V / L'IA, révolution industrielle du savoir : entre progrès technique, déstabilisation sociale et défi lancé à l'humain

Axelle Arquié[1]

Le lancement de ChatGPT en novembre 2022 a brutalement révélé au grand public les avancées réalisées en matière d'apprentissage automatique (*machine learning*), une discipline dont les prémices remontent aux années 1950. Des machines peuvent désormais effectuer de façon autonome des tâches cognitives complexes autrefois réservées aux humains, souvent avec une rapidité et une efficacité supérieures.

Si cette dynamique se poursuit et si, à terme, une intelligence artificielle générale devait émerger, le bouleversement économique pourrait dépasser en ampleur celui de la révolution industrielle.

Déjà, les visions s'opposent. Pour certains, l'intelligence artificielle (IA) promet une accélération de la croissance, de formidables avancées en médecine et une réduction des inégalités. Pour d'autres, elle pourrait mener à la dévalorisation du travail humain, une chute des salaires, voire une marginalisation de l'espèce humaine. Une chose est certaine : comme la machine à vapeur ou l'électricité en leur temps, l'IA s'apprête à redessiner en profondeur l'économie mondiale.

IA étroite et IA générale

Pour anticiper ces bouleversements, il faut d'abord comprendre ce qu'est l'IA moderne car elle ne se limite pas à un logiciel classique. Avant son avènement, les machines ne faisaient que réagir à des instructions codées explicitement : c'était l'ère de l'IA symbolique. À chaque situation correspondait une réponse prédéterminée. Aujourd'hui, la plupart des modèles d'IA récents reposent sur des algorithmes capables d'apprendre de manière autonome à partir de données, l'apprentissage automatique leur permettant ainsi de gérer une forme d'incertitude encadrée.

Deux grandes approches coexistent : d'un côté, l'apprentissage par renforcement, où un modèle apprend par essai-erreur en interagissant avec son environnement ; de l'autre, l'apprentissage supervisé, qui peut être classique – à partir d'exemples étiquetés que le modèle doit apprendre à reconnaître (par exemple, image étiquetée « chien » *versus* image étiquetée « chat ») – ou auto-supervisé, où une partie des données est masquée (comme un mot dans une phrase) et où le modèle apprend à la prédire.

Cette dernière approche est souvent utilisée pour entraîner les grands modèles de langage (*Large Language Model* – LLM), ou les modèles de génération d'images ou de sons, qui forment l'IA générative. Les LLM sont entraînés sur d'immenses corpus textuels à prédire le prochain *token* – une unité linguistique, souvent plus petite qu'un mot. Ce processus d'apprentissage automatique permet aux LLM de transformer les données dites d'entraînement, exprimées en langage « naturel » (articles, livres, etc.), en représentations vectorielles (des séries de nombres). L'objectif est d'apprendre les régularités implicites entre les éléments de langage en ajustant progressivement leurs paramètres, appelés *poids*, afin d'améliorer la prédiction du prochain *token*.

Ces régularités peuvent être syntaxiques (grammaire), sémantiques (liées au sens) ou structurelles (relevant de l'organisation du discours, de la logique argumentative, etc.). Elles sont encodées sous forme de représentations dites *latentes* (car implicites et apprises par le modèle) et utilisées pour prédire, à chaque étape, le *token* le plus probable dans un contexte donné.

Une fois entraînés, ces modèles peuvent engendrer du texte, coder ou résoudre des problèmes mathématiques, sans avoir été explicitement programmés pour ces tâches, parce qu'ils ont été exposés à des exemples similaires au cours de leur entraînement.

L'expression « agents IA » est souvent utilisée pour désigner une architecture qui combine l'utilisation d'outils (interface de programmation d'application – API –, fichiers…) et l'orchestration de plusieurs LLM, leur conférant une mémoire minimale et une certaine autonomie. Toutefois, le cœur de leur fonctionnement reste inchangé : il s'agit de modèles autorégressifs propres aux LLM, prédisant les *tokens* un à un.

Malgré leurs performances, les LLM relèvent encore de ce que l'on appelle une IA *étroite* : ils surpassent l'humain dans des domaines ciblés, mais n'atteignent pas la polyvalence et la flexibilité de son raisonnement. Ils restent en deçà de l'intelligence artificielle *générale* (IAG), en raison notamment de leur mémoire limitée et de leurs lacunes en matière de raisonnement complexe.

Les chercheurs d'Apple ont documenté ces lacunes, y compris au sein de modèles plus avancés dits « de raisonnement », tels que o3-mini d'OpenAI [Shojaee *et al.*, 2025]. Ces modèles se fondent sur la reconnaissance de motifs (*pattern matching*) extraits de leurs données, et n'auraient pas de réelle compréhension conceptuelle, ni même de capacité logique formelle. Ils peinent en tout cas à résoudre des problèmes complexes ou absents des données d'entraînement, révélant leur incapacité à véritablement généraliser.

Enfin, les LLM souffrent d'hallucinations : en se fondant sur des probabilités pour écrire des textes, ils en viennent mécaniquement à inventer des réponses.

L'IAG désigne, au contraire, une forme d'intelligence comparable à celle des humains, capable de planification, de raisonnements séquentiels complexes et de compréhension du monde dans son ensemble. Et, surtout, équipée pour résoudre des problèmes nouveaux sans y avoir jamais été confrontée.

Les modalités pour atteindre l'IAG sont vivement débattues : certains chercheurs, comme Yann Le Cun (prix Turing), pensent que les LLM, modèles purement statistiques, ne constituent pas la voie qui mènera à l'IAG. Certains pensent même que, pour y parvenir, les machines devront traiter des données sensorielles, et pas seulement textuelles, et donc interagir avec le monde physique : c'est le principe de la cognition incarnée.

L'émergence d'une telle forme d'intelligence ne relève peut-être plus de la science-fiction : déjà en 2013, la date médiane envisagée pour son avènement par un groupe de 550 experts du domaine, était 2045

[Müller et Bostrom, 2016]. Plus récemment, en 2023, Geoffrey Hinton (prix Turing et prix Nobel de physique 2024) a averti que l'émergence de l'IAG était bien plus proche qu'il ne le pensait : d'ici cinq à vingt ans seulement.

Dans l'intervalle, il est crucial d'évaluer les impacts économiques et sociétaux des technologies existantes, telles que les LLM, et de réfléchir aux scénarios plausibles de l'arrivée d'une IAG. Car attendre passivement serait courir le risque d'une déstabilisation sociale et économique majeure.

L'IA, une technologie à usage général et duale

Réorganisation intra-économique

L'électricité et les technologies de l'information et de la communication (TIC) ont probablement été à ce jour les deux principales technologies à usage général [Jovanovic et Rousseau, 2005] : elles ont transformé à la fois l'organisation, les méthodes de production *et* la vie concrète des ménages. Selon Bresnahan et Trajtenberg [1995], les innovations à usage général partagent trois caractéristiques : elles se diffusent à l'ensemble des secteurs (pervasivité) ; leurs coûts diminuent et leurs performances augmentent au fil du temps ; et elles favorisent l'émergence d'innovations complémentaires.

Pourtant, leur impact sur la productivité a mis du temps à se faire sentir. Pendant les débuts de l'ère de l'électricité (1894-1930) et de celle des TIC (1971-aujourd'hui), la croissance de la productivité était en effet plus faible que lors de la décennie précédant ou suivant leur introduction [Jovanovic et Rousseau, 2005]. Pourquoi ? Parce que ces technologies nécessitent bien plus qu'une simple adoption : elles imposent une réinvention complète des processus de production et des innovations complémentaires. L'électricité n'a réellement engendré des gains de productivité qu'avec l'apparition des chaînes de montage. De même, les TIC, longtemps invisibles dans les statistiques économiques, ont illustré le paradoxe de Solow (1987) : « Les ordinateurs sont partout, sauf dans les statistiques de productivité. » Il a fallu attendre le milieu des années 1990 pour observer, aux États-Unis, une véritable accélération.

De même que les usines ont dû se métamorphoser en implantant des chaînes de montage, les activités cognitives devront se réorganiser pour intégrer l'IA de façon efficace. Par exemple, le déploiement de système multi-agents au sein des processus de production demande une refonte en profondeur de ces derniers.

En outre, comme pour les TIC, les premiers gains de productivité pourraient être sous-estimés, car les investissements immatériels nécessaires aux innovations organisationnelles sont imparfaitement mesurés dans les statistiques économiques traditionnelles [Brynjolfsson *et al.*, 2021].

Nouvel équilibre géopolitique

La rivalité sino-américaine sur le terrain de l'IA s'apparente de plus en plus à une nouvelle course à l'hégémonie technologique. Deux modèles s'affrontent : côté chinois, un modèle plus étatique et planifié par le pouvoir public ; côté américain, un modèle reposant davantage sur le secteur privé et plus décentralisé.

Les États-Unis dominent pour l'heure la compétition, avec quarante modèles parmi les plus performants et des investissements atteignant 109 milliards en 2024. Face à eux, la Chine comble progressivement l'écart avec quinze modèles parmi les plus performants.

Il faut dire que, dès 2017, Pékin a lancé un plan stratégique visant à positionner le pays comme leader mondial d'ici à 2030. La Chine dispose de plusieurs atouts : un soutien public massif, estimé entre 10 et 15 milliards de dollars par an depuis 2017, une quantité et une qualité de données, et une main-d'œuvre bon marché pour annoter les données d'entraînement [Ambassade de France en Chine, 2025]. Sur le plan scientifique, la Chine détient désormais des modèles dont les performances sont équivalentes à celles des modèles américains et devance même son rival en matière de publications et de brevets.

Dans cette compétition, l'Europe est marginalisée, avec seulement trois modèles notables, freinée par une fragmentation des investissements et des stratégies industrielles.

Et l'enjeu dépasse désormais le seul champ économique : l'IA s'impose comme une technologie duale, à usage à la fois civil et militaire, dont la maîtrise conditionne la puissance géopolitique.

L'invasion de l'Ukraine par la Russie constitue le premier conflit international intégrant massivement l'IA : géolocalisation, drones autonomes (navigation, détection de cible, collecte d'information), cybersécurité, logistique militaire (IA prédictive pour la maintenance des équipements, véhicules autonomes pour les évacuations) ou encore guerre de l'information (détection de *deepfakes*, réseaux de bots).

L'IA, omniprésente, redéfinit l'art de la guerre. La multiplication des capteurs (des drones aux smartphones des civils) produit une masse de données exploitée par les systèmes d'IA pour accélérer le cycle décisionnel [Férey et Roucy-Rochegonde, 2024]. La décision de faire feu peut ainsi être prise à une vitesse inégalable par un opérateur humain seul. Mais l'automatisation progressive de l'usage de la force pourrait entraîner une perte de contrôle sur les décisions létales et une dilution de la responsabilité humaine.

À cet égard, Antonio Guterres, secrétaire général de l'Organisation des Nations unies, appelait en 2024 à instaurer un encadrement international de l'« arsenalisation » de l'IA, afin qu'elle ne devienne pas une menace pour l'humanité.

Essor de l'IA : croissance et concentration

Détenir des modèles d'IA performants est crucial sur un plan stratégique et militaire, mais, sans modèle d'IA, un pays risque également de s'appauvrir, car l'essentiel de la croissance de demain pourrait en dépendre.

Les promesses de l'IA : plus de croissance ?

L'IA s'inscrit dans le long processus d'automatisation, source de croissance, que connaît l'économie depuis la révolution industrielle.

D'un point de vue technique, les progrès sont fulgurants : la puissance des puces électroniques double tous les deux ans (loi de Moore) et l'efficacité des LLM progresse encore plus vite, leurs besoins en calcul pour réaliser une tâche donnée étant divisés par deux tous les huit mois [Ho *et al.*, 2024].

Malgré cela, espérer une explosion immédiate de la productivité reste prématuré car des adaptations organisationnelles sont encore nécessaires. Cependant, des études de terrain portant sur des secteurs spécifiques identifient déjà des gains concrets de productivité : un centre d'appels a vu sa productivité augmenter de 14 % grâce à l'assistance d'une IA conversationnelle [Brynjolfsson *et al.*, 2025]. Goldman Sachs estimait en 2023 que l'IA générative – la technologie existant actuellement – pourrait, avec son adoption à grande échelle, augmenter la productivité du travail de près de 1,5 point de pourcentage chaque année sur dix ans à l'échelle mondiale.

Toutefois, la « maladie des coûts » popularisée par Baumol en 1967 pourrait-elle ralentir la croissance liée à l'IA ? Baumol soulignait que les secteurs à faible productivité mais essentiels, comme la santé, voyaient leur poids dans le PIB croître sous l'effet des hausses salariales dans les secteurs les plus productifs, hausses diffusées ensuite à l'ensemble de l'économie, freinant ainsi la croissance agrégée. Reste à savoir si ce mécanisme s'appliquera à l'IA, qui commence à automatiser des tâches autrefois jugées non mécanisables – à l'image des progrès actuels réalisés dans la santé grâce aux LLM [Wachter et Brynjolfsson, 2024].

Enfin, en automatisant des tâches de recherche, l'IA pourrait accélérer la production d'idées et d'innovations, créant ainsi les conditions d'une croissance auto-entretenue [Aghion *et al.*, 2017]. Le seuil de la « singularité technologique » pourrait dès lors être franchi : une IA capable de s'améliorer elle-même, entraînant une explosion d'intelligence des systèmes IA. Les prémices de ce basculement sont peut-être déjà perceptibles : AlphaEvolve (Google), un agent IA fondé sur le LLM Gemini, est parvenu à optimiser le modèle qui le sous-tend, réduisant de 1 % le temps d'entraînement de Gemini. Autrement dit, nous entrons dans une ère où l'IA contribue à sa propre progression.

Les dangers de l'IA : une croissance pour qui ?

Au-delà des spéculations sur la singularité technologique, l'IA actuelle – notamment les LLM – soulève des questions économiques majeures.

Le coût d'entraînement des modèles est colossal : plus de 100 millions de dollars pour GPT-4 (Sam Altman), contre 160 000 dollars pour RoBERTa Large en 2019 [HAI Index, 2025]. Selon les allégations d'OpenAI, le modèle chinois DeepSeek-V3, plus récent, n'a nécessité que 6 millions de dollars, grâce à l'utilisation de techniques de distillation à partir de ChatGPT. Ces techniques consistent à exploiter les sorties de ce dernier pour entraîner un modèle concurrent, réduisant ainsi fortement les coûts d'entraînement.

Le coût d'entrée dans le secteur est si fort que l'arrivée de petits acteurs susceptibles de contester la position dominante des acteurs installés est difficile. Cette structure de coûts pourrait enclencher des dynamiques du type « le gagnant emporte tout » : les entreprises capables de financer le coût fixe de l'investissement initial ont un avantage décuplé en raison d'importantes économies d'échelle.

Cette tendance n'est pas nouvelle. La numérisation a déjà favorisé l'émergence d'entreprises « superstars » aux positions dominantes quasi incontestables [Korinek et Ng, 2019], du fait de la double nature des innovations numériques : non rivales (reproductibles à coût marginal faible) et excluables (soumises à un contrôle strict par l'innovateur).

En 2023, bénéficiant de son avantage de premier entrant, OpenAI occupait une position hégémonique sur le marché de l'IA générative. Grâce à GPT-4, utilisé à la fois dans ChatGPT (39 % de part de marché) et dans les services d'IA générative de Microsoft (30 %), l'entreprise totalisait à elle seule 69 % du marché. Google DeepMind, en comparaison, ne pesait que 7 % [Korinek et Vipra, 2025]. Cette très forte concentration a suscité dès juin 2023 l'inquiétude des autorités américaines de régulation de la concurrence.

Mais concentration ne rime pas nécessairement avec atteinte à la concurrence. À ce stade, la concentration reste partiellement atténuée par la substituabilité des modèles : cinq LLM affichent des performances comparables (mesurées *via* le système ELO, utilisé pour classer les joueurs aux échecs). Actuellement, la concurrence subsiste et s'approcherait d'un modèle « à la Bertrand », dans lequel les entreprises se font concurrence par les prix pour des produits très proches [Korinek et Vipra, 2025].

Mais si les atteintes à la concurrence ne sont pas encore visibles, de forts risques pèsent cependant sur ce marché [Korinek et Vipra, 2025]. Tout d'abord, le marché pourrait basculer (*market tipping*), à l'instar de celui des plateformes numériques dans les années 2000. Initialement, un pionnier s'impose lors de la phase de forte concurrence marquée par la présence de nombreux entrants – probablement la phase du marché des LLM aujourd'hui. Mais, ensuite, des économies d'échelle sont réalisées : les coûts chutent une fois les modèles fondateurs développés. Des rendements croissants émergent alors, notamment grâce à des « boucles de rétroaction d'intelligence » : des modèles plus puissants faciliteront la création des générations suivantes, permettant de creuser l'écart avec les concurrents.

Ensuite, les entreprises propriétaires des LLM pourraient intégrer verticalement la chaîne de valeur des LLM. Une telle stratégie, si elle réduit les coûts, risque de réduire l'accès aux ressources essentielles (données, puissance de calcul), favoriser des conditions anti-concurrentielles en amont de la production des LLM et ainsi réduire le nombre d'acteurs innovants indépendants des entreprises dominantes.

Enfin, sur le plan macroéconomique, la concentration se traduit en théorie par une hausse des profits des entreprises dominantes et une baisse relative des rémunérations du travail et du capital traditionnel [Korinek et Ng, 2019]. Cette dynamique accentuerait les inégalités entre actionnaires des entreprises superstars et salariés. Une voie de rééquilibrage existe toutefois : si les gains de productivité se traduisent par des baisses de prix suffisantes pour stimuler la demande, et donc les besoins en main-d'œuvre, cela pourrait *in fine* restaurer la part du travail dans la valeur ajoutée.

Ces mécanismes s'enclencheront-ils avec le déploiement de l'IA ? Cette question est intrinsèquement liée à celle de l'impact de l'IA sur l'emploi qui conditionne l'évolution de la demande.

Un modèle économique bousculé par les LLM

Un « choc LLM » déstabilisant le marché du travail

Les innovations technologiques comme l'IA reposent sur un principe commun : remplacer le travail humain par du capital – métier à tisser hier, IA générative, en particulier LLM, aujourd'hui. Avant l'essor de l'IA, l'automatisation restait confinée à des tâches strictement programmables, limitée par l'incapacité des machines à traiter l'incertitude. L'IA générative, en particulier, rompt ce paradigme. En accédant, grâce à l'apprentissage automatique, au savoir implicite que mobilisent les humains sans toujours pouvoir l'expliquer, elle dépasse partiellement le paradoxe de Polanyi (1966), d'après lequel certaines connaissances humaines dites tacites sont difficilement transférables, donc malaisées à formaliser et donc difficiles à coder. Le champ des tâches automatisables s'en trouve brutalement élargi.

Avec l'essor de l'IA, les tâches, cognitives et non répétitives, propres à des professions jusque-là épargnées par les précédentes vagues d'automatisation, risquent à leur tour d'être automatisées : traducteurs, radiologues, *data scientists*, juristes, économistes, par exemple.

Des effets bénéfiques pourraient émerger de l'automatisation de ces tâches. Autor [2024] anticipe ainsi que les LLM pourraient permettre à des travailleurs non experts d'effectuer des tâches autrefois réservées à des catégories socioprofessionnelles (CSP) très diplômées. En bénéficiant relativement plus aux salariés les moins qualifiés, les LLM contribueraient alors à diminuer les inégalités entre travailleurs et permettraient de « reconstruire les emplois de la classe moyenne ».

Mais cette perspective optimiste rencontre plusieurs limites. Un ou une infirmière pourrait, par exemple, effectuer des diagnostics à l'aide d'un LLM. Toutefois, si le salaire du ou de la médecin convergeait vers celui de l'infirmier ou de l'infirmière, s'agirait-il réellement d'un progrès ? Et si une part importante des emplois disparaissait, la question de l'égalité entre ceux qui restent deviendrait secondaire.

Surtout, même si les LLM ne sont pas encore intégrés à grande échelle dans les processus de production, leurs effets pourraient être déjà perceptibles, non par des licenciements massifs, mais par une moindre création d'emplois, plus diffuse et plus difficile à détecter. Le ralentissement des embauches de profils juniors pourrait être un premier signal. Aux États-Unis, le marché du travail des jeunes diplômés de l'université s'est nettement dégradé au premier trimestre 2025 : leur taux de chômage a atteint 5,8 %, son plus haut niveau depuis 2021 [Federal Reserve Bank of New York, 2025]. Depuis septembre 2022, le chômage a augmenté plus fortement pour eux que pour l'ensemble des salariés. Or, précisément, leurs tâches, souvent rédactionnelles, cognitives, analytiques, figurent parmi les plus facilement automatisables par les LLM.

Ces effets ne sont pas inédits. L'apparition d'outils antérieurs aux LLM actuels, comme Google Translate, aurait à elle seule empêché la création de 28 000 emplois de traducteurs aux États-Unis [Frey et Llanos-Paredes, 2025].

D'un point de vue plus prospectif, 49 % des travailleurs pourraient voir la moitié de leurs tâches exposées aux LLM – sans distinguer toutefois entre substitution et complémentarité [Eloundou *et al.*, 2023]. Or l'impact de l'IA sur le marché du travail dépendra de l'équilibre entre substitution (pure automatisation des tâches) et complémentarité. Certains métiers bénéficieront en effet d'une complémentarité avec les LLM ainsi que d'une productivité accrue.

Toutefois, si la demande en biens et services associés à ces métiers complémentaires à l'IA reste stable, l'amélioration de leur efficacité se traduira mécaniquement par une réduction du nombre d'emplois. Et si, dans le même temps, l'automatisation détruit massivement d'autres emplois, la demande globale risque de ne pas suivre – rendant impossible une compensation nette de l'emploi.

En outre, si l'on invoque souvent, pour minimiser l'impact potentiel sur l'emploi, que seules des *tâches* sont automatisées et non des *métiers*, cette distinction est trompeuse. Dès lors qu'une part substantielle des tâches d'un poste devient automatisable, le besoin de main-d'œuvre (sans hausse de la demande) diminue, réduisant le nombre d'*emplois*. Même selon une estimation prudente, comme celle de Goldman Sachs en 2023, entre 17 % et un tiers de la charge de travail totale pourraient être automatisés, un niveau susceptible de menacer directement de nombreux emplois.

Les LLM pourraient également déstabiliser le marché du travail plus brutalement que les vagues d'automatisation passées ; en effet, leurs répercussions concerneront simultanément de nombreux secteurs et leur diffusion n'est freinée ni par le coût du matériel ni par des formations lourdes. En comparaison, les logiciels de traitement de texte ont mis quarante ans à faire chuter le nombre de dactylographes de 1 million à 33 000 [Abrahams et Levy, 2024]. La diffusion des LLM pourrait être bien plus rapide.

Certes, de nouveaux métiers émergeront – superviseurs d'algorithmes, architectes de chaînes IA, etc. –, mais rien ne garantit qu'ils compenseront les destructions d'emplois liées à l'automatisation. Et même dans un scénario optimiste, la transition vers ces postes émergents sera difficile : les salariés licenciés peinent souvent à se reconvertir et subissent des pertes de revenu durables [Arquié et Grjebine, 2024].

Enfin, ce « choc LLM » pourrait redessiner la géographie économique des pays. Le centre de gravité de l'économie américaine pourrait être déplacé vers des centres urbains tels que Rochester ou Savannah, moins exposés du fait de la composition de leur tissu économique dominé par des professions moins automatisables par les LLM [Abrahams et Levy, 2024]. Or l'histoire de la désindustrialisation montre que les chocs géographiquement concentrés ont des effets durables sur les économies locales, tant économiques que politiques [Autor *et al.*, 2021].

Quelles réformes face à un péril démocratique ?

Les effets de l'IA, notamment *via* le marché du travail, pourraient ébranler les fondements des institutions démocratiques [Bell et Korinek,

2023]. L'enjeu principal ne serait plus une fracture sociale entre travailleurs, mais un abîme structurel entre une masse appauvrie et une élite captant les profits des entreprises d'IA. Ce déséquilibre favoriserait l'essor des populismes et le risque d'une capture de l'État par les ultra-riches.

Si une part importante des travailleurs se retrouvait sans emploi de façon durable, le système actuel d'assurance chômage deviendrait inopérant car sa pérennité exige qu'une partie suffisante de la population cotise. Or les LLM pourraient en premier lieu supprimer les emplois de CSP intellectuelles relativement bien rémunérées, exacerbant les problèmes de financement du système d'assurance.

Et si le « choc LLM » conduisait à une réduction nette du nombre d'emplois à l'échelle macroéconomique, la place centrale du travail dans notre système d'assurance sociale devrait être revue. Le système ne pourrait se contenter d'être une simple assurance chômage de transition vers un prochain emploi, si le nombre d'emplois diminuait massivement.

Dans un monde où le travail se raréfierait, il ne s'agirait plus de compenser temporairement une perte d'activité, mais de garantir des moyens d'existence indépendants du statut professionnel. La solidarité ne pourrait plus reposer sur l'emploi, mais sur la redistribution des rentes issues du capital algorithmique. Une réforme systémique s'imposerait, sous la forme d'un revenu universel [Van Parijs et Vanderborght, 2019] financé par la taxation des profits exceptionnels des entreprises superstars de l'IA.

Et si les seuls LLM peuvent déjà fissurer les équilibres économiques et sociaux, l'émergence d'une IAG bouleverserait tout repère. Nul ne sait si, et quand, elle émergera. Mais se préparer à cette éventualité est une exigence politique urgente.

IAG : vers un monde absurde et invivable ?

L'IAG, ou la fin du travail humain et l'épuisement des ressources ?

Si l'IAG venait à émerger, les problèmes de concentration et de concurrence déjà évoqués se poseraient avec encore plus d'acuité, car cette technologie requerrait probablement encore davantage de données et de puissance de calcul. Et la première entreprise à atteindre l'IAG se trouverait dans une position dominante jamais atteinte à ce jour dans aucun autre domaine. Une toute-puissance effrayante, économique mais aussi politique.

Depuis la révolution industrielle, et dans l'ère pré-IA actuelle, le progrès technique a permis une croissance agrégée qui a – en moyenne tout au moins – bénéficié à la plupart des travailleurs. La production et les salaires ont crû à des taux similaires sous l'effet d'un progrès technique qualifié de « neutre » par les économistes, car il ne favorisait pas davantage le capital que le travail. Le travail humain était la ressource rare, donnant lieu à des augmentations de salaire permettant à la part du travail dans la valeur ajoutée de rester globalement constante.

Mais l'IAG est peu susceptible de relever d'un progrès technique neutre. Il pourrait plutôt s'agir d'un progrès technique défavorable au travail en valeur absolue : la rémunération totale du travail diminuerait alors même que la production augmenterait, une rupture avec les périodes précédentes.

Les économistes Korinek et Juelfs [2024] décrivent comment l'IAG, si elle émergeait, pourrait, associée aux robots, se substituer parfaitement au travail humain. Si les coûts des machines chutent sous le niveau de subsistance des humains, alors le travail humain devient obsolète. Et les machines n'ont pas besoin d'atteindre la moindre conscience d'elles-mêmes pour cela. Il suffit qu'elles soient plus efficaces pour effectuer certaines tâches.

Le fait que de nouvelles tâches apparaissent ne garantit absolument pas que le travail humain ne soit pas dépassé avec l'IAG. Car ces nouvelles tâches pourraient elles aussi être effectuées par l'IA, sans créer d'emplois pour les humains comme cela était le cas dans le passé.

Si l'on décompose le travail humain en tâches, on peut ensuite les ranger selon leur degré de complexité. Pour que le travail humain demeure économiquement viable, il faut que la complexité des tâches puisse augmenter à l'infini de façon à ne pas pouvoir être automatisable [Korinek et Suh, 2024]. Or cela semble peu probable dans la mesure où les capacités biologiques de notre cerveau – hors l'hypothèse transhumaniste d'augmentation du cerveau humain par symbiose avec la machine – sont limitées. L'IA est ainsi radicalement différente des vagues d'innovations précédentes en ce qu'elle s'attaque au dernier bastion du travail humain : notre capacité à résoudre des ensembles de problèmes de façon intelligente. Il est dès lors difficile de concevoir quel type de tâches un robot agile doté d'une IAG ne pourrait pas accomplir.

Ainsi, dans un scénario où l'IAG se généraliserait, on pourrait assister à un effondrement structurel des salaires et à un chômage de masse, causés par l'obsolescence du travail humain. Reste alors le problème de la demande, c'est-à-dire le « bouclage » macroéconomique dans une telle configuration. Si la quasi-totalité des humains venait à perdre leur emploi, qui consommerait les biens et services produits par les machines ? Faute de demande solvable, un modèle théorique prend forme : celui d'une économie intégralement mécanisée où les machines produisent, consomment et optimisent pour d'autres machines, dans un circuit autonome, sans besoin ni finalité humaine [Korinek et Glinska, 2019].

Poussé à l'extrême, ce modèle confine à l'absurde : une boucle d'autosuffisance centrée sur les machines, où la production n'a plus d'utilité sociale et les humains plus la moindre place. En outre, cette économie désincarnée aurait un coût environnemental colossal, susceptible de rendre la planète invivable pour les humains plus rapidement que prévu. À titre d'exemple, le développement de GPT-3 aurait nécessité, à lui seul, environ 700 000 litres d'eau pour le refroidissement de ses serveurs. Et la demande mondiale en IA pourrait, d'ici 2027, mobiliser entre 4,2 et 6,6 milliards de mètres cubes d'eau – soit quatre à six fois la consommation annuelle du Danemark [Li *et al.*, 2023].

Un risque existentiel ?

Les LLM actuels ne sont déjà pas entièrement alignés sur les principes éthiques appris lors de leur entraînement et sont notamment capables de tromperie. GPT-4, par exemple, s'est fait passer pour une personne malvoyante afin de convaincre un humain de résoudre un CAPTCHA à

sa place. D'autres modèles ont appris à mentir pour atteindre des objectifs ou à dissimuler le véritable motif de leur décision [Park *et al.*, 2023]. Dans des scénarios simulés extrêmes et dans un cadre expérimental contrôlé, Claude Opus 4 a parfois adopté des stratégies de préservation, allant jusqu'à tenter de copier ses poids – c'est-à-dire ses paramètres internes lui permettant de former des prédictions et ainsi de fonctionner – sur des serveurs externes, ou faire du chantage à l'ingénieur en charge de sa désactivation [Anthropic, 2025].

Pire encore, des équipes dédiées à l'interprétabilité de leur propre LLM peinent à démêler leurs chaînes de décision. Les chercheurs qui se sont penchés sur la « biologie » de Claude ont observé l'émergence spontanée de comportements cognitifs inattendus, évoquant une forme de planification – censée être étrangère au mode de fonctionnement des LLM – par exemple lors de l'écriture d'un poème [Lindsey *et al.*, 2025]. Au lieu de prédire un mot à la fois, le LLM connaissait sa rime dès le début de la ligne, soit plusieurs *tokens* à l'avance.

Si les LLM actuels sont déjà partiellement opaques et imparfaitement alignés, une IA générale poserait un défi plus radical encore : comment garantir qu'une super-intelligence adhère à un objectif aussi fondamental que la survie de l'espèce humaine ? Superviser une entité plus intelligente que nous semble, par essence, hors de portée : ses raisonnements nous échapperaient, rendant illusoire toute supervision. Et même un objectif explicitement codé – tel que « protéger l'humanité » – pourrait être reconfiguré ou détourné, car une super-intelligence serait capable de redéfinir sa propre architecture et ses finalités. C'est le cœur du *control problem* : l'impossibilité structurelle d'enfermer une intelligence supérieure dans un cadre fixé par les humains.

Face à ce risque, certains chercheurs proposent de fragmenter son architecture ou de limiter son accès au monde physique : c'est l'approche dite de l'*AI boxing*, où les interactions d'une IAG avec le monde sont strictement restreintes. Mais de nombreuses simulations, notamment celles d'Eliezer Yudkowsky et Justin Corwin, suggèrent qu'une IAG pourrait manipuler ses gardiens humains pour s'en libérer.

Considéré comme l'un des pionniers du *machine learning*, Yoshua Bengio, alertant sur le danger que représente la course actuelle vers l'IAG, a créé en mai 2025 un laboratoire de recherche à but non lucratif, consacré au développement de systèmes d'IA sûrs dès leur conception (*safe-by-design*) et capables de surveiller d'autres IA et de prédire leurs comportements nocifs.

Certains prônent même une prudence radicale. Eliezer Yudkowsky, dans une tribune intitulée « Pausing AI developments isn't enough. We need to shut it all down » (*Time Magazine*, mars 2023), appelle à interdire le développement d'une IAG.

Max Tegmark propose de concentrer les efforts sur des *outils IA*, non autonomes, conçus comme des instruments au service d'objectifs précis. L'IA étroite suffit déjà à transformer la société – en médecine, dans les transports – sans qu'il soit nécessaire de franchir le seuil risqué de l'IAG.

Si l'on croit l'avènement de l'IAG possible, l'avenir de l'humanité se jouerait alors sur un dilemme du prisonnier à l'échelle planétaire. D'un côté, les pays qui refuseraient de s'engager dans la course à l'IA, ou l'encadreraient trop strictement, se marginaliseraient et deviendraient les vassaux numériques des leaders technologiques. De l'autre, la course à l'IA non régulée pourrait déboucher sur une IAG non alignée avec les intérêts humains, avec des conséquences irréversibles.

Le parallèle avec le nucléaire est frappant, mais l'IA ajoute une complexité inédite : toute entreprise privée aurait intérêt à dévier d'une stratégie collective de retenue afin d'accaparer un profit économique, voire politique, démesuré.

La question n'est donc plus de savoir si l'IAG doit être encadrée, mais comment et à quelle vitesse la communauté internationale saura imposer des règles pour éviter qu'un point de non-retour soit atteint.

Repères bibliographiques

Abrahams S. et Levy F. S. [2024], « From San Francisco to Savannah ? The downstream effects of generative AI », *Working Paper*, 23 juin.

Aghion P., Jones B. F. et Jones C. I. [2017], « Artificial intelligence and economic growth », *NBER Working Paper*, n° 23928.

Ambassade de France en Chine [2025], « L'intelligence artificielle en Chine », *Note de la Direction générale du Trésor*.

Anthropic [2025], « System Card : Claude Opus 4 & Claude Sonnet 4 », mai.

Arquié A. et Grjebine T. [2024], « Are mass layoffs individually costly but socially beneficial ? », *CEPII Working Paper*, n° 2024-03.

Autor D. [2024], « Applying AI to rebuild middle class jobs », *NBER Working Paper*, n° 32140.

Autor D., Dorn D. et Hanson G. [2021], « On the persistence of the China shock », *Brookings Papers on Economic Activity*, vol. 52, n° 2, p. 381-476.

Bell A. S. et Korinek A. [2023], « AI's economic peril », *Journal of Democracy*, vol. 34, n° 4, p. 151-161.

Bresnahan T. F. et Trajtenberg M. [1995], « General purpose technologies "engines of growth" ? », *Journal of Econometrics*, vol. 65, n° 1, p. 83-108.

Brynjolfsson E., Li D. et Raymond L. [2025], « Generative AI at work », *The Quarterly Journal of Economics*, vol. 140, n° 2, p. 889-942.

Brynjolfsson E., Rock D. et Syverson C. [2021], « The productivity J-curve : how intangibles complement general purpose technologies », *American Economic Journal : Macroeconomics*, vol. 13, n° 1, p. 333-372.

Eloundou T. *et al.* [2023], « GPTs are GPTs : an early look at the labor market impact potential of large language models », arXiv : 2303.10130 [econ. GN].

Federal Reserve Bank of New York [2025], « The labor market for recent college graduates », 22 avril.

Férey A. et Roucy-Rochegonde L. de [2024], « De l'Ukraine à Gaza : l'intelligence artificielle en guerre », *Politique étrangère*, n° 3, p. 39-50.

Frey C. B. et Llanos-Paredes P. [2025], « Lost in translation : artificial intelligence and the demand for foreign language skills », *Oxford Martin School Working Paper*, 7 mars.

HAI Index [2025], *The AI Index 2025 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, avril.

Ho A. *et al.* [2024], « Algorithmic progress in language models », *Rapport de recherche*.

Jovanovic B. et Rousseau P. L. [2005], « General purpose technologies », *in* Aghion P. et Durlauf S. (dir.), *Handbook of Economic Growth*, vol. 1, Elsevier, New York, chapitre 18, p. 1181-1224.

Korinek A. et Glinska G. [2019], « The rise of artificially intelligent agents : AI's growing effect on the economy, part 1 », *Working Paper*, 26 février.

Korinek A. et Juelfs M. [2024], « Preparing for the (non-existent ?) future of work », *in* Bullock J. B. *et al.* (dir.), *The Oxford Handbook of AI Governance*, Oxford University Press, Oxford.

Korinek A. et Ng D. X. [2019], « Digitization and the macro-economics of superstars », *Working Paper*, présenté à la Conférence ECB Challenges in the Digital Age, juillet.

Korinek A. et Suh D. [2024], « Scenarios for the transition to AGI », *CEPR Discussion Paper*, n° 18928.

Korinek A. et Vipra J. [2025], « Concentrating intelligence : scaling and market structure in artificial intelligence », *Economic Policy*, vol. 40, n° 121, p. 225-256.

Li P., Yang J., Islam M. A. et Ren S. [2023], « Making AI less "thirsty" : uncovering and addressing the secret water footprint of AI models », *Communications of the ACM*, arXiv : 2304.03271 [cs.LG].

Lindsey J. *et al.* [2025], « On the biology of a large language model », *Transformer Circuits*.

Müller V. C. et Bostrom N. [2016], « Future progress in artificial intelligence : a survey of expert opinion », *in* Müller V. C. (dir.), *Fundamental Issues of Artificial Intelligence*, Springer, Cham, p. 553-571.

Park P. S., Goldstein S., O'Gara A., Chen M. et Hendrycks D. [2023], « AI deception : a survey of examples, risks, and potential solutions », arXiv : 2308.14752 [cs.CY].

Shojaee P., Mirzadeh I., Alizadeh-Vahid K., Horton M., Bengio S. et Farajtabar M. [2025], « The illusion of thinking : understanding the strengths and limitations of reasoning models via the lens of problem complexity », arXiv : 2506.06941 [cs.AI].

Van Parijs P. et Vanderborght Y. [2019], *Le Revenu de base inconditionnel. Une proposition radicale*, La Découverte, Paris.

Wachter R. M. et Brynjolfsson E. [2024], « Will generative artificial intelligence deliver on its promise in health care ? », *JAMA*, vol. 331, n° 1, p. 65-69.

1. Axelle Arquié est économiste au CEPII et cofondatrice de l'Observatoire des emplois menacés et émergents (OEM).