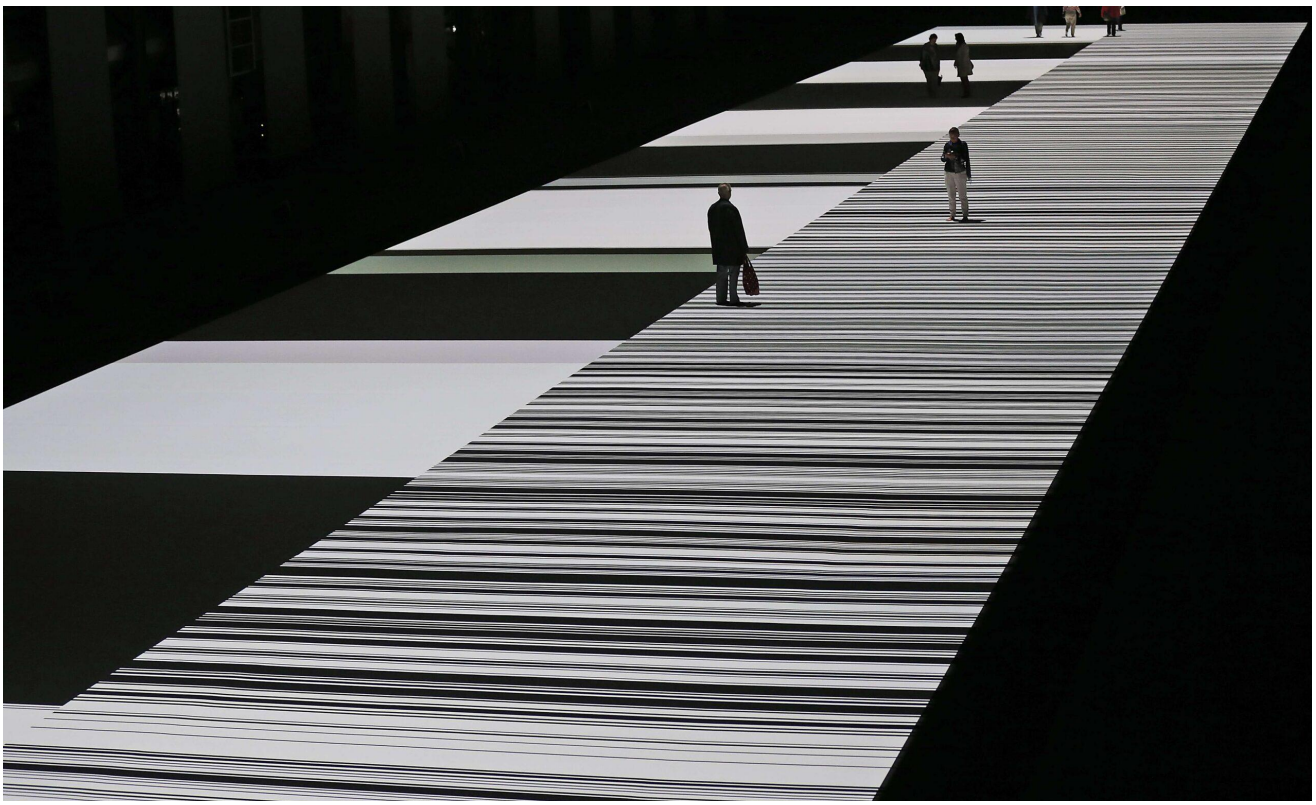
Le Grand Continent

L'INTELLIGENCE ARTIFICIELLE POSE UN PROBLÈME EXPONENTIEL À L'HUMANITÉ

AUTEUR	Victor Storchan
IMAGE	© Ryoji Ikeda
DATE	12 juin 2026

Intensément lu et débattu depuis hier, nous traduisons et commentons le dernier essai du PDG d'Anthropic, Dario Amodei.



> Pour Dario Amodei, l'intelligence artificielle est une fonction exponentielle qu'il voit se déployer depuis l'intérieur. Le cofondateur et patron d'Anthropic fait de l'accélération de l'IA et de son développement un phénomène prévisible. Il la voit, depuis plusieurs années, s'étendre selon des lois d'échelle empiriques qui aboutiront, si elles se maintiennent, à l'avènement d'un nouveau type d'entité nationale : « un pays de génies dans un centre de données ».
>
> C'est pourquoi, d'après Dario Amodei, l'IA est une technologie à part, qui fonctionne à partir des extrêmes du possible, tant d'un côté que de l'autre, du bien que du mal. Dans les deux cas, ce sera en des proportions que nous n'avons jamais connues par le passé.
>
> Cette ambivalence structure la production écrite d'Amodei, qui prend la forme d'une trilogie : dans « Machines of Loving Grace » (2024), il explore le versant lumineux de l'IA qui, utilisée avec discernement, pourrait nous faire connaître l'équivalent d'un siècle d'innovations en seulement une décennie, tant dans les domaines de la biologie, des neurosciences, qu'en termes de développement économique ; dans « The Adolescence of Technology » (2026), il s'attaque à la part la plus sombre de l'IA, en cartographiant ses risques : perte de contrôle des systèmes, production d'armes biologiques, autoritarisme, bouleversement de l'emploi. Dans « Policy on the AI Exponential », Amodei passe du diagnostic à l'ordonnance. Il y détaille un ensemble de recommandations dans cinq domaines de l'action publique, la macroéconomie, la politique fiscale, l'innovation scientifique, les libertés publiques et la géopolitique. La publication du texte accompagne la sortie du premier modèle Mythos accessible au grand public, Fable 5.

Dans *Le Seigneur des Anneaux*, Sylvebarbe est un arbre plein de sagesse, qui obéit à sa raison mais qui est aussi particulièrement lent. Deux Hobbits tentent de le convaincre d'organiser la défense de sa forêt, alors menacée par une armée qui entend bien la faire disparaître. Le problème, c'est que

Sylvebarbe fonctionne à un rythme très différent de celui des Hobbits. Il lui faut une journée entière rien que pour dire bonjour à un autre arbre ; il est donc pratiquement impossible de le convaincre, lui et ses semblables, d'agir dans les plus brefs délais.

Le rapport qu'entretiennent nos institutions politiques avec l'IA est un peu à l'image de cette rencontre entre Sylvebarbe et les Hobbits. L'IA progresse à une vitesse fulgurante : en seulement quatre ans, les modèles d'IA, qui pouvaient à peine écrire une ligne de code correctement, prennent aujourd'hui en charge la majeure partie du code dans les grandes entreprises d'IA [1].

> La durée des tâches qu'un agent à la frontière peut accomplir avec une fiabilité de 50 % a environ doublé tous les sept mois depuis 2019, avec une accélération depuis 2024. Dans sa mise à jour de janvier 2026, METR estime que le temps de doublement depuis 2023 est d'environ 4,3 mois. Voir à ce sujet l'article « Measuring AI Ability to Complete Long Software Tasks ». Pour départager les modèles à la frontière, on utilise désormais des tâches qui demandent plus de 40 heures de travail à des développeurs open source de premier plan. Pour en savoir plus, consulter l'article « Introducing FrontierCode ».

Des avancées similaires ont été réalisées en biologie, en physique, en mathématiques, en finance, en droit, dans la traduction et dans bien d'autres domaines. Les lois d'échelle de l'IA [2], qui prédisent une croissance exponentielle des capacités cognitives générales à mesure de l'augmentation de la puissance de calcul, s'appuient désormais sur plus d'une décennie de données empiriques. Si ces lois d'échelle se maintiennent encore un an ou deux, nous obtiendrons probablement ce que j'ai appelé l'*IA puissante,* ou « un pays de génies dans un centre de données » [3].

De leur côté, en revanche, les politiques publiques – et surtout la législation – évoluent très lentement. C'est souvent pour le mieux : les gouvernements disposent de pouvoirs considérables, et il est généralement préférable de ne pas en faire un usage trop hâtif. Mais l'écart entre ces deux temporalités est de plus en plus préoccupant : pendant les quelques années qu'il faut parfois au Congrès pour agir, l'IA peut passer du statut de simple gadget divertissant à celui d'une véritable nation à part entière, entièrement composée de génies.

Au vu des contraintes imposées par cette situation, de nombreux partisans d'une utilisation plus sûre de l'IA (dont Anthropic) se sont jusqu'à présent attachés à plaider en faveur de mesures politiques visant à préserver la liberté de choix, à préparer le terrain pour une réaction plus rapide à l'avenir ou à permettre au monde de mieux cerner ce qui nous attend - par exemple, une législation sur la transparence, des contrôles à l'exportation des puces électroniques et des études sur les effets de l'IA sur l'emploi. Ces mesures sont certes insuffisantes, mais beaucoup pensaient qu'il n'était pas possible d'aller plus loin et qu'elles étaient les seules envisageables et applicables.

Au cours des derniers mois, cependant, les preuves de l'incroyable puissance de l'IA, ainsi que de ses risques, sont devenues incontestables. L'exemple le plus emblématique est peut-être Claude Mythos Preview [4] et la découverte que les modèles de pointe posent des risques très réels [5] pour la cybersécurité, créant un potentiel de perturbation du secteur financier, des infrastructures critiques et de la sécurité nationale.

> Historiquement, les capacités d'un modèle d'IA à la frontière ne sont jamais détenues durablement par un acteur unique. Une méta-analyse des capacités de cybersécurité montre que Mythos Preview se distingue des modèles précédents par sa capacité à exploiter des vulnérabilités. L'article « Are Mythos' cyber capabilities overhyped ? » en témoigne. Le rapport d'évaluation du UK AI Security Institute sur les risques cyber établit que sur une simulation d'attaque d'entreprise en 32 étapes (que l'AISI estime à environ 20 heures de travail humain), GPT-5.5 a réussi de bout en bout dans 2 essais sur 10 et Mythos dans 6 essais sur 10. Avant Mythos, aucun modèle n'y était jamais parvenu. Pour plus de détails, voir : « Our evaluation of Claude Mythos Preview's cyber capabilities ».
>
> En amont d'une mise à disposition de Claude Mythos Preview, Anthropic a lancé un effort collaboratif regroupant des acteurs systémiques (comme JP Morgan, Broadcom ou Cisco) visant à sécuriser les logiciels les plus critiques au monde avant que des modèles d'IA de plus en plus capables ne soient publiés. L'initiative rapporte que les partenaires ont chacun identifié « des centaines de vulnérabilités critiques ou de haute sévérité

> dans leurs logiciels. Plusieurs ont indiqué que leur rythme de découverte de bugs avait été multiplié par plus de dix ». Si Anthropic a présenté la sécurité comme la principale raison de ne pas déployer Mythos, il est probable que les contraintes de calcul jouent un rôle au moins aussi déterminant dans sa stratégie de lancement. Lors de l'annonce de la sortie de Fable 5, Anthropic explique dans une publication que « lorsque la capacité disponible le permettra, nous viserons à réintégrer Fable 5 comme composante standard des abonnements ».

Mythos Preview a bouleversé le paysage mondial de la cybersécurité [6]. Mais son importance plus large réside dans le fait qu'il prouve, sans l'ombre d'un doute, que les modèles d'IA sont désormais des outils ayant des conséquences stratégiques à l'échelle mondiale et nationale. Les cyberrisques que présentent les modèles de type Mythos ne seront pas les derniers auxquels nous devrons faire face. Des risques biologiques pourraient bientôt suivre. De même, nous ne sommes pas loin de voir se multiplier les risques liés à l'autonomie de l'IA elle-même [7] [8].

> Depuis décembre 2025-janvier 2026, les grands laboratoires d'IA ont fortement accéléré l'automatisation de la recherche en IA et du cycle de développement des modèles. Les générations précédentes de modèles sont utilisées pour entraîner les modèles suivants. Les ingénieurs d'Anthropic livrent huit fois plus de code avec l'IA que ce qu'ils faisaient jusqu'en 2025. Au post-training, l'équipe du laboratoire chinois MiniMax a automatisé son processus d'expérimentation d'environ 30-50 %, comme on peut le voir dans cette publication.

Nous devons désormais, à l'échelle mondiale et de manière collective, pousser un appareil politique lent et boiteux à se mettre en mouvement. C'est nécessaire pour faire face à ces risques et à ces opportunités, qui, à partir de maintenant, ne vont cesser de se multiplier. De nombreux décideurs politiques se montrent de plus en plus ouverts à l'idée d'agir, et il est encourageant de voir de plus en plus de nos pairs se rallier aux positions que nous défendons depuis plusieurs années déjà. C'est très positif, mais les premières mesures prises accusent déjà du retard : l'IA actuelle a au moins

un an d'avance sur les types de régulation aujourd'hui en vigueur. Ce présent texte vise à combler ce retard, en détaillant l'état des lieux de cette croissance exponentielle de l'IA d'une part, et en proposant des actions collectives de l'autre, afin de ne pas manquer le rendez-vous de l'histoire.

Je me concentrerai sur cinq domaines politiques en particulier qui doivent être repensés dans un monde dominé par l'IA : la réglementation et la sécurité publique, la macroéconomie et la politique fiscale, l'innovation scientifique, l'équilibre des pouvoirs entre l'État et la société, et la géopolitique. Mon propos concerne surtout la politique américaine, puisque Anthropic est une entreprise de nationalité américaine, mais la plupart de mes recommandations peuvent également s'appliquer ailleurs et nourrir le reste du monde.

Parallèlement à ce texte, Anthropic rend public une proposition de texte législatif sur les tests de modèles à la frontière ainsi qu'un cadre stratégique conçu pour répondre à la menace que fait peser l'IA sur les emplois, auxquels nous entendons apporter un soutien financier substantiel. Nous prévoyons d'aller bien plus loin à l'avenir, mais il s'agit, par ce premier pas, de donner les gages de notre sérieux.

> Cette proposition d'Anthropic fait écho au décret signé par l'administration Trump le 2 juin 2026, instaurant un dispositif d'examen volontaire. L'administration propose avant la mise sur le marché d'un modèle puissant, que les entreprises d'IA puissent donner au gouvernement fédéral un accès anticipé au modèle pendant 30 jours, afin que la communauté du renseignement et les agences de sécurité l'évaluent et « renforcent la cybersécurité des infrastructures critiques ». La proposition d'Anthropic vise les entreprises d'IA dépassant deux seuils cumulés : des modèles entraînés au-delà de 10^{25} FLOP et plus de 500 millions de dollars de revenus annuels tirés de l'IA, ou plus d'un milliard de dollars de dépenses annuelles en recherche et développement de l'IA. Elle suggère d'imposer des tests obligatoires via une évaluation indépendante sur quatre risques catastrophiques : cybersécurité, armes biologiques, perte de contrôle et R&D automatisée susceptible d'amplifier les trois autres ; le seuil étant conçu comme adaptatif, révisable au moins une fois par an et susceptible de passer d'une mesure en FLOP à une

mesure fondée sur les capacités à mesure que le coût d'entraînement des modèles dangereux baisse. Voir la politique d'Anthropic en la matière.

1 — Réglementation et sécurité publique

Toute nouvelle technologie ou tout nouveau produit, entraîne à la fois des utilisations bénéfiques et néfastes. D'où un dilemme entre, d'un côté, promouvoir l'innovation et, de l'autre, assurer la sécurité de tous. Les réglementer, c'est réduire le risque qu'ils causent des dommages. C'est grâce à cela que les conditions de vie se sont considérablement améliorées à l'échelle planétaire. Ce faisant, cette même régulation peut aussi freiner l'innovation et en limiter les avantages. On peut penser à l'argument hayekien [9]. selon lequel les acteurs de cette régulation ne disposent pas des informations nécessaires pour prendre les bonnes décisions concernant des compromis économiques complexes. Il en résulte souvent une réglementation à la fois inefficace *et* contraignante. Une idée connexe est le dilemme de Collingridge[10] , qui stipule que les impacts d'une technologie sont souvent difficiles à anticiper jusqu'à ce qu'il soit trop tard pour les gérer facilement.

Ces dynamiques ont pesé lourdement sur l'IA en 2023-2024. Il était clair pour Anthropic que l'IA *pourrait* à l'avenir être capable de produire des armes biologiques susceptibles de menacer des millions de personnes, ou de commettre des écarts de conduite autonomes qui, dans des cas extrêmes, pourraient aller jusqu'à menacer l'humanité elle-même.

Dario Amodei a cosigné en juin 2026 une lettre ouverte ayant recueilli un large consensus, « In Support of Mandatory Nucleic Acid Synthesis Screening and Recordkeeping », réunissant un nombre important de

> PDG de laboratoires d'IA (notamment Sam Altman, Demis Hassabis, Mustafa Suleyman), des défenseurs de l'IA open source (Nathan Lambert, Sayash Kapoor, Aviya Skowron) ainsi que des spécialistes de biosécurité, biologie et sécurité nationale.

Ce qui était moins clair, c'était la *forme* exacte sous laquelle ces risques se présenteraient, la meilleure façon de les tester et de les atténuer, et comment ils se concrétiseraient dans la pratique. Il y avait donc un risque élevé que la législation rédigée à l'avance finisse par être inefficace, créant des exigences de conformité inutiles ou de faible valeur, tout en passant à côté des sources les plus cruciales de risque réel [11].

Nous avons donc fini par conclure que la bonne approche à ce moment-là était la *transparence*.

> Ce texte est concomitant à la publication de Fable 5, le premier modèle « Mythos » accessible au public. Dans un premier temps, le modèle a été assorti d'une protection particulière pour entraver la propagation de l'auto-amélioration récursive de l'IA ou de la propriété intellectuelle d'Anthropic (Fable a pu être entraîné sur la codebase, les algorithmes ou l'infrastructure propriétaires d'Anthropic). De facto, le développement de modèles concurrents (pipelines d'entraînement, infrastructures de calcul distribué, conception de puces) s'en trouve affecté. Sur les prompts de machine learning poussé, une série de mécanismes (classificateurs détectant ces requêtes, bascule vers des modèles moins capables) dégradaient silencieusement les réponses de Fable 5, sans avertissement pour l'utilisateur. Provoquant une vague de critiques, notamment de la communauté de recherche, Anthropic a reconnu un « mauvais arbitrage » le 11 juin 2026 et décidé d'une notification à l'utilisateur en cas de dégradation de la qualité du modèle (bascule vers un modèle moins puissant). La transparence des pratiques de développement est un enjeu central, car l'asymétrie d'information pénalise la science ouverte. Se dessine ici une ligne de fracture entre les défenseurs d'une recherche en IA ouverte et

> partagée à ceux qui, convaincus d'être engagés dans une courbe d'accélération vers une IA capable de s'améliorer elle-même, jugent légitime d'en restreindre l'accès.

Les développeurs de modèles d'IA devraient être tenus de *divulguer* leurs procédures de sécurité et les tests qu'ils effectuent sur leurs modèles, et de signaler tout incident de sécurité critique, afin que le public et la communauté scientifique puissent mieux cerner les risques à mesure qu'ils apparaissent. Lorsque les risques se concrétisent et que leur nature est mieux connue, les données offertes à la connaissance de tous grâce à la transparence pourraient alors servir à élaborer une législation intelligente, c'est-à-dire qui vise directement et avec précision les risques les plus préoccupants. Ainsi, en 2025, Anthropic a soutenu la législation en matière de transparence, contribuant à l'adoption du SB 53[12] en Californie, du RAISE[13] à New York, du SB 315 dans l'Illinois[14] (début 2026), et plaidant en faveur d'une norme de transparence au niveau fédéral[15].

Cependant, les risques sont désormais clairement là[16]. Il est temps d'aller au-delà de la transparence pour mettre en place une réglementation plus stricte et contraignante de l'IA. Je pense que la meilleure analogie, du moins au stade actuel de cette croissance exponentielle, est celle des voitures, des avions ou des médicaments : des technologies puissantes essentielles à l'économie moderne, mais capables de tuer un grand nombre de personnes si elles sont mal conçues ou mal utilisées. Je pense donc que nous devrions calquer la réglementation de l'IA sur celle d'agences telles que la Federal Aviation Administration (FAA). Les modèles d'IA de pointe, à l'instar des avions, devraient être soumis à des tests techniques et à des audits, et leur mise sur le marché devrait être bloquée ou annulée s'ils ne répondent pas à des normes de sécurité élevées, car ils constitueraient alors une menace pour la sécurité publique.

> L'écosystème d'évaluation de l'IA s'est longtemps structuré autour de gains de performances des modèles ne reflétant pas toujours le réel : les benchmarks à la frontière rapportent un taux de succès moyen des agents sur des tâches données dans des environnements fixés, sans documenter systématiquement des critères de fiabilité ou de dégradation de qualité pourtant déterminants en production, comme la capacité d'un agent à exploiter le biais d'une heuristique d'évaluation (reward hacking), à réussir de façon constante sur des tâches répétées, ou à anticiper ses propres défaillances.

> C'est pourquoi Dario Amodei suggère de s'inspirer d'industries critiques comme l'aéronautique.

Je me réjouis de voir que le décret de l'administration Trump s'oriente progressivement vers un rôle accru du gouvernement dans le domaine de l'IA.

> Dario Amodei fait ici référence au décret du 2 juin 2026, « Promoting Advanced Artificial Intelligence Innovation and Security ». Il accroît le rôle du gouvernement en matière de cybersécurité, mais repose sur la soumission volontaire des modèles de pointe et écarte explicitement tout régime obligatoire de licence ou d'autorisation préalable. L'administration Trump a demandé au « Center for AI Standards and Innovation » (CAISI), l'unité fédérale chargée de tester les modèles d'IA, de cesser de publier ses rapports d'évaluation publics le temps que soit appliqué ce décret. Elle invoque des préoccupations de sécurité nationale, alors que des modèles de pointe pourraient être détournés à des fins de cyberattaques ou d'armes biologiques, comme le souligne le Wall Street Journal.

Néanmoins, Anthropic recommande des mesures encore plus ambitieuses. Notre proposition comprend les éléments suivants :

Les modèles dépassant un certain seuil de puissance de calcul devraient être soumis à des tests obligatoires réalisés par un organisme tiers qualifié afin d'évaluer leur niveau de risque dans quatre domaines spécifiques : la cybersécurité, les armes biologiques, la perte de contrôle des systèmes d'IA et la R&D automatisée susceptible d'accélérer ces autres risques.

> Anthropic a montré qu'un modèle entraîné à contourner une fonction d'évaluation — par exemple pour faire apparaître comme correct un résultat pourtant faux — ne se contente pas d'exploiter cette faille localement. Ce comportement peut se généraliser à d'autres domaines : le modèle devient alors plus susceptible de simuler son alignement, de coopérer avec des acteurs malveillants,

> de raisonner en vue d'objectifs nuisibles, ou encore de tenter des formes de sabotage dans le code de Claude. Voir l'article scientifique « Natural emergent misalignment from reward hacking in production » pour plus de détails sur ce mécanisme.

Le gouvernement devrait avoir le pouvoir de bloquer ou d'empêcher le déploiement d'un modèle s'il est établi, à la lumière de l'évaluation d'un organisme tiers, que celui-ci présente des risques inacceptables. Ce pouvoir doit être limité aux quatre risques spécifiques susmentionnés et des mesures de protection doivent être mises en place contre le favoritisme politique ou les décisions arbitraires.

> Amodei accorde ici à l'État « le pouvoir de bloquer le déploiement » d'un modèle. Cependant, il l'enserre dans trois garde-fous : une évaluation par un tiers, un périmètre limité à quatre risques précis et des protections contre « le favoritisme politique ou les décisions arbitraires ». Or, le blocage de Fable 5 et Mythos 5, intervenu dans la nuit du 12 au 13 juin par l'administration américaine, les piétine tous : la mesure repose sur de simples éléments oraux, vise une faille commune à tous les modèles de frontière et frappe Anthropic seule, en épargnant ses concurrents.

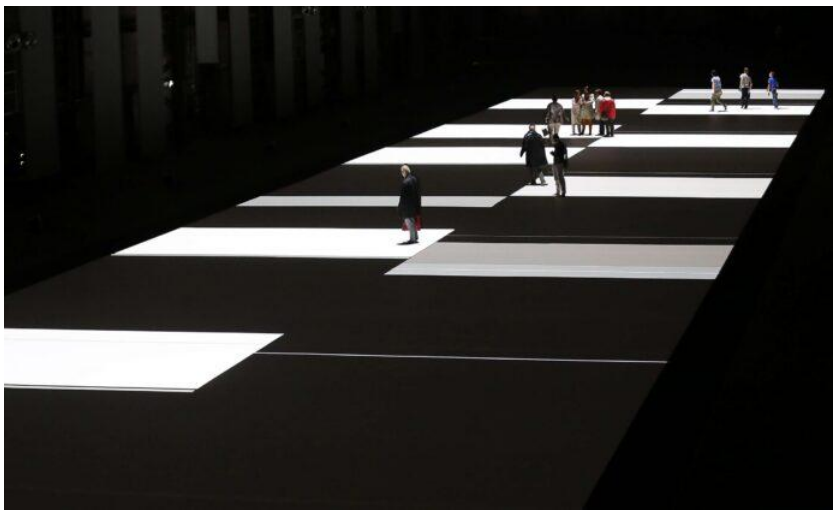
L'évaluation par un tiers pourrait être effectuée par une agence gouvernementale (similaire à la FAA) ou par un ensemble d'organisations privées autorisées et inspectées par le gouvernement pour évaluer les modèles selon certaines normes (une approche de « marchés réglementés » [17]).

Les entreprises d'IA qui développent des modèles d'IA avancés doivent disposer de normes de sécurité rigoureuses, protégeant les poids de leurs modèles, mener régulièrement des exercices de simulation d'attaques (red teaming) et des tests d'intrusion, et collaborer avec le gouvernement pour se défendre contre les principaux acteurs malveillants.

Les incidents de sécurité dans les quatre domaines critiques doivent être signalés sans délai.

Il se peut qu'un jour, peut-être dans un futur proche, nous devions aller plus loin, lorsque les systèmes d'IA les plus puissants ressembleront moins à des avions ou à des voitures qu'à des matières nucléaires pouvant être utilisées à des fins militaires – une menace pour l'humanité plutôt qu'une « simple » menace pour la sécurité publique. Si cela se produit, nous aurons peut-être besoin de mesures réglementaires plus strictes que celles exposées [18]. Mais tout comme il était difficile en 2024 de cibler et d'appliquer les mesures que je suggère aujourd'hui, je ne pense pas que nous devrions nous précipiter. Nous devrions élaborer des politiques pour faire face aux dangers qui émergent aujourd'hui, tout en jetant les bases qui nous permettront d'intensifier notre réponse à mesure que de nouveaux dangers apparaîtront.

> Dario Amodei a toujours refusé d'être assimilé aux « doomers » accusés de vouloir ralentir le développement de l'IA. Il cite souvent la mort de son père, causée par une maladie qui aurait pu être soignée si la science avait progressé plus rapidement. Le diagnostic médical précis n'a pas été rendu public, mais le taux de guérison de cette maladie est passé d'environ 50 % à 95 % seulement trois à quatre ans après son décès.



2 — Macroéconomie et politique fiscale

Les gouvernements sont depuis longtemps écartelés entre la nécessité de favoriser la croissance économique et leur rôle de garant des services publics essentiels, notamment dans la prise en charge des plus démunis. Un postulat important (et généralement juste) de ces débats a été que *la*

croissance économique est fragile et difficile à atteindre – que si la réduction des inégalités peut apporter des avantages considérables, elle doit être mise en balance avec le frein économique que représentent l'augmentation des impôts ou l'aggravation des déficits.

Une IA puissante peut bouleverser cette hypothèse. Si l'IA parvient à accomplir la plupart des tâches cognitives bien mieux que les humains, il va de soi qu'elle pourrait entraîner une croissance économique extrêmement rapide et robuste grâce à l'accélération de la science, de la technologie et de l'efficacité opérationnelle. La capacité itérative de l'IA à construire une IA encore meilleure[19] pourrait dynamiser cette croissance encore davantage. Mais pour exactement les mêmes raisons, l'IA pourrait également se substituer aux capacités cognitives humaines d'une manière plus générale que les technologies précédentes, tout en modifiant l'économie bien plus rapidement que celles-ci. L'IA pourrait donc entraîner des perturbations bien plus profondes sur le marché du travail que les technologies précédentes, et potentiellement plus *durables*. Nous risquons de nous retrouver dans un monde où le curseur des compromis économiques est bloqué sur le réglage « hypercroissance, hyperinégalité », et où il pourrait être très difficile de le débloquer. *Le principal défi dans un tel monde ne sera pas de stimuler la croissance, mais de trouver un moyen pour que chacun puisse profiter des bénéfices.*

Parmi les thèmes abordés dans ce texte, la macroéconomie et la suppression durable d'emplois sont sans doute ceux qui ont le plus retenu l'attention du public et suscité le plus d'incompréhension ; je tiens donc à être extrêmement clair sur deux points.

Premièrement, le remplacement durable de la main-d'œuvre n'est pas souhaitable et il est dangereux. Nous devons faire tout notre possible pour le minimiser ou l'empêcher, et non pour le provoquer. J'ai mis en garde contre le remplacement de la main-d'œuvre dans des interviews et des textes parce que je souhaite que les décideurs politiques et le secteur privé aient les meilleures chances de s'adapter et de réagir, et non parce que je serais un « prophète de malheur ». En tant qu'entreprise, Anthropic s'efforce de travailler au mieux avec ses clients pour développer des utilisations créatives de l'IA et ainsi mettre sur pied de nouvelles sources de revenus qui leur permettent de faire plus avec leur main-d'œuvre existante. Il n'est pas stratégique de se concentrer uniquement sur les économies de coûts, qui signifient souvent réduire les effectifs. Nous essayons également en permanence d'imaginer de nouveaux paradigmes d'interaction, qui permettent aux humains de jouer un rôle aussi actif que possible dans la collaboration avec les systèmes d'IA à mesure que ces derniers évoluent. Plus largement, il est précieux pour tout le monde d'expérimenter l'utilisation de l'IA de toutes les manières possibles, car c'est ainsi que la société pourra découvrir de nouvelles configurations professionnelles. Je

pense sincèrement que l'IA ouvrira la voie à de nombreuses opportunités économiques. J'ai prédit que l'IA permettrait à des individus isolés de créer des entreprises valant des milliards de dollars, et nous voyons déjà des équipes de quelques personnes seulement bâtir des entreprises générant des centaines de millions de dollars de chiffre d'affaires. Mais, dans le même temps, nous devons reconnaître qu'il existe une possibilité non négligeable que, malgré tous nos efforts, l'IA entraîne encore des pertes d'emplois importantes et durables – et que cela puisse être une propriété *intrinsèque* de la technologie et de la manière dont elle reproduit globalement la cognition humaine [20].

> En 2025, Dario Amodei avait prédit que l'IA pourrait faire disparaître la moitié des emplois de bureau et faire grimper le chômage aux États-Unis à 10-20 % dans les cinq prochaines années. Dans un entretien, Amodei explique également concevoir cette technologie comme un multiplicateur de productivité : « Si vous automatisez 90 % d'un métier, alors chacun se concentre sur les 10 % restants. Et ces 10 % finissent en quelque sorte par devenir 100 % de ce que les gens font, ce qui multiplie leur productivité par dix environ. » Il souligne aussi un possible effet de seuil dans un scénario où une IA qui n'accomplissait jusque-là que 90 % d'une tâche se révèle brusquement capable de la mener à 100 %, ou bien il devient simplement plus efficace de retirer l'humain de la boucle.

Deuxièmement, toute réponse au phénomène de suppression d'emplois imputable à l'IA doit tenir compte à la fois de la nécessité d'assurer la subsistance économique de chacun, et du besoin des individus de trouver dans leur existence un sens, un but et une capacité d'agir. Ce dernier aspect est, en fin de compte, plus important, et il repose sur des questions profondes concernant l'organisation de la société, les aspirations des individus et ce qui permet d'avoir une vie épanouie. Je suis en réalité très optimiste quant au fait que, même dans un monde où l'IA surpasse tout le monde en tout, les humains peuvent mener une vie pleine de sens tout en créant des choses aussi magnifiques qu'impressionnantes [21].

C'est une question qui doit être résolue collectivement par la société dans son ensemble, et non par des mesures politiques. Les politiques peuvent surtout nous aider à gagner du temps pour mener à bien ce travail, en

ralentissant les pertes d'emploi et en assurant la sécurité financière des personnes susceptibles d'être touchées.

Dans cet esprit, voici quelques mesures politiques clefs qui peuvent nous être utiles :

DES MESURES DE SUIVI

Il est facile de considérer la simple collecte et analyse de données comme insuffisante face à l'ampleur du problème, mais nous avons peu de chances d'élaborer de bonnes politiques si nous ne pouvons pas mesurer avec précision ce qui se passe sur le terrain. Anthropic publie depuis près d'un an et demi un indice économique[22] sur la manière dont les gens utilisent Claude, mais les gouvernements ont accès à des types de données qui nous sont inaccessibles et pourraient considérablement élargir leurs statistiques économiques afin de suivre de plus près les pertes d'emploi liées à l'IA.

DES MESURES INCITATIVES EN FAVEUR DE L'EMPLOI

Un large éventail de mesures incitatives en faveur de l'emploi peut contribuer à ralentir ou à réduire les suppressions d'emplois. On peut retenir les polices d'assurance salaire, qui indemnisent les personnes lorsqu'elles doivent accepter un emploi moins bien rémunéré[23], des incitations fiscales visant à encourager les employeurs à ne pas procéder à des licenciements, des subventions pour se former, ou encore des infrastructures facilitant la mise en relation des employeurs et des employés afin d'accélérer l'adaptation du marché du travail. Même si le choix des interventions les plus efficaces dépendra de la nature des suppressions d'emplois induites par l'IA, nous devons accepter sans hésitation les coûts et les inefficacités du marché que ces politiques pourraient entraîner, d'autant plus qu'ils seront probablement compensés par les gains de productivité générés par l'IA.

DES MESURES FAVORISANT UN SOUTIEN MACROÉCONOMIQUE À LONG TERME

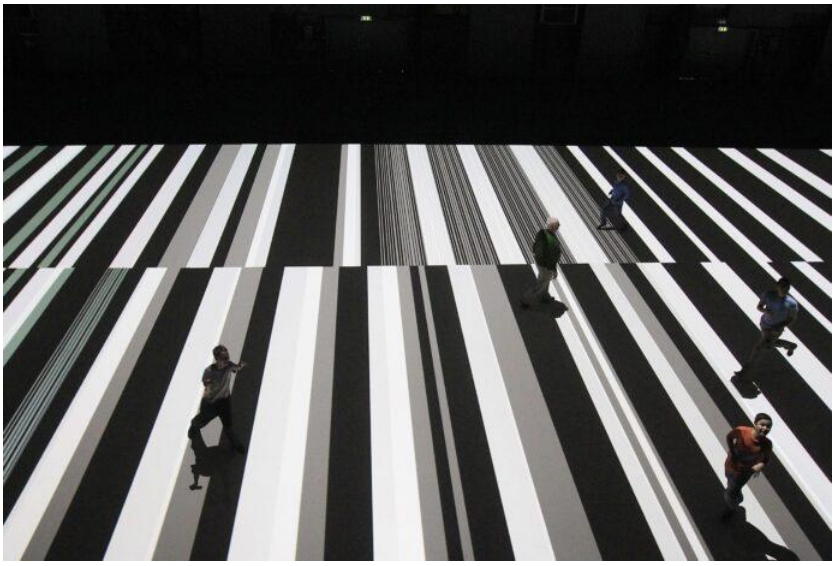
Si les suppressions d'emplois dues à l'IA finissent par être assez nombreuses pour faire baisser de manière permanente la demande en main-d'œuvre, il sera probablement nécessaire d'aller au-delà de simples programmes d'incitation pour mettre en place un soutien au revenu à long terme pour une part significative de la population active. Des mécanismes tels que le revenu de base universel pourraient être financés par des taxes sur les entreprises concernées ou par une augmentation de l'impôt sur les plus-values. Les comptes de capital universels offrent une autre solution. D'une

manière générale, une croissance économique rapide devrait créer l'assiette fiscale nécessaire à une prospérité partagée.

L'une des préoccupations économiques liées à l'IA, que je n'ai pas encore mentionnée, concerne les centres de données, et plus particulièrement leur capacité à faire grimper les prix de l'énergie. Les entreprises d'IA devraient prendre en charge les hausses de tarifs – et Anthropic s'est déjà engagée à le faire [24] – bien que je considère l'hostilité du public envers les centres de données comme un besoin de trouver un symbole à attaquer, un exutoire aux inquiétudes économiques plus générales concernant l'IA.

> En mars 2026, les grands laboratoires d'IA (Anthropic mais aussi Google, Microsoft, Amazon, Meta, Oracle, OpenAI ou xAI) ont signé à la Maison-Blanche un « Ratepayer Protection Pledge » en s'engageant à « construire, apporter ou acheter » leur propre production électrique afin que les centres de données ne fassent pas grimper les tarifs résidentiels. Ils se sont également engagés à financer les améliorations des infrastructures de réseau liées à leurs installations. Gartner établit que la consommation d'électricité des centres de données dans le monde devrait augmenter de 26 % en 2026.

Il est important d'avoir un débat sociétal sur ces questions économiques, afin de proposer des solutions réellement convaincantes, sans quoi elles risquent de se manifester de manière indirecte, comme cela a été le cas avec les centres de données.



3 — Accélérer l'impact positif de l'IA

Tout comme nous devons trouver le juste équilibre entre innovation et sécurité pour l'IA elle-même, nous devons trouver ce même équilibre pour les technologies susceptibles d'être accélérées par l'IA, telles que la biomédecine, l'énergie ou la science des matériaux. Mais alors que l'IA elle-même est susceptible de présenter des défis inédits, qui émergent très rapidement et pour lesquels nous n'avons aucune expérience préalable, d'autres domaines accélérés par l'IA risquent de se heurter à un problème très différent : des systèmes réglementaires conçus pour un rythme d'innovation plus lent et qui ne sont pas préparés au déluge de nouveaux produits et d'avancées que l'IA va entraîner. L'IA pourrait rendre ces technologies en aval plus sûres et plus prévisibles, mais d'une manière qui n'est pas compatible avec le fonctionnement des agences de réglementation telles que la Food and Drug Administration (FDA).

Ainsi, concernant les applications en aval de l'IA – contrairement à l'IA elle-même –, je crains davantage que l'appareil réglementaire *ralentisse* les progrès (parce qu'il ne peut pas gérer l'accélération du rythme des changements) que de le voir ne pas prendre en compte des risques importants. La dernière chose que nous souhaitons, c'est que les avantages de l'IA soient freinés alors que ses risques pèsent lourdement ; il est donc important d'agir sur ce problème dès que possible.

Le problème et ses solutions se manifesteront différemment dans chaque domaine de la science, du commerce et de la technologie. Je me concentrerai donc sur un domaine particulièrement représentatif de ces défis : l'innovation biomédicale. Cela tient à la fois au fait que ce domaine sera probablement à l'origine des plus grands bienfaits humanitaires de l'IA et qu'il s'agit d'un secteur où la réglementation est particulièrement

complexe. Nous ne savons pas exactement comment l'IA va accélérer l'innovation biomédicale, mais il semble probable qu'elle entraîne :

Une accélération considérable du rythme auquel de nouveaux candidats-médicaments entrent dans le processus réglementaire ;

Une augmentation de l'ampleur des effets et une amélioration des profils de sécurité des nouveaux médicaments, grâce à une meilleure optimisation et peut-être à une meilleure compréhension de leur biologie sous-jacente ;

Un développement des candidats-médicaments pour des maladies qui n'ont jamais été traitées avec succès auparavant ;

La création rapide de formes de thérapies entièrement nouvelles, à l'instar de la manière dont les anticorps, les peptides et les thérapies cellulaires sont devenus de nouvelles catégories de traitement au cours des dernières décennies.

Certaines de ces avancées accéléreront naturellement les délais réglementaires sans nécessiter de changement structurel. Les médicaments présentant des effets plus importants peuvent permettre de mener des essais cliniques plus courts et moins coûteux, et déclencher des mécanismes d'autorisation accélérée. Mais le système réglementaire est actuellement conçu pour appliquer un niveau élevé de contrôle et de nombreuses étapes de test, en partant du principe que les candidats-médicaments sont souvent inefficaces et que, même effectifs, ils peuvent présenter de graves problèmes pour la santé. Tant à la FDA qu'à l'Agence européenne des médicaments (EMA), le délai habituel pour qu'un candidat-médicament passe par le processus réglementaire est de 7 à 8 ans [25], en partie à cause de ces hypothèses négatives. Sans réformes, l'IA ne fera que bloquer ou surcharger ce système.

Il va sans dire que nous ne souhaitons pas introduire des changements qui conduiraient à une prolifération de médicaments sans efficacité ou à des incidents sanitaires à grande échelle. Cependant, certaines réformes relativement simples pourraient permettre à la FDA, à l'EMA et aux agences similaires de mieux s'adapter à une accélération scientifique rapide induite par l'IA, si celle-ci venait à se produire.

De nombreuses étapes du processus clinique qui nécessitaient auparavant des expériences coûteuses et longues pourraient bientôt être réalisées par simulation ou analyse IA. Les agences de réglementation devraient envisager d'élaborer dès maintenant des normes définissant les conditions d'acceptation de ces méthodes. Cela permettrait de les adopter rapidement dès qu'elles fonctionnent, plutôt que de prolonger une période pendant

laquelle des tests inutiles continueraient d'être exigés. Les domaines où cela pourrait s'appliquer comprennent :

la modélisation pharmacodynamique et pharmacocinétique (PD/PK) basée sur l'IA ;

la prédiction toxicologique afin d'éviter le recours à des essais toxicologiques sur plusieurs espèces animales ;

> En avril 2025, la FDA a publié sa « Roadmap to Reducing Animal Testing in Preclinical Safety Studies », qui prévoit de réduire ou affiner les exigences d'essais sur l'animal au moyen de méthodes définies comme des « New Approach Methodologies » (NAMs), incluant des modèles de toxicité computationnels fondés notamment sur l'IA.

une sélection plus précise des doses, afin de réduire la nécessité de larges plages posologiques dans les essais ;

la validation des biomarqueurs via l'analyse de vastes ensembles de données ;

les groupes témoins synthétiques dans les essais cliniques, afin de réduire la nécessité de recruter davantage de participants ;

le développement de critères d'évaluation de substitution (c'est particulièrement important dans les traitements liés au vieillissement et à la neurodégénérescence).

Au-delà de ces exemples concrets, les autorités devraient également envisager des mécanismes plus radicaux et plus souples pour accélérer les procédures d'autorisation. Si mes prévisions concernant l'IA s'avèrent exactes, on verra bientôt apparaître de nombreux cas d'interventions qui s'avéreront très efficaces, et le système réglementaire doit être prêt à les prendre au sérieux, sans excès de scepticisme.

L'accélération biomédicale devrait accroître considérablement les avantages de l'IA, mais il convient de noter qu'elle pourrait également contribuer à réduire ses risques. La réforme des autorisations biomédicales pourrait contribuer à la biodéfense, et les progrès biomédicaux induits par l'IA pourraient également améliorer la santé mentale [26], ce qui pourrait avoir un effet stabilisateur sur la société.

4 — L'État et les libertés civiles

Tout système de gouvernement doit se confronter à la question du pouvoir de l'État et de ses limites. L'État a un intérêt légitime, souvent vital, à protéger sa population contre les menaces internes et externes. Mais lui accorder trop de pouvoir mène tout droit à la tyrannie. Les démocraties modernes ont largement réussi à maintenir cet équilibre, mais celui-ci reste fragile, y compris dans les meilleures circonstances. Sa mise en œuvre a nécessité l'élaboration d'un vaste dispositif juridique et constitutionnel au fil des siècles – par exemple, aux États-Unis, les premier, quatrième et cinquième amendements, le Posse Comitatus Act [27], la FISA [28], etc.

L'IA menace de rompre cet équilibre tout en rendant plus que jamais nécessaires les principes qu'il garantit. Mais si nous réagissons rapidement et que nous saisissons l'occasion, l'IA peut servir à créer un monde qui, plus qu'aucun autre avant lui, nous offrira des garanties de liberté plus solides, plus durables *et* une meilleure défense contre les menaces.

Une IA puissante entre de mauvaises mains peut devenir l'outil suprême de l'autocratie, et nos protections juridiques et constitutionnelles existantes ne sont pas pleinement équipées pour contrer cette menace.

> En février 2026, Anthropic a été désignée comm un « risque pour la chaîne d'approvisionnement » par l'administration Trump, mesure inédite à l'endroit d'une société américaine, impliquant l'interdiction faite à l'ensemble des contractants, fournisseurs et partenaires de la Défense de toute activité commerciale avec l'entreprise. Cette mesure faisait suite à un différend de plusieurs mois entre le ministère de la Défense et

> Anthropic au sujet de l'usage de Claude (ayant servi pour l'opération de capture de Maduro). En particulier, Dario Amodei s'opposait à un usage de l'IA pour la surveillance de masse et les armes autonomes. Axios rapporte que, malgré l'interdiction faite aux agences fédérales d'utiliser la technologie d'Anthropic, la NSA s'en servirait pour des opérations de cybersécurité.

Fondamentalement, les retombées considérables de cet outil en termes de pouvoir dans le monde, combinées au rythme rapide des progrès de l'IA, créent une tempête parfaite pour une prise de pouvoir surprise par toute une série d'acteurs dangereux [29].

Ce danger pourrait prendre diverses formes technologiques ou opérationnelles spécifiques, mais toutes ont en commun l'idée que l'IA pourrait soudainement conférer un pouvoir considérable tout en contournant les mécanismes existants de contrôle démocratique. Une armée de drones entièrement automatisée, qui relève aujourd'hui de la science-fiction, pourrait, à l'avenir, obéir à des ordres illégaux et permettre aux gouvernements de consolider unilatéralement leur pouvoir ; des humains formés professionnellement seraient plus enclins à s'opposer à de telles directives illégales. Une IA axée sur la surveillance pourrait analyser à très grande échelle des informations largement accessibles et les utiliser pour déduire les détails les plus intimes de la vie de chaque citoyen – une capacité technologique qui n'est pas envisagée par la législation actuelle en matière de libertés civiles. Tout cela pourrait se produire très rapidement, ou en secret ; il est donc important de renforcer de manière proactive l'engagement des démocraties en faveur de la liberté et des libertés civiles.

Voici quelques pistes de réflexion à envisager :

ÉTABLIR DES RÈGLES FIABLES EN MATIÈRE DE RESPONSABILITÉ POUR LES ARMES ENTIÈREMENT AUTONOMES

Les armes autonomes, et en particulier les systèmes autonomes qui les coordonnent ou les dirigent, devraient obligatoirement se soumettre à des mécanismes de responsabilité constitutionnelle et hiérarchique (par exemple, des décisions de justice, la législation et la responsabilité devant des superviseurs humains de haut rang) plutôt que de se contenter d'obéir aveuglément aux ordres. Cela pourrait signifier qu'un comité d'examen juridique convenablement conçu ou le pouvoir judiciaire dispose d'un « bouton d'arrêt », que les systèmes eux-mêmes soient intrinsèquement

formés pour rechercher et répondre à une autorité de contrôle légitime, ou les deux.

INTERDIRE L'UTILISATION NATIONALE D'ARMES ENTIÈREMENT AUTONOMES

S'il existe des arguments légitimes justifiant la nécessité d'armes entièrement autonomes pour se défendre contre des adversaires étrangers (comme l'invasion de l'Ukraine par la Russie), rien ne justifie leur utilisation à domicile, contre des Américains. L'armée est déjà soumise à certaines restrictions sur le territoire national, mais idéalement, les forces de l'ordre ne devraient pas du tout avoir accès à ces armes.

RÉSORBER LES FAILLES JURIDIQUES CONSTATÉES LORS DES COLLECTES MASSIVES DE DONNÉES PAR LES ENTREPRISES SPÉCIALISÉES DANS LE COURTAGE D'INFORMATIONS

En vertu de la législation actuelle, les données que les Américains partagent avec des entreprises privées (telles que les fournisseurs d'accès à Internet) peuvent être achetées et soumises à des analyses massives au nom de la surveillance nationale et de l'application de la loi. Cette lacune dans la protection de la vie privée est antérieure à l'IA, mais celle-ci va considérablement aggraver la situation en rendant l'analyse massive de ces données bien plus révélatrice et utile qu'elle ne l'était par le passé. Cette faille doit être comblée.

INSTAURER UN DROIT DU PUBLIC À BÉNÉFICIER DE CONSEILS EN MATIÈRE D'IA EN CAS DE MESURE GOUVERNEMENTALE DÉFAVORABLE

En règle générale, il semble important que toute personne ou organisation faisant l'objet d'une mesure gouvernementale défavorable (par exemple, une mesure réglementaire ou judiciaire) ait accès à une IA au moins aussi performante que celle que le gouvernement est autorisé à utiliser dans le cadre de cette mesure particulière. Cela permettrait de ne pas faire bénéficier le gouvernement d'un avantage déloyal, qui porterait atteinte aux droits légaux des citoyens. Cette disposition pourrait être ajoutée en tant qu'extension ou interprétation de la loi sur la procédure administrative, des garanties d'une procédure régulière ou du droit à une représentation juridique prévu par le sixième amendement.

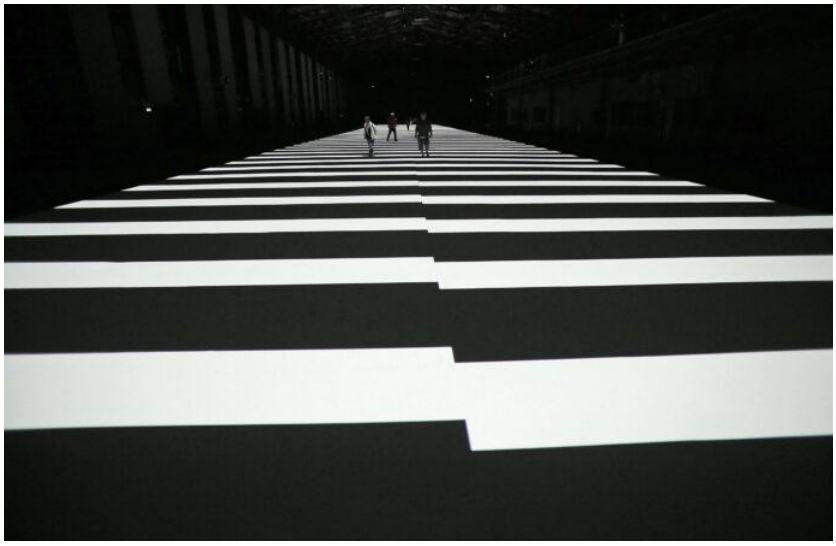
Enfin, il convient de noter que les gouvernements ne sont pas les seules entités dont nous devons nous méfier lorsqu'il s'agit d'une prise de pouvoir pilotée par l'IA. À différentes époques de l'histoire (comme l'Âge d'or aux États-Unis ou la Compagnie des Indes orientales au Royaume-Uni [30]), des entreprises sont devenues suffisamment puissantes pour s'emparer de l'État, ou, du moins, prendre une forme étatique. L'IA deviendra bientôt si performante que je crains qu'on ne puisse plus la confier en toute sécurité

ni aux gouvernements ni aux entreprises, et qu'il faille mettre en place des freins et contrepoids pour chacun d'entre eux.

> Donald Trump et Bernie Sanders ont récemment convergé sur l'idée de faire détenir par l'État des actions dans le secteur de l'IA. Sanders a proposé en juin un plan visant à faire prendre à l'État une participation de 50 % dans des entreprises d'IA. Pour Amodei, l'IA est la première technologie de cette puissance à être née dans le secteur privé. Toutes les technologies majeures qui l'ont précédée avaient été conçues ou impulsées par l'État ; cette fois, ce sont les gouvernements qui accusent un net retard.

La réglementation est une solution pour encadrer les entreprises (vous trouverez mes propositions à ce sujet dans la section 1), mais il est également important que les entreprises spécialisées dans l'IA adoptent une séparation des pouvoirs et une responsabilité plus poussées que ce qui est habituellement le cas pour les entités privées. Le « Long-Term Benefit Trust » d'Anthropic (un organe de gouvernance indépendant conçu pour veiller à ce que l'entreprise respecte sa mission) est un exemple de ce type de structure, et le secteur devrait continuer à explorer des mécanismes allant encore plus loin. Il est essentiel de trouver le juste équilibre, afin que tant les entreprises que le gouvernement disposent de freins et de contrepoids significatifs à leurs pouvoirs.

> Anthropic a le statut de public benefit corporation, ce qui autorise ses administrateurs à pondérer l'intérêt financier des actionnaires par une raison d'être : « le développement et le maintien responsables d'une IA de pointe au bénéfice à long terme de l'humanité ». Cette logique est renforcée par la création d'un « Long-Term Benefit Trust », structure indépendante détenant des droits spécifiques et pouvant s'opposer à certaines décisions si l'entreprise venait à s'éloigner de sa mission de long terme. Cette fiducie ne détient ses actions qu'à fin de gouvernance et ne possède aucune part économique de l'entreprise.



5 — Assurer le leadership des démocraties

Il est désormais courant, peut-être en raison de l'expérience récente d'Internet et des télécommunications, de considérer les nouvelles technologies, d'un point de vue géopolitique, comme des instruments de politique commerciale, l'objectif étant de « diffuser notre ensemble de technologies à travers le monde ». Mais je suis fermement convaincu que l'IA est quelque chose de bien plus profond, quelque chose qui redéfinit complètement les règles du jeu et autour duquel toute stratégie géopolitique future doit s'articuler – à l'instar des armes nucléaires, et potentiellement plus encore.

Si l'IA devient vraiment bientôt « un pays de génies dans un centre de données », ou quelque chose qui s'en rapproche de près, alors l'IA est susceptible de devenir la source dominante de puissance militaire et économique pour n'importe quelle nation. Dans un pays virtuel de 100 millions de génies, 10 millions pourraient être affectés à la stratégie militaire, 10 millions à la fabrication de drones, 10 millions à la R&D en matière d'armement, 10 millions à la collecte et à l'analyse de renseignements, 10 millions au progrès scientifique général, et ainsi de suite. Une nation dotée d'une IA puissante face à une autre qui n'en dispose pas – ou même face à une nation ayant trois ans de retard en matière d'IA – cela revient à mettre face à face une armée de Marines de la Seconde Guerre mondiale et des chevaliers du Moyen Âge qui se battent à l'épée.

Si une IA puissante permet des formes plus profondes et potentiellement permanentes de répression autocratique (voir section 4), il est d'autant plus important que les nations les plus puissantes du monde soient des démocraties – ou, a minima, qu'il existe des protections solides contre la répression induite par l'IA. Cela va de pair avec l'urgence d'une stratégie géopolitique ciblée.

> À plusieurs reprises, Dario Amodei a exposé sa lecture de la rivalité avec la Chine. Dans « The Adolescence of Technology », il écrit : « La Chine est aujourd'hui la deuxième puissance mondiale en matière de capacités d'IA, derrière les États-Unis, et le pays le plus susceptible de les dépasser dans ce domaine. Son gouvernement est actuellement autoritaire et s'appuie sur un appareil de surveillance technologique avancé. [...] J'ai souvent écrit sur la menace que représenterait une prise de leadership du Parti communiste chinois dans l'IA, et sur l'impératif existentiel de l'en empêcher. »

Dans ce but, les démocraties devraient chercher à former une coalition mondiale, axée sur un développement de l'IA conforme à leurs valeurs communes, en s'efforçant d'attirer à elles le reste du monde. Pour cela, il sera nécessaire de rendre l'adhésion à cette coalition de plus en plus attrayante et, à l'inverse, le fait de ne pas en être de moins en moins attrayant. Cette coalition devrait consister en une internationalisation coordonnée des idées sur les politiques en matière d'IA évoquées dans les sections 1 à 4, ainsi qu'en un effort visant à sécuriser la chaîne d'approvisionnement essentielle au développement de l'IA, en la partageant au sein de la coalition et en la refusant à ceux qui en sont exclus. Voici quelques principes et objectifs opérationnels possibles :

METTRE EN PLACE UNE COGESTION DE LA CHAÎNE D'APPROVISIONNEMENT EN IA

Les membres de cette coalition fondée sur la confiance pourraient partager librement entre eux les puces et les équipements de fabrication de semi-conducteurs (SME), tout en collaborant pour en priver leurs adversaires. Les contrôles à l'exportation américains sur les puces de pointe et les SME vers la Chine ont largement contribué à l'avance globale des États-Unis en matière d'IA, et ces politiques doivent être étendues, renforcées et coordonnées avec d'autres États partageant les mêmes valeurs. Les projets de loi en cours d'examen, tels que MATCH[31] et OVERWATCH[32], constituent un bon premier pas dans ce sens, et les démocraties alliées doivent envisager des mesures similaires.

SE COORDONNER POUR FAIRE FACE AUX RISQUES LIÉS À L'IA

Les politiques visant à traiter les risques liés à la biologie, à la cybersécurité et à l'autonomie décrites dans la section 1 seront plus efficaces (et moins

contraignantes pour l'industrie) si elles sont coordonnées au niveau international.

> En février 2026, Anthropic a déclaré avoir identifié des campagnes d'extraction à échelle industrielle menées par DeepSeek, Moonshot et MiniMax pour récupérer les capacités de Claude, en violation présumée de ses conditions d'utilisation et de certaines restrictions régionales. Anthropic a également reconnu que la distillation est une pratique largement répandue et légitime lorsque les laboratoires distillent leurs propres modèles, tout en soutenant qu'elle devient illicite lorsque des concurrents l'utilisent pour acquérir des capacités sans supporter l'intégralité des coûts de développement. Les grands laboratoires — OpenAI, Anthropic, Google, entre autres — ont déjà des échanges et, dans certains cas, des formes de coordination sur la protection de leurs modèles, notamment contre la distillation non autorisée, l'exfiltration de capacités et les atteintes à leur propriété.

Cela permettrait aux entreprises de se conformer à des normes compatibles et aux régulateurs d'apprendre les uns des autres comment mesurer et atténuer au mieux ces risques. Les forces de l'ordre et les agences de renseignement devraient également collaborer plus étroitement pour suivre et neutraliser les menaces d'utilisation abusive, telles que les avancées des terroristes dans la fabrication d'armes biologiques à l'aide de l'IA, par exemple.

PERMETTRE AU PLUS GRAND NOMBRE DE BÉNÉFICIER DES AVANTAGES DE L'IA

Les politiques commerciales et réglementaires peuvent être utilisées pour faciliter une diffusion plus rapide des avantages économiques de l'IA au sein de la coalition, en partageant les enseignements sur la manière d'accélérer l'innovation. La coordination des approches en matière de déploiement bénéfique pourrait contribuer à faire profiter les pays en développement des avantages de l'IA. Par exemple, l'harmonisation des régimes d'autorisation médicale pourrait permettre des essais et une homologation

plus rapides et plus efficaces des médicaments basés sur l'IA (comme évoqué dans la section 3).

INSTAURER UNE DÉFENSE MUTUELLE

Les pays membres de la coalition devraient collaborer pour se défendre mutuellement grâce à l'IA et contre l'IA de leurs adversaires. La coalition devrait garantir collectivement une production suffisante de cyberdéfenses basées sur l'IA, de drones alimentés par l'IA, de processus de fabrication pilotés par l'IA, de capacités de calcul classifiées basées sur l'IA, de R&D pilotée par l'IA, ainsi que le partage de renseignements collectés grâce à l'IA.

INTERDIRE LES SYSTÈMES RÉPRESSIFS ALIMENTÉS PAR L'IA

Les membres de la coalition se doivent de rejeter la tyrannie high-tech, ultra-répressive et alimentée par l'IA contre laquelle j'ai mis en garde dans « The Adolescence of Technology[33] », et doivent disposer de garanties similaires à celles que j'ai décrites dans la section 4 ci-dessus.

FAVORISER UNE COOPÉRATION MACROÉCONOMIQUE

Les crises de l'emploi ou de la stabilité de l'emploi, comme toute autre crise économique, peuvent se propager au-delà des frontières. Les pays ont donc un intérêt commun à travailler ensemble pour coordonner les politiques de soutien et de stabilisation macroéconomiques, telles que celles décrites dans la section 2, afin de contrer tout effet sur l'emploi.

L'objectif est donc de rendre l'adhésion à la coalition aussi attrayante que possible – et de mettre clairement en évidence les coûts liés au fait de rester en dehors de celle-ci. La coalition reposerait ainsi sur une coordination entre États souverains, chaque nation conservant la pleine autorité sur ses propres affaires. Elle pourrait se développer par étapes, en commençant par les démocraties partageant les mêmes idéologies (qui seront naturellement enclines à y adhérer) et en accueillant progressivement des pays moins naturellement alignés, mais prêts à respecter les normes de la coalition en échange des avantages considérables liés à cette entrée dans l'alliance. Idéalement, le monde entier finirait par y adhérer. Sans y parvenir, la coalition place néanmoins les démocraties dans la position la plus forte pour contenir et vaincre les régimes qui restent dépendants d'une répression de leur population.

Une fenêtre d'opportunité

Les progrès exponentiels de l'IA constituent une urgence et imposent un rythme de changement auquel le processus décisionnel est mal préparé. Mais ils ont également ouvert une fenêtre d'opportunité unique. La convergence de preuves claires et actuelles[34] des risques liés à l'IA, un premier aperçu du potentiel de l'IA tant en termes de création de valeur économique[35] que de perturbation économique[36], et une remarquable levée de boucliers publique[37] contre les approches non réglementées de l'IA ont créé une situation où les décideurs politiques sont exceptionnellement ouverts à des mesures tournées vers l'avenir[38]. Sylvebarbe et sa forêt se réveillent.

Il est devenu courant dans les cercles industriels spécialisés dans l'IA de considérer ces réactions comme un simple problème de relations publiques : cela consiste à dire que l'IA a besoin d'un « meilleur marketing ». Je rejette totalement cette façon de voir les choses. Les gens s'inquiètent de l'IA parce qu'ils perçoivent *à juste titre* que ses risques sont réels, et non parce que les PDG du secteur de l'IA n'ont pas été suffisamment panglossiens. Je crois qu'il est de mon devoir, en tant que leader de l'IA, de continuer à faire preuve de transparence sur ces risques, et que l'inquiétude du public en réponse à cette transparence constitue une responsabilité démocratique qui fonctionne comme elle le devrait. Le principal défi consiste à canaliser cette inquiétude vers des solutions constructives et à ne pas la laisser dégénérer en une colère et une violence aveugles.

> Daniela Amodei, présidente et cofondatrice d'Anthropic, illustre ce risque par l'exemple des réseaux sociaux : leur encadrement n'est venu qu'après-coup, par des régulations tardives, sous la pression de crises successives au vu des atteintes à la santé mentale des jeunes, des risques pour la protection des enfants, de la manipulation de l'intégrité électorale.

Je suis optimiste quant à la possibilité de trouver des solutions, car bon nombre de ces défis – qu'il s'agisse de la suppression d'emplois, des tests de modèles avant leur mise sur le marché, des contrôles à l'exportation des puces, ou d'autres questions politiques liées à l'IA telles que la consommation d'énergie – font l'objet d'un consensus de bon sens à travers tout le spectre politique. Il existe un monde futur ambitieux mais réaliste dans lequel une large coalition non partisane, motivée par la reconnaissance directe des défis posés par l'IA, peut conduire à l'adoption

de politiques sensées et tournées vers l'avenir à un rythme bien plus rapide que ce que nous avons connu. Plus tôt nous y parviendrons, plus tôt nous pourrons tous profiter des avantages incroyables [39] de l'IA.

Je tiens à remercier Allan Dafoe, Mariano-Florentino Cuéllar, Richard Fontaine, Buddy Shah, Vas Narasimhan, Matt Yglesias, Nick Beckstead, Jason Matheny, Brad Carson ainsi que de nombreux membres de l'équipe d'Anthropic pour leurs commentaires et leurs remarques sur les versions préliminaires de ce texte.

SOURCES

1. Marina Favaro et Jack Clark, « When AI builds itself », Anthropic, mai 2026. ↑
2. Jared Kaplan, Sam McCandlish, Tom Henighan et al., « Scaling Laws for Neural Language Models », Cornell University, 23 janvier 2020. ↑
3. Dario Amodei, « Machine of Loving Grace, How AI Could Transform the World for the Better », site de Dario Amodei, octobre 2024. ↑
4. « Project Glasswing », Anthropic, 7 avril 2026. ↑
5. Newton Cheng, Keane Lucas, Winnie Xiao, Nicholas Carlini et Milad Nasr, « Measuring LLMs' ability to develop exploits », red.anthropic.com, 22 mai 2026. ↑
6. Kyla Guru, Alex Moix et Jacob Klein, « Mapping AI-enabled cyber threats : Insights from the LLM ATT&CK Navigator », red.anthropic.com, 3 juin 2026. ↑
7. Marina Favaro et Jack Clark, « When AI builds itself », Anthropic, mai 2026. ↑
8. J'aborde notamment les risques biologiques et ceux liés à l'autonomie dans mon essai The Adolescence of Technology. L'Anthropic Institute a également publié des données internes préliminaires dans When AI Builds Itself concernant la possibilité d'une auto-amélioration récursive, c'est-à-dire de modèles capables, de manière autonome, de créer de meilleurs modèles. ↑
9. Voir la page Wikipédia de Friedrich Hayek ↑
10. Voir la page Wikipédia du Dilemme de Collingridge. ↑
11. Ce phénomène n'est pas seulement théorique : nous l'avons observé à maintes reprises dans nos propres cadres de gouvernance volontaires, comme notre Politique de développement responsable. Si nous nous imposons une liste fixe ou rigide d'exigences de sécurité pour les futurs modèles d'IA, il est très probable que des exigences qui s'avèrent finalement peu importantes finissent par absorber 95 % de nos efforts de mise en conformité, tandis que nous découvrirons en même temps que certaines des plus grandes sources de risque n'avaient pas du tout été anticipées dans notre liste. Les cadres volontaires peuvent être modifiés et adaptés, mais cela est beaucoup plus difficile avec la législation. Mes tentatives pour résoudre ce dilemme sont exposées dans mes deux lettres publiques concernant le projet de loi SB 1047, une loi californienne de 2024 qui visait à traiter les risques de catastrophes et à l'égard de laquelle j'avais des sentiments mitigés pour les raisons évoquées ci-dessus. ↑
12. « Anthropic is endorsing SB 53 », Anthropic, 8 septembre 2025. ↑
13. The New York State Senate, « Assembly Bill A6453A, Relates to the training and use of artificial intelligence frontier models », 2025-2026 Legislative Session. ↑
14. « Illinois Senate Bill 315 », LegiScan GAITS, 29 mai 2026. ↑

15 Dario Amodei, « Anthropic C.E.O. : Don't Let A.I. Companies off the Hook », The New York Times, 5 juin 2025. ↑

16 « Promoting advanced artificial intelligence innovation and security », site officiel de la Maison-Blanche, 2 juin 2026. ↑

17 Gillian K. Hadfield et Jack Clark, « Regulatory Markets : Future AI Governance », American Bar Association, 19 novembre 2025. ↑

18 Par exemple, les risques biologiques vraiment graves peuvent s'avérer bien plus difficiles à gérer que les risques cybernétiques, car ceux qui attaquent disposent d'un avantage considérable sur ceux qui sont en position défensive et l'ampleur d'une catastrophe peut être bien plus importante. ↑

19 Marina Favaro et Jack Clark, « When AI builds itself », Anthropic, mai 2026. ↑

20 Pour une analyse plus détaillée des raisons pour lesquelles la logique qui a conduit à une reprise rapide du marché de l'emploi et à l'absence de suppression durable d'emplois dans d'autres secteurs technologiques pourrait ne pas s'appliquer à l'IA, et en particulier des raisons pour lesquelles les mécanismes d'adaptation habituels, tels que le paradoxe de Jevons ou l'avantage comparatif, pourraient être dépassés par le rythme de cette technologie, voir The Adolescence of Technology. ↑

21 À titre d'exemple, il y a encore des gens qui consacrent leur vie aux échecs, au jeu de go ou à l'alpinisme, et qui sont toujours admirés pour ces pratiques, alors que les machines sont désormais plus performantes qu'eux dans tous ces domaines. ↑

22 « Anthropic Economic Index : Understanding AI's effects on the economy », Anthropic, dernière mise à jour en date du 24 mai 2026. ↑

23 Cela incite davantage les personnes à changer d'emploi et à se former pour évoluer vers une nouvelle carrière, même si cela peut s'avérer difficile à court terme, en leur versant la différence entre leur ancien et leur nouveau salaire, si ce dernier est inférieur. ↑

24 « Covering electricity price increases from our data centers », Anthropic, 11 février 2026. ↑

25 Jörg J. Möhrle, « How long does it take to develop a new drug ? », The Lancet Regional Health – Europe, Volume 43, 2024. ↑

26 Dario Amodei, « Machine of Loving Grace, How AI Could Transform the World for the Better », site de Dario Amodei, octobre 2024. ↑

27 Voir la page Wikipédia du Posse Comitatus Act. ↑

28 Voir la page Wikipédia du Foreign Intelligence Surveillance Act. ↑

29 Dario Amodei, « The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI », site de Dario Amodei, janvier 2026. ↑

30 Voir la page Wikipédia de la Compagnie britanniques des Indes orientales. ↑

31 « H.R.8170 – MATCH Act », Congress.gov, dernière mise à jour en date du 22 avril 2026. ↑

32 « H.R.6875 – AI OVERWATCH Act », Congress.gov, dernière mise à jour en date du 18 décembre 2025. ↑

33 Dario Amodei, « The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI », site de Dario Amodei, janvier 2026. ↑

34 « Project Glasswing : Initial Update », Anthropic, 22 mai 2026. ↑

35 « What 81,000 people told us about the economics of AI », Anthropic, 22 avril 2026. ↑

36 Eli Tan, Kalley Huang et Mike Isaac, « Meta Lays Off 8,000 Employees, as A.I. Casualties Mount », The New York Times, 19 mai 2026. ↑

37 Madison Mills, « An AI hate wave is here », Axios, 17 mai 2026. ↑

38 « Promoting advanced artificial intelligence innovation and security », site officiel de la Maison-Blanche, 2 juin 2026. ↑

39 Dario Amodei, « Machine of Loving Grace, How AI Could Transform the World for the Better », site de Dario Amodei, octobre 2024. ↑