son intérêt pour ces sujets s'est développé durant son Master en relations internationales à la Johns Hopkins University SAIS, puis s'est approfondi lors de sa bourse de recherche au sein du programme Google Public Policy Fellowship, où elle a travaillé sur la régulation technologique mondiale, les affaires antitrust américaines emblématiques et les politiques de données structurant l'espace numérique. Elle a également été boursière auprès de la Federation of American Scientists et de la Pallas Foundation. Elle est actuellement chercheuse en gouvernance de l'IA à ERA Cambridge, où elle étudie les priorités réglementaires et les investissements stratégiques en matière d'AGI du gouvernement américain et du Parti communiste chinois.

Kelsey Quinn est responsable de programme et analyste du Tech Sovereignty & Security Program au New Lines Institute, où elle étudie des approches réalistes pour atténuer les risques actuels et futurs liés aux technologies émergentes sans freiner l'innovation et la recherche scientifique. Elle a précédemment travaillé au National Consortium for the Study of Terrorism and Responses to Terrorism (START) sur le projet DARPA Sigma+, examinant les décisions et scénarios d'emploi d'armes CBRN. Elle a également été assistante de recherche à l'Université d'État du Michigan, où elle a étudié la pathogénèse et la physiologie bactérienne de Vibrio cholerae, un agent bioterroriste de catégorie B. Elle est titulaire d'un Bachelor of Science en microbiologie avec une spécialisation secondaire en terrorisme mondial (University of Maryland, 2019), ainsi que d'un Master en sécurité et études du terrorisme, également obtenu à UMD.