UN REGARD D'ÉCONOMISTE SUR L'INTELLIGENCE ARTIFICIELLE

Christian BIALÈS
Professeur honoraire de Chaire Supérieure en Économie et Gestion
Secrétaire élu du Conseil fédéral de la Ligue de l'Enseignement Hérault

Conférence faite le 18 juin 2026 au Gazette Café à Montpellier
À l'initiative de la Vigie de la Laïcité 34

« Sur la terre, deux choses sont simples : raconter le passé et prédire l'avenir. Y voir clair au jour le jour est une autre entreprise » (Armand Salacrou)

« Le présent serait plein de tous les avenirs, si le passé n'y projetait pas une histoire « (André Gide)

Salacrou a grandement raison : quand on entend et voit tout ce qui se passe aujourd'hui à propos de l'intelligence artificielle (le salon VivaTech qui se tient en ce moment à la porte de Versailles, la volonté du Premier Ministre exprimée avant-hier de généraliser l'usage de l'IA dans les services publics et d'allouer 655 millions supplémentaires alloués à l'IA via le plan d'investissement France 2030, et bien entendu, comme si c'était un fait exprès, le numéro de la Gazette de Montpellier qui vient de sortir avec un important dossier sur l'IA, ce qui me donne très facilement l'occasion de remercier le Gazette-Café de nous accueillir ici ce soir)), l'IA est très à la mode mais, surtout, elle devient de plus en plus à la fois un argument de guerre géopolitique et une question majeure de philosophie et d'éthique ; on ne peut donc être que saisi et complètement désorienté par l'ampleur que prend le sujet au jour le jour.

Alors, pour tenir compte des réflexions d'Armand Salacrou et d'André Gide, et je traiterai en tant qu'économiste la question de l'IA en deux parties. Dans une 1ère partie, je me poserai la question plutôt historique : « l'IA est-elle une révolution industrielle comme les autres ? » et dans la seconde j'envisagerai les conséquences que cette histoire nous permet d'envisager pour l'avenir proche, car il y a trop d'incertitudes pour aller au-delà.

1ère partie : l'IA est-elle une révolution industrielle comme les autres ?

I- Avant de répondre à cette question, rappelons-nous ce qu'écrivait au début du chapitre 3 de sa « Grammaire des civilisations » parue en 1987 Fernand Braudel :

« une des responsabilités essentielles de l'Europe est d'avoir réalisé la révolution industrielle qui a couru et court l'univers. Ce formidable lancement technique est son œuvre, une œuvre récente à l'échelle de l'histoire des civilisations puisqu'elle date de deux siècles à peine.
Jusque-là, la brillante Europe ne fait figure, sur le plan matériel, que de pays sous-développé, non par rapport au monde qui l'entoure, mais par rapport à ce qu'elle allait devenir elle-même.
Alors, comment a-t-elle réussi à franchir le seuil industriel ? Comment sa civilisation a-t-elle réagi aux conséquences de sa propre création ?
Telles sont les questions qui se posent d'entrée de jeu.
Leur intérêt est actuel :
• Elles exigent des explications préalables sur l'état de l'Europe avant son industrialisation. Or cet Ancien régime économique est encore celui de bien des régions du monde qui tentent de le dépasser.
• La révolution industrielle est un phénomène compliqué ; nulle part, elle ne s'est produite en une seule fois. Des secteurs sont restés longtemps à la traîne, ainsi l'industrie lanière du Yorkshire ou la quincaillerie autour de Birmingham, jusqu'au milieu du XIXe siècle, pour ne prendre des exemples que dans le pays pionnier que fut

l'Angleterre. Ces discordances visibles aujourd'hui, en Amérique du Sud, par exemple, sont normales dans chaque pays en voie d'industrialisation.
• L'exemple de l'Europe prouve que l'industrialisation pose, dès les prémices de sa réussite, de graves problèmes sociaux. Le pays qui entreprend de s'industrialiser doit envisager, en même temps, la révision de ses structures sociales, s'il veut éviter la longue gestation idéologique et révolutionnaire qui a travaillé et fait souffrir l'Europe. Quatre Révolutions industrielles classiques, celle de la vapeur, celle de l'électricité, celle du moteur à explosion, celle de l'énergie nucléaire, se succèdent, et s'ajoutent les unes aux autres. »

Quelques années seulement après la parution de ce livre de Fernand Braudel, une cinquième révolution industrielle s'est déclenchée, celle des Techniques de l'information et de la communication, les TIC.
Et la question qui s'impose est de savoir si l'IA n'est qu'un prolongement de la RI des TIC ou si elle correspond à une RI à part entière, ce qui serait donc la 6ème RI.

Dressons **d'abord un panorama de l'ensemble des RI sous 4 angles, les buts, les causes, les processus et les conséquences.**

• Du point de vue d'abord des buts :
Le point commun à toutes les RI est d'accroître la puissance disponible et de diminuer l'effort que les hommes doivent déployer pour réaliser la satisfaction de leurs besoins ; autrement dit pour augmenter la productivité de leur travail et, par-là, d'augmenter la production des biens et les services susceptibles de satisfaire leurs besoins, donc la croissance économique qui est l'une conditions, et non la moindre, de la prospérité. Sachant qu'on appelle besoin en économie tout sentiment de manque, les RI ont bien pour but de répondre à une pénurie perçue de force mécanique, d'énergie, de mobilité ou encore de puissance de calcul.
C'est sur le type de manque que les RI se distinguent :

RI	But dominant	Nature du dépassement
Machine à vapeur (1780)	Remplacer la force musculaire (humaine et animale)	Physique / mécanique
Moteur électrique (1880)	Distribuer l'énergie à distance et éclairer	Physique / réseau
Moteur à explosion (1880)	Libérer la mobilité individuelle et le transport de masse	Physique / spatial
Énergie nucléaire (1945)	Produire une énergie quasi-illimitée (et une puissance militaire absolue)	Énergétique / stratégique
Electronique/Informatique (1960) puis TIC	Traiter et transmettre l'information à l'échelle planétaire	Cognitive / communicationnelle
IA	Automatiser le raisonnement et la décision	Cognitive / agentive

On observe une **bascule historique majeure** entre les quatre premières RI et les deux dernières : les premières visent à amplifier la puissance *physique et énergétique* ; les deux dernières visent à amplifier — voire supplanter — la puissance *intellectuelle*.
C'est ce que Erik Brynjolfsson & Andrew McAfee (*The Second Machine Age*, 2014) appellent le passage de l'automatisation musculaire à l'automatisation cognitive.
Une deuxième ligne de fracture sépare le nucléaire de toutes les autres : c'est la seule RI dont le but militaire a précédé et conditionné le but civil — la bombe avant le réacteur.

• Du point de vue des causes :
Dans son ouvrage important publié en 2002, « Technological Revolutions and Financial Capital », Carlotta Perez (Professeur émérite à l'University College de Londres ainsi qu'à l'Université du Sussex) identifie un schéma que l'on retrouve dans toutes les RI parce que toutes naissent de la convergence entre
1) une innovation technique de rupture,
2) une disponibilité nouvelle de ressources
et 3) une demande économique insatisfaite.

Autrement dit, tout IA concrétise une rencontre entre une demande et une offre. La question d'ailleurs, qui peut se poser, de l'offre et de la demande, quelle est celle qui vraiment le moteur ?

On peut détailler ces points de rencontre innovation de rupture / ressource-clé / demande motrice pour chaque RI :

RI	Innovation de rupture	Ressource clé	Demande motrice
Vapeur	Machine de Watt (1769-1782)	Charbon, fer	Textile, mines, transports
Électricité	Dynamo, alternateur (Faraday, Tesla, Edison)	Cuivre, charbon	Éclairage, industrie lourde
Moteur à explosion	Cycle Otto (1876), Diesel (1892)	Pétrole	Mobilité, guerre
Nucléaire	Fission (Hahn, Meitner, 1938)	Uranium	Guerre, puis énergie de base
TIC	Transistor (1947), microprocesseur (1971), TCP/IP (1983)	Silicium, cuivre/fibre	Gestion, communication, mondialisation
IA	Deep learning / AlexNet (2012) (architecture de réseau neuronal convolutif)	Data, GPU, électricité	Productivité, automatisation, compétition géopolitique

Mais si toutes les RI ont ces points de ressemblance stratégiques, il y a malgré tout des différences importantes à souligner :

1)Les quatre premières RI sont d'abord causées par des contraintes physiques (épuisement de la force animale, besoin d'énergie mobile) alors que les deux dernières le sont par des contraintes informationnelles (explosion du volume de données à traiter).

2)Le rôle de l'État est croissant : marginal pour la vapeur (initiative privée britannique), central pour le nucléaire (Manhattan Project, CEA), et à nouveau mixte pour l'IA, mais avec une dimension géopolitique inédite (rivalité USA-Chine).

3)La recherche fondamentale précède de plus en plus longtemps l'application : quelques décennies pour la vapeur, un siècle pour les réseaux de neurones artificiels (Warren McCulloch & Walter Pitts, 1943 → ChatGPT, 2022).

• Point de vue des processus :

Sur ce point aussi, l'analyse de Carlotta Perez est instructive.
Selon elle, le processus des RI est assimilable à un cycle en deux temps :
. Un temps d'installation avec 3 moments : irruption technologique, spéculation financière et infrastructure de base ;
. Un temps de déploiement avec là aussi 3 moments : diffusion généralisée, régulation, redistribution des gains.

On peut vérifier le raisonnement pour les différentes RI :

RI	Épicentre géographique	Durée de diffusion	Acteurs dominants	Rôle des standards
Vapeur	Angleterre	~80 ans	Entrepreneurs privés, banques locales	Faible (artisanal)
Électricité	USA / Allemagne	~50 ans	Grands groupes (GE, Siemens, AEG)	Fort (fréquences, voltage)
Moteur à explosion	France / USA / Allemagne	~60 ans	Constructeurs + États (guerre)	Moyen (normes routières)
Nucléaire	USA puis URSS	~30 ans (militaire), civil encore en cours	États exclusivement	Très fort (traités internationaux)
TIC	USA (Silicon Valley)	~30 ans	Mix public/privé, puis GAFAM	Décisif (TCP/IP, HTML, GSM)
IA	USA / Chine	En cours (~10 ans)	Hyperscalers privés ; architectures numériques	En construction (AI Act, etc.)

Par-délà ces points communs significatifs, il y a malgré tout des spécificités à relever :

1)La vapeur et le moteur à explosion se diffusent via des infrastructures physiques visibles (usines, routes, rails) qui structurent le territoire. L'IA se diffuse via des infrastructures invisibles (data centers, modèles dans le cloud), ce qui rend sa pénétration moins perceptible mais potentiellement plus rapide et universelle.

2)La concentration du pouvoir économique s'accentue au fil des RI : de nombreux petits entrepreneurs textiles au XVIIIe siècle, à une poignée d'hyperscalers (Microsoft/OpenAI, Google, Meta, Anthropic) pour l'IA.
Alors que les toutes premières RI se développent dans un marché atomistique, au fur et à mesure que la structure capitalistique s'alourdit avec les RI suivantes, les entreprises se développent en un environnement de plus en plus moléculaire avec une tendance marquée à la formation d'oligopoles et à la construction de barrières à l'entrée.

3)Le rythme s'emballe : ChatGPT atteint 100 millions d'utilisateurs en deux mois (2023), là où l'électricité a mis des décennies à pénétrer les foyers.

• Point de vue des conséquences :
Ce dernier point d'analyse révèle lui aussi des invariants et des différences entre les différentes RI :

Dans toutes les RI, on note :

1)Un phénomène de destruction créatrice pour reprendre l'expression de Schumpeter (1883-1950), économiste spécialiste des fluctuations économiques et de la croissance qui donne à l'entrepreneur et à l'innovation les rôles principaux : chaque RI détruit des métiers et en crée de nouveaux — le tisserand manuel, le palefrenier, le télégraphiste, le mineur de fond ont successivement disparu ou été marginalisés.

2)Carlotta Perez déjà citée place, comme d'ailleurs aussi le Prix Nobel français Philippe Aghion, ses recherches dans un cadre néo-schumpétérien qui met l'accent sur l'interaction entre technologie, écosystème entrepreneurial, innovation, politiques publiques, politique et société
.
3)Creusement initial des inégalités, suivi d'une redistribution partielle après régulation sociale et syndicale.

4)Recomposition géopolitique : chaque RI redéfinit les puissances dominantes (Angleterre → USA avec la vapeur et l'électricité ; bipolarité USA-URSS avec le nucléaire ; USA-Chine avec l'IA).

5)Nouvelles pathologies sociales : accidents industriels, pollution, dépendance aux nouvelles technologies, isolement numérique.

Dimension	Vapeur	Électricité	Moteur à explosion	Nucléaire	TIC	IA
Emploi	Exode rural, prolétariat urbain	Nouveaux métiers techniques	Industrie automobile de masse	Peu d'emplois directs	Tertiarisation, télétravail	Menace les professions qualifiées
Environnement	Début de la pollution industrielle	Moindre (énergie propre localement)	Grande pollution (CO_2, particules)	Risque nucléaire, déchets	Impact énergétique croissant	Empreinte carbone massive
Pouvoir	Bourgeoisie industrielle	Oligopoles électriques	États + constructeurs	États souverains	GAFAM, diffusion horizontale (, ateliers mécanographiques, informatique distribuée)	Reconcentration oligopolistique
Savoir	Techniques empiriques	Essor de la science appliquée	Ingénierie de masse	Big science, recherche d'État	Accès universel à l'information	Risque de pollution informationnelle
Corps social	Urbanisation brutale	Confort domestique	Individualisme mobile	Angoisse existentielle (bombe)	Société en réseau	Autonomie humaine en question

Cela dit, il y a une différence fondamentale en ce qui concerne les conséquences du développement de ces différentes RI : si les cinq premières RI transforment le rapport de l'humain à la matière et à l'énergie, l'IA est la première à aller jusqu'à transformer son rapport à elle-même. C'est pourquoi elle soulève des questions sans précédent historique : qu'est-ce que le travail intellectuel ? Qu'est-ce que la créativité ? Où s'arrête l'outil et où commence l'agent ?
Stuart Russell (*Human Compatible*, 2019) et Yoshua Bengio posent ces questions en termes d'*alignement* qui consiste à faire en sorte que l'IA produise des résultats conformes aux objectifs éthiques ou autres de ses concepteurs. En cela, l'alignement fait partie du domaine de la sûreté des IA.

Au total, il faut souligner trois lignes de fracture historiques dans ce qui vient d'être montré :

1. Local → Planétaire (accélération continue)
   La vapeur reste d'abord britannique ; l'IA est simultanément mondiale dès son émergence.
2. Diffus → Concentré (tendance croissante)
   La vapeur naît dans des milliers d'ateliers ; l'IA est entre les mains de cinq ou six acteurs mondiaux.
3. Physique → Cognitif (après les TIC)
   Les quatre premières RI amplifient la puissance corporelle et énergétique ; les deux dernières amplifient jusqu'à contester la puissance intellectuelle.

II- Même si on a compris que l'histoire des révolutions industrielles correspond à un enchaînement de bouleversements techniques ayant beaucoup de similitudes entre eux et qui non seulement se succèdent mais aussi se cumulent, *maintenant est venu le moment de répondre à la question posée au début : l'IA est-elle une RI en tant que telle ; ou serait-elle plutôt un prolongement de la RI des TIC débutée plusieurs années avant son irruption ?*

Autrement dit, *y a-t-il continuité ou discontinuité* ?

J'ai bien connu la révolution des TIC qui a commencé grâce à la convergence des progrès de l'informatique (avec son passage de l'informatique centralisée à l'informatique distribuée) et ceux des techniques de télécommunication. Quand j'ai enseigné au début des années 1970 à l'ENS que l'on appelle aujourd'hui ENS - Paris-Saclay, j'assurais dans le département d'économie et gestion, entre autres, l'enseignement de traitement de l'information, ce qui m'a amené à confronter les étudiants aux machines comptables et aux équipements à cartes perforées. Et quelques années après il m'a fallu les initier à la programmation en langage assembleur pour l'ordinateur individuel Olivetti choisi par le ministère de l'Éducation nationale pour qu'il entre dans les classes de technologie tertiaire de nos lycées. Et pour aider les collègues, j'ai écrit en 1976 deux articles publiés par le Ministère (CNDP-Centre National de Documentation Pédagogique). Puis, encore quelques années plus tard, j'ai utilisé le Cobol, ce langage dit évolué, spécialement conçu pour faire de la gestion. Et au début des années 1990, j'ai écrit un long article qui se trouve sur mon site personnel, intitulé « La nouvelle économie en questions», ce qui correspond en réalité à l'économie numérique.

•Arguments en faveur de la *continuité* : les TIC font partie de l'économie numérique et l'IA aussi.
Il y a d'autres bonnes raisons de voir l'IA — y compris les LLM (Large Language Model : grand modèle de langage) — comme une extension de la révolution numérique :

- Les LLM reposent sur des infrastructures TIC : GPU, cloud, internet, données numériques massives
- Le machine learning existe depuis les années 1980-90 — ce n'est pas conceptuellement nouveau
- Les gains de productivité restent pour l'instant sectoriels, comme dans les premières phases des TIC
- Robert Gordon ou le Prix Nobel 2024 Daron Acemoglu inviteraient à la prudence : la révolution TIC elle-même a déçu en termes de productivité agrégée.

Dans le cadre de cette réponse à notre question, l'IA serait une phase de déploiement approfondi de la même révolution — ce que Carlotta Perez appellerait le « turning point » vers la maturité.
J'en profite ici pour bien préciser que *les LLM n'ont rien à voir avec les langages machine, assembleur et évolué* évoqués plus haut : un LLM n'est pas un langage de programmation. Ce sigle désigne le langage humain naturel considéré comme objet de traitement.
La différence entre eux est triple :
1.Les langages classiques exécutent des règles explicitement écrites par un humain.
Un LLM a *appris* des régularités statistiques à partir de centaines de milliards de mots — personne n'a écrit les règles : elles ont émergé de l'entraînement (apprentissage ici, contre programmation là)
2. Un programme COBOL exécute toujours exactement la même chose dans les mêmes conditions (déterminisme).
Un LLM produit des réponses probabilistes — il génère le token le plus vraisemblable selon son modèle du langage, ce qui introduit une variabilité fondamentale (probabilisme).
3.COBOL manipule des données selon une syntaxe formelle stricte.
Un LLM manipule du sens — il peut paraphraser, inférer, raisonner, traduire, résumer. C'est précisément le saut qualitatif évoqué tout à l'heure.

• Ce qui vient d'être dit montre déjà des différences entre TIC et IA : il y a en effet des arguments en faveur de la *discontinuité*
Seulement, ce n'est pas une discontinuité seulement quantitative ; elle est également et fondamentalement qualitative. La différence n'est donc pas seulement une question de degré, c'est aussi et surtout une question de nature. **Et c'est précisément la nature de cette différence qui plaide pour dire que l'IA est une RI à part entière.**
En effet,

a) L’IA peut faire ce que les TIC ne faisaient pas

La révolution numérique des TIC automatisait des tâches routinières et codifiables (David Autor, Franck Levy, Richard Murnane, 2003). Les LLM attaquent des tâches cognitives non routinières — rédaction, raisonnement, diagnostic, code — longtemps considérées comme le refuge humain face à l'automatisation.

b) La nature de la technologie

Les TIC sont des outils de traitement et transmission de l'information. Les LLM sont des outils de production et de transformation du sens — c'est un changement de registre, pas seulement d'échelle.

c) La General Purpose Technology (GPT = TUG, technologies à usage général)

On pourrait dire que l'IA est une GPT qui améliore les autres GPT — elle accélère la recherche en biologie, en matériaux, en chimie. C'est une propriété que les TIC n'avaient pas à ce degré.

Les technologies à usage général (GPT = TUG) sont des technologies capables d'affecter l'ensemble d'une économie. Elles ont le potentiel de transformer radicalement les sociétés par leur impact sur les structures économiques et sociales préexistantes. C’est pourquoi certains expliquent le sigle GPT par « Génial mais Plutôt Terrifiant »).

d) Le travail intellectuel comme cible

Les révolutions industrielles précédentes (vapeur, électricité, TIC) substituaient du capital au travail physique ou administratif. L'IA s'attaque au travail intellectuel et créatif — ce qui est anthropologiquement d'une autre nature.

Pour résumer, on pourrait dire :

Phase	Technologie centrale	Ce qui est automatisé
Révolution industrielle classique	Vapeur, électricité	Travail physique
Révolution numérique / TIC	Informatique + Internet	Travail administratif et routinier
Révolution IA / LLM	Apprentissage profond, transformers	Travail cognitif et symbolique

Chaque phase s'appuie sur la précédente sans s'y réduire. L'IA n'est pas pensable sans les TIC, mais elle franchit un seuil qualitatif : la cognition symbolique comme objet d'automatisation.

Ceux qui font comme moi de la musique seront sensibles à l’image que je vais prendre : les langages classiques sont comme les partitions qu’interprète un musicien classique alors que les langages LLM sont comme les improvisations que fait un jazzman expérimenté qui compose sur l’instant un beau morceau à partir de tout ce qu’il a internalisé comme types d’accords et types de rythmes.

e) Le LLM est un processus de production spécifique.

Pour l’économiste, un processus de production comporte une boîte noire (l’usine) dans laquelle se réalise la transformation d’inputs (entrants) pour produire les outputs (extrants) recherchés.

Appliquons ce schéma à l'IA.

•***Côté Inputs :***

Input	Nature
Électricité	Énergie pour alimenter les serveurs et les puces
Eau	Refroidissement des datacenters
Matériel informatique	GPU/TPU, serveurs, câbles, stockage
Données d'entraînement	Des milliards de textes (livres, web, code…) utilisés *en amont* pour construire le modèle
Le prompt (tokens d'entrée)	Le texte que l'on envoie, converti en tokens numériques
Les « poids » du modèle	Les paramètres figés issus de l'entraînement — c'est la "connaissance" stockée
Capital humain	Ingénieurs, chercheurs, annotateurs humains (RLHF)

Note : Le GPU est un processeur polyvalent optimisé pour les calculs graphiques et parallèles, adapté à diverses applications IA, tandis que le TPU est un processeur spécialisé développé par Google exclusivement pour accélérer les calculs tensoriels des réseaux de neurones profonds, avec un gain en vitesse et efficacité pour ce type précis de tâches.

Remarque : nombreux sont les experts qui estiment que la bataille de l'IA se gagnera par les ressources en électricité ; plus précisément, la course à l'IA ne se gagnera pas tellement grâce à la machine la plus performante mais grâce à l'énergie la moins chère. Pour l'électricité, la France est très bien placée ; et la Chine l'est mieux que les États-Unis. L'eau a une grande importance aussi, au point que l'on réfléchit à immerger les data centers dans l'océan, à certaines profondeurs.
Remarque dans la remarque : cette utilisation de ressources naturelles, de même que les composants nécessaires pour les machines ont des conséquences négatives en termes écologiques.

•***Intérieur de la boîte noire : les facteurs de production internes***

Le cœur de la boîte noire d'une IA actuelle est son **architecture d'apprentissage avec son outil principal qu'est le « transformer »**.
Je vais tenter d'expliquer le plus simplement possible ce que fait le « « transformer », non pour empiéter sur les plates-bandes des technologues mais parce que cela a des implications économiques considérables.

1. Le problème de départ est l'importance du contexte

Imaginons qu'on lit les deux phrases suivantes :
« La banque est sur la place du marché » vs *« J'ai déposé de l'argent à ma banque »*
Le mot **«** banque **»** a un sens différent dans ces deux contextes. Les anciens systèmes de traitement du langage ne savaient pas gérer ça — ils assignaient un sens fixe à chaque mot.
Le Transformer résout précisément ce problème, grâce au « mécanisme de l'attention ».

2. Le mécanisme de l'« attention » est le cœur du transformer

L'idée intuitive : quand le modèle traite un mot, il regarde simultanément tous les autres mots de la phrase et se demande : *« Quels autres mots sont pertinents pour comprendre celui-ci ? »*
Pour « banque » dans la première phrase, il va accorder beaucoup d'attention à « place » et « marché ». Dans la seconde, à « argent » **et** « déposé »**.**
C'est comme si chaque mot votait pour les autres mots qui l'aident à se définir — et le modèle pondère ces votes pour construire une représentation du sens.
C'est fondamentalement différent d'une chaîne de montage où l'information passe mot par mot, séquentiellement — ce que faisaient les systèmes précédents (les RNN, *Recurrent Neural Networks, réseaux neuronaux récurrents*).

3. Pourquoi c'est une rupture

L'« architecture Transformer » (article fondateur de Vaswani et al., chercheurs chez Google, 2017 — intitulé sobrement *« Attention is All You Need » L'attention est tout ce dont avez besoin) permet de dépasser la technique précédente des RNN grâce à la parallélisation :

Ancienne approche (RNN)	Transformer
Traitement séquentiel mot à mot	Traitement **parallèle** de toute la séquence
La mémoire du début de phrase s'efface	**Tout le contexte** reste accessible
Lent à entraîner	Massivement *parallélisable* sur GPU
Contexte limité	Contexte très long possible

La « parallélisation » est cruciale : elle a permis de faire exploser la taille des modèles — des milliards de paramètres entraînés sur des milliards de mots — ce qui n'était tout simplement pas faisable avec les architectures précédentes.

La parallélisation : traitement en parallèle vs traitement en séquentiel)

1. Le problème avec le traitement séquentiel

Imagine que tu dois traduire cette phrase :

« Le chat mange la souris »

Avec les anciens systèmes (RNN), la machine traitait les mots un par un, dans l'ordre :

1 traite « Le » → résultat

2 prend ce résultat + « chat » → nouveau résultat

3 prend ce résultat + « mange » → etc.

Chaque étape attendait la fin de la précédente. C'est comme une chaîne de montage où chaque ouvrier attend que le précédent ait terminé.

2. Ce que fait le Transformer :

Le Transformer traite tous les mots en même temps, en parallèle et non plus en séquentiel.

Il calcule simultanément :

- la relation de « chat » avec tous les autres mots
- la relation de « mange » avec tous les autres mots
- etc.

3. Pourquoi c'est possible techniquement : les GPU

La parallélisation n'est pas une idée nouvelle : ce qui a changé, c'est le matériel disponible.

Les GPU (Graphics Processing Units), initialement conçus pour les jeux vidéo, sont des processeurs capables d'effectuer des milliers d'opérations mathématiques simultanées. Ils sont architecturalement faits pour la parallélisation.

CPU classique	*GPU*
Quelques cœurs très puissants	Des milliers de petits cœurs
Idéal pour tâches séquentielles complexes	Idéal pour tâches parallèles simples et massives
Comme un chef cuisinier expert	Comme une armée de préparateurs

Le mécanisme d'attention du Transformer se réduit à des multiplications de matrices, exactement ce pour quoi les GPU sont optimisés.

4. La conséquence économique directe

C'est ici que le lien avec l'économie devient très concret :

- La parallélisation a démultiplié la vitesse d'entraînement : on peut entraîner des modèles mille fois plus grands qu'avant
- Mais elle exige des milliers de GPU en parallèle, d'où des coûts colossaux qui constituent une barrière à l'entrée sur le marché pour dissuader tout nouveau venu éventuel d'entrer dans le club.
- Ce qui explique la concentration du marché car seules quelques entreprises peuvent financer ces infrastructures
- Et la bataille géopolitique autour des semi-conducteurs avancés (NVIDIA, TSMC, restrictions américaines à l'export vers la Chine)

Le lien de causalité est direct et vérifiable :

architecture Transformer → parallélisation → besoin de GPU → coûts massifs → concentration oligopolistique → enjeux géopolitiques sur les semi-conducteurs

Et cette chaîne illustre parfaitement ce que j'ai analysé dans mon article des années 2000 : une technologie générique restructure non seulement les marchés mais aussi les rapports de force entre nations. La révolution TIC avait produit la domination des GAFA. La révolution des Transformers est en train de reproduire, et peut-être d'amplifier, cette logique de concentration, avec les mêmes questions sur la souveraineté européenne qu'envisagent bien les rapports Draghi et Letta.

5. Alors, que fait concrètement un LLM avec la technique du transformer et son mécanisme d'attention ?

Un LLM est essentiellement des couches de Transformers empilées, des dizaines de couches d'attention successives, chacune affinant la représentation du sens.

À chaque couche, le modèle construit une compréhension de plus en plus abstraite :

- Couches basses → syntaxe, grammaire
- Couches intermédiaires → sémantique, relations entre concepts
- Couches hautes → raisonnement, cohérence globale

Et à la fin, **il** prédit le prochain token le plus vraisemblable (n'oublions pas qu'il s'agit d'un modèle probabiliste et non déterministe). Mais, répété des milliards de fois sur des données massives, ce mécanisme simple produit des capacités émergentes considérables.

Remarque : il est tentant de voir dans le mécanisme de l'attention un mécanisme de « compréhension ». C'est une métaphore utile, mais en réalité c'est fondamentalement du calcul de similarités vectorielles — de l'algèbre linéaire sophistiquée. Le sens *nous* à semble émerger à nous, mais le modèle ne manipule que des nombres. La question de savoir si cela constitue une forme de compréhension reste entière — et c'est précisément ce que John Searle contestait en 1980 avec sa *chambre chinoise*.

Comme on n'arrête jamais le progrès, on est passé récemment de la phase de l'Architecture Transformer à celle du « Transformer génératif pré-entraîné » (en anglais, le *Generative Pre-trained Transformer, donc GPT, mais attention, ce n'est pas la* General Purpose Technology)

Un GPT est une spécialisation de l'Architecture Transformer, avec trois caractéristiques supplémentaires :

Dimension	Ce que ça signifie
Transformer	L'architecture de base qui vient d'être décrite
Pré-entraîné (*Pre-trained*)	Entraîné sur des masses de texte généralistes *avant* toute spécialisation
Génératif (*Generative*)	Conçu pour *produire* du texte, pas seulement le classer ou l'analyser

Si se rajoute une 4ème dimension, celle de l'affinage (fine-tuning) sur des tâches spécifiques ou via du renforcement par retour humain, on a alors affaire à une technique de Machine Learning, le RLHF (Renforcement Learning from Human Feedback) ; le modèle génératif brut est transformé en assistant conversationnel qui apparaît si bluffant à ceux qui utilisent des chatbots.

***L'approche par GPT a des implications économiques à souligner* :**

1. Parce que le pré-entraînement a un coût colossal

Entraîner un grand modèle comme GPT-4 ou moi-même coûte des **centaines de millions de dollars** — en calcul GPU, en énergie, en données. Cela crée :

- des barrières à l'entrée massives que permet la concentration du marché sur quelques acteurs (OpenAI, Google, Anthropic, Meta). Ces barrières à l'entrée concernent non seulement le capital matériel et technique mais aussi le capital humain (question des compétences et des talents) et il y a aussi un « capital » d'actifs complémentaires. Par exemple, Google a comme produit principal Gemini mais aussi Google Search, Google WorkPlace, Google Cloud et tout un système publicitaire ; chaque brique renforçant les autres, cela rend difficile de quitter la plateforme et surtout cela dissuade ceux qui voudraient pénétrer le marché ;
- une structure industrielle oligopolistique inédite dans l'histoire du logiciel ;
- des questions sur la souveraineté technologique : l'Europe n'a pas encore de modèle « frontier » comparable, c'est-à-dire de modèle qui se trouve sur la frontière technologique, lieu où se trouvent les modèles à la pointe extrême du progrès.

2. Parce que le RLHF introduit du travail humain invisible
L'affinage par retour humain (*RLHF*) repose sur des *annotateurs humains*, souvent des travailleurs précaires dans des pays à bas salaires, qui évaluent et corrigent les réponses du modèle. C'est une chaîne de valeur cachée, documentée notamment par des enquêtes sur les sous-traitants de Meta et OpenAI au Kenya ou aux Philippines.
3. Parce que les Transformers sont la technologie qui concrétise la discontinuité
J'ai dit tout à l'heure qu'il y a une rupture qualitative des LLM par rapport aux TIC. On comprend maintenant que cette rupture repose techniquement sur le Transformer : c'est lui qui a rendu possible l'automatisation des tâches cognitives non routinières dont parlent Eloundou et ses collègues dans « GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models » (22 avril 2023).
4. Parce que l'architecture détermine les usages et donc les marchés
Le fait que les GPT soient génératifs et pré-entraînés crée un modèle économique spécifique :

- le coût marginal d'une requête est quasi nul une fois le modèle entraîné
- la valeur est captée à l'amont (entraînement) et à l'aval (API, abonnements)
- les données des utilisateurs deviennent un input stratégique pour les cycles d'amélioration suivants

C'est une économie de rendements croissants très marquée — ce qui rejoint précisément les caractéristiques des GPT (*General Purpose Technologies*) que j'analysais déjà pour les TIC dans les années 2000.
En un mot : la technologie et l'économie sont ici indissociables — ce qui est rarement aussi visible que dans le cas des LLM.
Une preuve de plus est la suivante : Yan LeCun se plait à dire que le LLM est moins intelligent qu'un chat Technologiquement il n'a pas tort mais la question n'est même pas là car en tant que technologue il est normal qu'il se pose cette question, mais ce n'est pas celle de l'économiste et des hommes d'affaires du monde de l'IA. Leur question est : « qu'est-ce qui est utile, maintenant ? ». Pour simplifier, on peut dire que le marché n'a jamais payé pour l'intelligence mais pour l'utilité. Mais qui commande l'utilité, l'offre ou la demande ?

•Côté outputs

Output	Nature
Tokens de sortie	La réponse, générée un token à la fois
Texte reconstruit	Les tokens retraduits en langage lisible
Chaleur	Dissipée par les serveurs (déchet thermique inévitable)
CO_2 (indirect)	Si l'électricité n'est pas 100 % renouvelable
Logs & données d'usage	Souvent réutilisés pour améliorer les modèles futurs
Compression du temps	Grâce à l'IA, le temps de réalisation de nombreuses tâches est réduit dans des proportions inédites.

Ce cadre montre quelque chose d'important : la phrase souvent entendue (notamment dans la bouche d'Arthur Mensch (co-fondateur et Président de Mistral AI) lors de son audition à l'Assemblée nationale), « l'IA transforme l'électricité en tokens » est vraie mais incomplète**.** Elle est tellement vraie que comme le coût marginal d'automatisation tend vers 0 à mesure que la puissance de calcul augmente, le token devient une unité de production dont le coût tend vers le coût de l'énergie. Mais l'électricité est le facteur *variable* et *visible* ; les facteurs les plus coûteux sont en réalité *en amont* — l'entraînement du modèle, qui a lui-même consumé des quantités massives d'énergie, de données et de travail humain, une fois pour toutes. L'inférence (ma question → sa réponse) n'en est que l'*exploitation*.

Note : le token, c'est quoi ?
Un token est l'unité de base avec laquelle un modèle de langage lit et produit du texte.
Ce n'est ni exactement un mot, ni une lettre : c'est un fragment de texte issu d'un découpage statistique du langage.
Concrètement :

- "bonjour" → 1 token
- "intelligence" → peut devenir 2 tokens : "intellig" + "ence"
- " !" → 1 token

- un espace fait souvent partie du token qui le suit.

En anglais, 1 token ≈ ¾ de mot. En français ou dans d'autres langues, c'est souvent moins efficace : il faut plus de tokens pour dire la même chose, car les tokenizers ont été entraînés majoritairement sur de l'anglais. La tokenization est l'algorithme qui décrit le processus de découpage du texte en unités élémentaires (tokens) que les LLMs peuvent traiter. Ce mécanisme détermine les coûts d'utilisation, les limites de contexte et la capacité du modèle à traiter différentes langues. Chaque modèle utilise son propre tokenizer.
Le modèle ne "comprend" jamais des mots : il manipule des identifiants numériques correspondant à ces fragments, puis prédit quel identifiant doit venir ensuite.

Deux remarques :
1) ce qui précède exprime trois points essentiels :

- du langage machine au langage naturel, c'est l'histoire de l'informatique
- des TIC aux LLM, c'est l'histoire des révolutions technologiques numériques
- de **l'**architecture technique **aux** implications économiques, c'est le lien indissociable entre technologie et économie

2) Les codes, qui sont le domaine pas seulement des développeurs professionnels mais aussi des communautés « open source » (qui sont une source importante des données d'entraînement), ainsi que des modèles d'IA eux-mêmes (d'où une boucle auto-récursive qui fait que l'IA produit du code qui l'améliore elle-même), sont aux trois étages de notre système de production, donc à la fois dans les inputs, la boîte noire et les outputs.
Signalons que dans une déclaration de ces derniers jours, le PDG de Google, Sundar Pichai, considère que les meilleurs ingénieurs arrêtent aujourd'hui d'écrire des codes et se transforment en chef d'orchestre pour gérer des ensembles d'AgentsIA (assistant numérique intelligent = programme conçu pour effectuer des tâches et prendre des décisions de manière autonome afin d'atteindre des objectifs spécifiques). Mais cela n'a pas empêché Elon Musk d'acheter Cursor très cher (60 Mds de $), parce que c'est le meilleur outil utilisé par les développeurs quand ils écrivent un code et que le monde se construit non pas avec du béton et de l'acier mais avec du code : c'est un code qui fait voler la fusée, c'est un code qui fait fonctionner un appareil à l'hôpital, etc… En plus, l'IA peut écrire du code et elle va pouvoir écrire son propre code et ainsi s'améliorer elle-même comme cela a déjà été dit. L'entreprise qui peut déclencher ce cycle en premier prend un avantage considérable sur ses concurrents. On comprend ainsi mieux la stratégie d'Elon Musk : en achetant Cursor, il achète l'un des plus grands réservoirs de données de codes au monde ; il a acheté plus qu'une entreprise, il a acheté le carburant du futur.
Un commentateur de la chose IA donne son point de vue : « Elon Musk était l'un des fondateurs d'OpenAI. Le mot « open » dans ce nom, c'est-à-dire « ouvert », était une promesse. L'intelligence artificielle devait être accessible à tous et pas aux mains de quelques entreprises, mais à celles de toute l'humanité. Mais cette promesse n'a pas été tenue. OpenAI s'est peu à peu fermé. Il est devenu une entreprise fermée, commerciale, contrôlée par quelques personnes. Son nom est resté « ouvert », mais ses portes se sont fermées à tous. Elon Musk y a vu une trahison et a pris ses distances. Aujourd'hui, son vrai souci se résume en une seule phrase. L'intelligence artificielle ne doit pas rester enfermée dans le coffre de quelques entreprises. Elle doit appartenir à tous. C'est ici que s'affrontent deux visions du monde. Alors, comment vaincre ces géants fermés et ouvrir l'intelligence artificielle à tous ? Cela passe par la création du modèle le plus puissant. Et le modèle le plus puissant se nourrit des meilleures données. L'une des plus grandes sources de ces données, c'est Cursor ».

Quand on décrit un processus de production de cette façon, une boîte noire qui, au moyen d'inputs, produit un output, on définit <u>la valeur ajoutée</u> ainsi réalisée comme étant à la fois la différence entre la valeur de l'Output et la valeur des Inputs d'une part et la somme des rémunérations des différents facteurs mis en œuvre pour produire cette VA d'autre part.

•Côté "différence de valeurs" (approche par les prix)

	Valeur
Output	Le token produit, vendu via un abonnement (Claude Pro, ChatGPT…) ou via l'API (prix au million de tokens)
Inputs consommés	Électricité, eau, amortissement du matériel, licences…
Valeur ajoutée brute	La différence — et elle existe, puisque ces entreprises génèrent des revenus

Remarques :

1)Mais il y a un problème de <u>mesure</u> : quelle est la valeur d'un token ? Elle est hétérogène : un token qui aide à diagnostiquer une maladie n'a pas la même valeur qu'un token qui rédige un SMS. Le prix facturé est uniforme mais la valeur créée ne l'est pas.

2)Quand un humain (expert ou annoteur) vérifie, corrige ou annote une sortie d'IA, il produit involontairement des données qui entraîneront son propre remplacement (des auteurs parlent de « malédiction du codificateur »). La VA de l'IA est donc doublement gonflée, par les données historiques non rémunérées et par le travail de vérification capté gratuitement.

3) Satya Nadella, le PDG de Microsoft vient en cette mi-juin de proposer un test pour déterminer la valeur que représente pour vous l'utilisation de votre IA : il vous propose de retirer votre système d'IA et si vous perdez toute valeur, c'est que le modèle faisait votre travail et que vous travailliez en réalité pour son fournisseur !

4) Datant également de quelques jours seulement, en lançant sa version 4, DeepSeek vient de changer la donne dans le monde de l'IA. Elle n'est pas dans le modèle et ses performances ; il ne s'agit pas de savoir si cette version « bat » GPT-5,5 ou Claude Opus 4,7. Elle est dans l'économie ; pour la 1ère fois, un modèle ouvert de niveau quasi frontalier est disponible à 1/10 du coût des modèles propriétaires leaders.
Autrement dit, l'IA se démocratise, l'intelligence devient abondante, c'est l'exécution qui reste rare ; la course ne porte plus sur les modèles, elle porte sur les systèmes. L'avantage passe de la techno à la stratégie.
Finalement, on a affaire à une évolution que l'on a déjà vue plusieurs fois : Linux a démocratisé les serveurs, Android a démocratisé les smartphones, le Cloud a démocratisé l'informatique, …

•Côté "somme des rémunérations" (approche par les revenus)

C'est là que ça devient philosophiquement délicat.

Facteur de production	Rémunération	Problème
Capital matériel (GPU, serveurs)	Amortissement, profit pour les actionnaires	✅ classique
Travail des ingénieurs	Salaires	✅ classique
Travail des annotateurs (RLHF, souvent externalisé)	Salaires très bas, souvent dans le Sud global	✅ mais invisibilisé
Les auteurs des données d'entraînement	**Rien** — ou presque	⚠️ rupture majeure
Le modèle lui-même	Rien — il ne perçoit aucune rémunération	🤔 question philosophique

À propos du travail des annoteurs, en fonction de ce qui a été dit, il y a une part qui est mesurable (le peu de salaire qui leur est versé) et une part non mesurable mais qui s'avère être un facteur important dans le processus de production : c'est non seulement un facteur de production invisibilisé mais c'est aussi le véritable facteur rare de la production à l'ère de l'IA d'aujourd'hui. On peut aller jusqu'à dire que les annoteurs sont en IA les travailleurs les moins visibles, géographiquement (il y en a un peu partout dans le monde, et souvent dans les pays en développement) et économiquement (la valeur ajoutée qu'ils apportent est captée de manière non comptabilisée), mais ils sont aussi, en même temps, les plus stratégiques.

La rupture fondamentale : la donnée comme input non rémunéré

C'est le nœud du problème. Dans tout processus de production classique, les détenteurs des inputs reçoivent une rémunération. Or ici, les auteurs des textes, livres, articles, codes qui ont servi à entraîner le modèle n'ont généralement pas été rémunérés — et ne l'ont souvent pas consenti non plus.
Cela signifie que la valeur ajoutée calculée par la méthode comptable classique est artificiellement gonflée : une partie de la "valeur ajoutée" d'Anthropic ou d'OpenAI est en réalité une captation de valeur créée par d'autres, non comptabilisée comme input.
C'est précisément ce que plaident les auteurs, journaux et artistes dans les procès actuels contre ces entreprises.

Et le modèle lui-même, facteur de production non rémunéré ?

Évidemment, le modèle *travaille* au sens fonctionnel. Mais il ne reçoit rien, n'a aucun statut juridique, ne peut rien revendiquer. C'est une **force productive sans facteur de production reconnu** — ce qui est sans précédent dans l'histoire économique, et que les cadres comptables actuels ne savent pas traiter.
En résumé : **oui, la valeur ajoutée s'applique**, mais elle révèle, en l'appliquant, une **anomalie structurelle** — des inputs massifs (les données, le travail créatif humain passé) qui entrent dans la boîte noire sans contrepartie, ce qui est à la fois un fait économique et un enjeu juridique et éthique central de notre époque.
Par conséquent, la grille économique classique, quand on l'applique sérieusement à l'IA, reste valable mais elle révèle des tensions que le droit et la comptabilité n'ont pas encore su résoudre.

D'ailleurs, il y a encore des questions délicates qui se posent en aval : qui devrait percevoir la "rente" des données ? Comment valoriser un modèle au bilan d'une entreprise ? L'IA va-t-elle modifier les parts capital/travail dans la distribution de la valeur ajoutée ?

1. Qui devrait percevoir la "rente" des données ?
Plusieurs modèles sont défendables :
Le modèle "droit d'auteur étendu" : chaque auteur dont l'œuvre a servi à l'entraînement perçoit une micro-redevance, à l'image des droits SACEM. Séduisant en théorie, cauchemardesque en pratique : identifier des milliards de sources, les attribuer, les micropayer...

Le modèle "bien commun" : les données du web ont été produites collectivement par l'humanité, donc la valeur qu'elles génèrent devrait revenir collectivement, via une taxe redistribuée. C'est la logique d'un souverain fonds de données.
Le modèle "consentement et opt-in" : seules les données explicitement licenciées entrent dans l'entraînement. Cela existe déjà (Adobe Firefly, par exemple), mais cela réduit considérablement la puissance des modèles.
Une combinaison de taxe sectorielle et de droits collectifs**,** gérée par des organismes similaires aux sociétés de gestion de droits, est probablement la voie la plus réaliste, certes imparfaite, mais praticable.

2. Comment valoriser un modèle au bilan d'une entreprise ?
C'est un vrai casse-tête comptable. Aujourd'hui, les modèles apparaissent comme des immobilisations incorporelles**,** au coût d'entraînement, un peu comme un logiciel.
Mais cette approche est insatisfaisante pour plusieurs raisons :

- Le coût d'entraînement ne reflète pas la valeur économique réelle (GPT-4 a peut-être coûté 100M$ à entraîner mais "vaut" des milliards)
- Le modèle se déprécie rapidement — un modèle de 2022 vaut peu en 2025
- Il est impossible à séparer de l'infrastructure et des données qui le font vivre.

Une valorisation par les flux futurs actualisés (c'est une bonne façon de calculer la valeur économique de tout actif) serait plus juste, mais introduit une incertitude et une subjectivité massives (comment concrètement déterminer les flux de revenus futurs successifs et comment fixer le taux d'actualisation ?). C'est un angle mort des normes comptables actuelles.

3. L'IA va-t-elle modifier les parts capital/travail dans la valeur ajoutée ?
C'est la question la plus lourde. L'histoire des grandes automatisations suggère deux scénarios :
Le scénario optimiste (thèse classique) : comme la machine à vapeur ou l'informatique, l'IA détruit des emplois routiniers mais en crée de nouveaux, augmente la productivité globale, et les gains finissent par se diffuser. La part du travail se stabilise à long terme.
Le scénario pessimiste (thèse de Daron Acemoglu, entre autres) : l'IA est différente des révolutions précédentes car elle attaque cette fois les tâches cognitives non routinières, là où le travail humain s'était jusqu'ici réfugié. Si le capital peut être substitué au travail jusque dans ses bastions, la part du travail dans la valeur ajoutée — déjà en baisse depuis les années 1980 — pourrait continuer à s'effondrer.
Le scénario pessimiste semble sous-estimé dans le débat public. Non pas que la croissance soit impossible, mais parce que la vitesse de la transition pourrait dépasser la capacité des travailleurs et des institutions à s'adapter — ce qui n'est pas un problème de long terme, mais un problème de la prochaine décennie.
Le fil conducteur de ces trois réponses est le même : les cadres existants — juridiques, comptables, économiques — ont été construits pour un monde où les facteurs de production avaient des propriétaires identifiables et des prix de marché**.** L'IA les met en défaut non pas parce qu'elle est magique, mais parce qu'elle repose sur un input, la connaissance humaine accumulée, que personne n'a jamais su vraiment tarifer.
Ce sont des questions que les économistes, les juristes et les décideurs publics vont devoir affronter sérieusement dans les années qui viennent.

Deux dernières remarques, au risque de vous mettre un peu de confusion dans l'esprit :

1)Les fameux LLM tournent sur des machines qui ne peuvent fonctionner qu'en étant programmées avec leur propre langage machine, cette programmation étant écrite en langage évolué (Python ou C++ par exemple). Et en sortie les LLM peuvent produire du code Cobol ou assembleur !
Finalement, pour reboucler un peu avec la thèse de la continuité, on peut dire qu'il y a une progression -plus ou moins régulière – dans l'ordre de l'abstraction en passant du langage machine puis à l'assembleur puis au langage évolué puis enfin au LLM, l'abstraction en ce sens que l'on a progressé de plus en plus en partant du langage de la machine pour se rapprocher le plus possible du langage humain. Avec le LLM, le langage naturel devient lui-même l'interface. *Alors qu'avant il fallait que l'homme apprenne à parler le langage de la machine, avec le LLM, c'est la machine qui apprend à parler comme nous*.

2)Pour rendre possible tout progrès technique, quel qu'il soit, il faut se fonder sur des **infrastructures** pertinentes. Celles-ci doivent pour cela répondre à trois conditions : *l'efficacité* dans le rapport de leurs résultats aux objectifs

poursuivis, *l'efficience* dans le rapport de leurs résultats aux ressources utilisées et *l'effectivité* dans le rapport de leurs résultats aux désirs de la société.
À l'ère industrielle, les nations qui ont bâti des infrastructures physiques telles que routes, voies ferrées, réseaux d'eau et d'électricité ont acquis des avantages concurrentiels durables.
Aujourd'hui, une transition comparable, et peut-être plus lourde de conséquences, est en cours grâce à l'infrastructure intelligente.
Pour me limiter à l'essentiel, je poserai seulement trois questions et en tirerai une conséquence, donc 4 points rapides :

1. Les infrastructures numériques présentent-elles des spécificités ?
Au sens classique, une infrastructure est un équipement collectif qui conditionne l'activité économique sans être directement consommé dans la production — les routes, les réseaux électriques, les ports.
Elle a trois caractéristiques :
•Non-rivalité partielle — plusieurs utilisateurs peuvent s'en servir simultanément
•Effets de réseau — sa valeur augmente avec le nombre d'utilisateurs (comme « plus on est de fous, plus on rit », « plus on est nombreux sur un réseau d'IA et plus tous ses utilisateurs en profitent » ; à condition toutefois qu'il n'y ait pas congestion du réseau).
•Irréversibilité — elle nécessite des investissements massifs et durables
Les infrastructures numériques suivent cette même logique, mais appliquée à la circulation de données et de calcul.
On peut les décomposer en trois couches :

Couche	*Exemples*
Physique	Câbles sous-marins, datacenters, antennes 5G, satellites
Logicielle	Systèmes d'exploitation, protocoles internet (TCP/IP), cloud (AWS, Azure…)
Cognitive	Modèles de langage, moteurs de recherche, systèmes de recommandation

C'est cette troisième couche qui est nouvelle et qui mérite notre attention parce que c'est la couche qui explique le qualificatif « intelligent » et qui fait de l'infrastructure un input supplémentaire et capital pour la boîte noire : l'infrastructure en devient en effet à la fois le squelette, le système nerveux et le système circulatoire.

2. On parle d'infrastructures "intelligentes". Qu'est-ce-à dire ?
L'adjectif "intelligent" est une prétention de type marketing, mais il pointe malgré tout quelque chose de réel : ces infrastructures ne se contentent plus de transporter ou de stocker, elles traitent, filtrent, priorisent et génèrent.
Un réseau routier est passif — il ne décide pas quelle voiture passe en premier.
Un réseau numérique intelligent, lui :
- Anticipe la demande (mise en cache prédictive)
- Optimise en temps réel (routage dynamique, gestion de l'énergie)
- Produit de la valeur propre (recommandations, traductions, diagnostics)

C'est ce saut — de l'infrastructure *passive* à l'infrastructure *active* — qui est conceptuellement nouveau.
En février 2026, plusieurs auteurs conduits par Christian Catalini, du MIT, ont signé un article marquant : « Some Simple Economics of AGI » où ils considèrent que l'exécution n'est pas la seule infrastructure critique ; la *vérification* l'est aussi, sans doute même la plus importante, ce qui déplace le centre de gravité de l'analyse, de la production vers la vérification. Ils le montrent en avançant le concept de « gap de mesurabilité » qui est l'écart entre ce que les IA peuvent exécuter (étendue des tâches automatisables) et ce que les humains peuvent vérifier (étendue des tâches contrôlables). Or, ce gap s'élargit structurellement puisque l'étendue des tâches automatisables croît exponentiellement avec les capacités de calcul alors que l'étendue des tâches contrôlables est bornée biologiquement. Il y a donc collision de deux courbes de coûts concurrentes, celle du coût marginal de production qui tend vers 0 avec l'augmentation de la capacité de calcul et l'accumulation des données, alors que celle du coût de vérification est dépendante des contraintes biologiques, du temps humain et de l'expérience vécue. De plus, si les effets des réseaux ont joué très favorablement dès le développement de l'IA pour leur activité de production, ce qui résiste aujourd'hui, ce sont les effets de réseau de vérification. On retrouve ici un schéma constant : la valeur ne réside jamais dans la technologie elle-même, qui finit toujours par se diffuser, mais dans ce que l'organisation sait en faire et que personne ne peut répliquer.

Enfin, remarquons que ce qui n'est pas vérifié ne peut pas être correctement valorisé et a fortiori rémunéré. Et ce problème de valorisation et de rémunération ne peut que s'aggraver : la capacité humaine à vérifier va devenir dans l'économie numérique de demain le facteur le plus rare et, hélas, le plus mal rémunéré.
En juin, donc ce mois-ci, est paru un autre article, produit par des chercheurs de Google DeepMind, « From AGI to ASI », ce qui correspond à un saut dans l'avenir, donc dans l'inconnu. Sauf que pour eux, en allant de plus en plus vers toujours plus de capacité de calcul, il ne s'agit pas d'augmenter un input économique, il s'agit carrément d'un vecteur de franchissement de seuils qualitatifs d'intelligence en direction de la limite théorique « incomputable ». C'est pourquoi, pour eux, le gap de mesurabilité introduit par Christian Catalini deviendrait incommensurable et la supervision humaine directe deviendrait totalement impossible. Alors, ils ont eu bien raison de mettre en exergue au début de leur article une phrase du livre écrit en 1950 par Turing (Computing Machinery and Intelligence) : « nous ne pouvons voir qu'à courte distance, mais nous voyons beaucoup de choses à faire ».

3. Comment intégrer l'infrastructure dans le processus de production que j'ai décrit en inputs - boîte noire – et outputs ?
On peut les situer les infrastructures numériques à trois niveaux :
Au niveau des inputs
Les câbles sous-marins, les datacenters, le cloud sont ce qui permet aux données d'entraînement d'exister, de circuler et d'être stockées. Sans eux, pas de modèle possible. Ils sont une sorte de méta-infrastructure, invisible mais fondatrice.
À l'intérieur de la boîte noire
L'inférence (le passage de la question à la réponse) s'exécute sur des infrastructures de calcul distribuées, des clusters (grappes) de GPU (processeurs graphiques pour faire des calculs à grande vitesse) répartis sur plusieurs datacenters, coordonnés en temps réel. L'infrastructure n'est pas séparable du processus de production : elle est *le facteur capital* au sens le plus concret.
Au niveau des outputs
La réponse que donne l'IA à la question qu'on lui pose parvient via des infrastructures réseau (fibre, 4G/5G, protocoles). Sans cette couche, l'output n'atteint jamais l'utilisateur — la valeur reste dans la boîte.

4. Une conséquence économique majeure : le capital devient infrastructure
C'est peut-être la transformation la plus importante. Historiquement, le capital de production (une usine, une machine) appartenait à une entreprise particulière. Les infrastructures, elles, étaient souvent publiques ou réglementées, car leur caractère de monopole naturel rendait la concurrence inefficace.
Or les infrastructures numériques intelligentes sont aujourd'hui privées et concentrées — AWS, Google Cloud, Microsoft Azure contrôlent l'essentiel du cloud mondial. Anthropic elle-même loue ses infrastructures à Amazon. Cela crée une structure inédite : une poignée d'entreprises contrôlent l'infrastructure sur laquelle repose la production de valeur de toute l'économie numérique, y compris leurs propres concurrents.
C'est une position de rente infrastructurelle d'une puissance sans précédent, qui pose des questions de régulation analogues à celles que posaient les chemins de fer au XIXe siècle, et que l'Europe commence seulement à saisir avec le Digital Markets Act, le fameux DMA entré en vigueur en mars 2024 qui veut mettre fin à la domination des géants du Net.

En résumé
Les infrastructures numériques intelligentes sont tout à la fois le squelette, le système nerveux et le système circulatoire du processus de l'IA. Les ignorer dans l'analyse économique de l'IA, c'est comme analyser la production industrielle classique en oubliant l'électricité et les routes : on décrit la machine, mais pas les conditions qui la rendent possible.
Et leur caractère "intelligent" — actif, prédictif, génératif — brouille la frontière classique qui existait entre infrastructure (neutre, passive) et acteur économique (qui crée et capte de la valeur).

Remarque en guise de conclusion de cette partie : en IA, il faut connaître la valse des étiquettes !

Valse des étiquettes parce que l'on parle de l'IA mais en réalité il y a plusieurs IA, plusieurs sortes d'IA, plusieurs niveaux d'IA.
Il est donc indispensable d'en dresser une classification. Le critère le plus utilisé est celui de la capacité générale :
IA étroite (ANI — Artificial Narrow Intelligence) Beaucoup d'IA actuelles appartiennent à cette catégorie.
Elles excellent dans une tâche unique ou un domaine circonscrit : reconnaissance d'images, traduction, jeu d'échecs, recommandation. Elles n'ont aucune capacité de transfert spontané vers un autre domaine.
IA générale (AGI — Artificial General Intelligence) Un système hypothétique capable d'accomplir n'importe quelle tâche cognitive qu'un humain peut accomplir, avec la même flexibilité et adaptabilité. Le débat sur le moment d'émergence de l'AGI est très actif — certains chercheurs pensent qu'elle est proche, d'autres qu'elle est fondamentalement différente de ce que nous construisons.
Jusqu'à ce niveau, faible autonomie et contrôlabilité satisfaisante
Mais existent maintenant de plus en plus des IA « *agentiques »* qui sont non seulement en mode réactif mais proactif, capables

- d'échanges longs et multiétapes
- de percevoir leur environnement,
- de planifier une séquence d'actions pour atteindre un objectif (elles dépassent le traitement de tâches isolées mais restent la plupart du temps domaine-dépendantes).
- d'agir de manière autonome dans le monde réel (exécuter un code, naviguer sur Internet, interagir avec des outils, passer des commandes, …) : les IA agentiques ont une autonomie d'action significative puisqu'elles ne se contentent pas de répondre, elles peuvent agir (elles font partie des IA actives, par opposition aux IA passives).
- de s'adapter en fonction des résultats intermédiaires

Les IA les plus connues sont ChatGPT, Claude ou encore des agents IA industriels.
L'IA agentique est actuellement le concept empiriquement le plus ancré dans la réalité, c'est pourquoi l'EU AI Act commence à beaucoup s'y intéresser, car ces IA agentiques sont considérées à haut risque (mais des risques opérationnels – non existentiels - comme des erreurs en cascade, des actions irréversibles, des détournements d'objectifs.
À partir du niveau suivant et a fortiori pour le dernier, autonomie de plus en plus grande et contrôlabilité de plus en plus faible
Niveau intermédiaire de transition : d'IA dite RSI (pour "recursive self-improvement"),
IA superintelligente (ASI — Artificial Super Intelligence) Concept encore plus spéculatif : une IA qui surpasserait l'intelligence humaine dans tous les domaines, y compris la créativité, le jugement social et la recherche scientifique. C'est le cœur des réflexions de Nick Bostrom ou Stuart Russell sur les risques existentiels.
Ces deux niveaux (RSI et ASI) présentent des risques existentiels et/ou des transformations civilisationnelles. C'est au sujet de ces deux niveaux que se pose en priorité la grave question de la sécurité de l'IA. Mais, pour l'instant des deux niveaux restent encore hautement spéculatifs, des extrapolations théoriques. Ce ne sont pas à l'heure actuelle des faits imminents mais des hypothèses de travail.

Soulignons que le passage de la RSI et l'ASI soulève maintes questions redoutables (à la fois au point de vue technologique mais aussi et surtout aux points de vue philosophie et éthique) puisqu'il déplace le regard de l'économie présente de l'IA vers sa trajectoire possible. Un article passionnant est paru ce mois-ci sur le sujet, « From AGI to ASI » , écrit par des universitaires des Universités de Londres, d'Australie et par des chercheurs de Google DeepMind. Ils se donnent, entre autres, pour but de placer l'ASI sur le continuum des différents types d'IA d'après le score que Shane Legg et Marcus Hutter ont développé dans leur article d'avril 2006 (il y a tout juste 20 ans !), « A Formal Measure of Machine Intelligence », score qui va de l'AGI jusqu'à l'IA Universelle (dite « AIXI »), qui correspond à la limite théorique « incomputable ». Les auteurs de l'article « From AGI to ASI » se posent plus particulièrement la question : les modèles entraînés sur des données humaines peuvent-ils un jour former des concepts véritablement nouveaux auxquels l'humain n'aurait pas encore pensé ? Leur réponse est incertaine mais leur hypothèse est troublante à partir d'un concept qu'ils introduisent, celui de « Abstraction

Barrier » : nos modèles actuels seraient prisonniers du plafond conceptuel de l'humanité. Ils recombinent ce que nous savons, mais ne peuvent sans doute pas penser ce que nous n'avons pas encore pensé. Pour franchir cette barrière, il faudrait une IA capable d'abstraire des concepts nouveaux directement depuis des données brutes du monde physique, ce qu'aucun système actuel ne fait.
Cela prolonge et nuance le propos tenu sur la valeur ajoutée : si l'IA est fondamentalement limitée par le stock de connaissance humaine dont elle est issue, alors sa valeur productive est bornée par ce stock ; ce qui rejoint notre analyse des données d'entraînement comme input non rémunéré, mais lui ajoute une dimension *cognitive* et pas seulement économique.

2ème partie : Quels scénarios pour l'avenir proche ?

I Deux questions vives au préalable

Avant de répondre à cette question de prospective, envisageons rapidement quelques autres questions que nous n'avons pas encore évoquées ou trop peu.

1ère question : Les infrastructures intelligentes ne sont pas seulement un défi conceptuel. C'est aussi un défi géopolitique pour l'Europe et donc pour nous tous (surtout qu'il y a des enjeux de cybersécurité) ; et le retard de l'Europe n'est même pas essentiellement technique mais carrément culturel.

Je viens de lire une réflexion faite par un ingénieur du secteur : « l'Europe veut une IA souveraine. Elle commence par décourager ceux qui pourraient la construire. Le discours européen sur l'intelligence artificielle devient totalement contradictoire. D'un côté, on répète qu'il faut bâtir une IA souveraine, créer des champions européens, protéger nos données, développer des clouds de confiance, financer du calcul, rattraper les États-Unis et la Chine. De l'autre, on empile les obstacles : AI Act, RGPD, doctrine CNIL, droit d'auteur, Data Act, NIS2, DORA, Cyber Resilience Act, SecNumCloud, hébergement des données de santé, CSRD, droit du travail, règles sectorielles dans la santé, la finance, l'éducation ou les services publics. Chaque texte peut se défendre isolément. Ensemble, ils créent une machine à ralentir. Le problème n'est pas qu'il y ait des règles. L'IA pose de vraies questions de sécurité, de responsabilité, de données et de droits fondamentaux. Mais l'Europe ne se contente pas d'encadrer. Elle transforme chaque projet IA en dossier juridique avant même qu'il ne devienne un produit. Avec l'AI Act, une startup doit classifier les risques, documenter son système, organiser la supervision humaine, prouver la qualité de ses données, prévoir le suivi post-déploiement. Avec le RGPD et la CNIL, elle doit justifier chaque traitement, chaque finalité, chaque base légale, chaque durée de conservation, chaque droit d'opposition. Avec le droit d'auteur, elle avance dans l'incertitude sur les données d'entraînement. Avec NIS2, DORA et le Cyber Resilience Act, elle ajoute audits, procédures cyber, contrôles fournisseurs, reporting incidents. Avec SecNumCloud ou HDS, elle voit son accès au cloud et aux données sensibles se complexifier encore.

Résultat : avant de recruter des ingénieurs, d'acheter du compute, de tester un produit ou de trouver ses clients, il faut déjà financer des juristes, des audits, des registres, des politiques internes et des procédures. Ce n'est pas neutre. C'est un choix industriel. Et c'est un mauvais choix si l'objectif est réellement la souveraineté. Car l'IA est une course de vitesse (le Co-fondateur et Président de Mistral AI, Arthur Mensch, a dit récemment devant la Commission parlementaire que si on ne rattrape pas notre retard dans les deux ans, ce retard sera totalement irrattrapable). L'IA exige du capital, des talents, des données, de l'énergie, du cloud, des GPU, de la commande publique, des clients et la capacité d'expérimenter rapidement. Or l'Europe ajoute du délai, du coût fixe et de l'incertitude précisément là où il faudrait de la vitesse. On veut des champions, mais on impose aux entrants des charges que seuls les grands groupes peuvent absorber. On veut de l'innovation, mais on commence par présumer le risque. On veut de la souveraineté, mais on freine l'accès aux données, au cloud, à l'énergie, au marché et au capital. C'est particulièrement absurde dans la santé, la finance ou l'éducation, qui sont justement les secteurs où l'IA pourrait produire des gains massifs. Chaque cas d'usage y devient un parcours réglementaire, un comité de conformité, un arbitrage juridique, une analyse de risque. L'Europe parle comme une puissance technologique, mais agit comme une administration de contrôle. La souveraineté ne se décrète pas. Elle se produit. Elle se produit avec des ingénieurs, des usines de calcul, une énergie abondante, des règles claires, des marchés accessibles et la liberté d'essayer vite. À force de vouloir être le continent de l'IA de confiance, l'Europe risque surtout de devenir le continent où personne ne peut construire assez vite. Nous aurons les règles, les audits, les labels, les chartes, les doctrines et les formulaires. Et les modèles seront ailleurs. Ce n'est pas une stratégie industrielle. C'est une bureaucratie qui rêve de puissance ».

À propos de Mistral, pour un autre spécialiste, on est en face d'un terrible cas de conscience : est-il encore efficace d'encourager et d'aider Mistral AI quand on entend plusieurs spécialistes du domaine dire que Pour l'un, « pour construire un modèle « SOTA » (c'est-à-dire étant au sommum de l'état de l'art au moment considéré), ce que l'on appelle aussi le modèle frontière, autrement dit encore le modèle de pointe, il faut accepter d'être un pirate comme tous le sont parce que c'est la condition d'entrée dans la course, et c'est la guerre, c'est risqué, c'est sanglant, ça demande de mettre une force que 99% des gens sont incapables d'imaginer. Tu engages tout, tu

prends des risques qui pourraient te tuer, et tu fonces quand même. Cette énergie-là, l'Europe la stérilise méthodiquement. Pourquoi ? Parce que chez nous, personne n'est libre. Les gens ne sont pas libres, et ils sont pauvres. Même les entrepreneurs sont pauvres, parce qu'on ponctionne, on encadre, on bride à chaque étape ».

2ème question : quelle est la structure de l'industrie de l'IA ?

Dans toute industrie, les entreprises tentent de s'approcher le plus possible de la situation particulièrement enviable du monopole. C'est ce qui explique que dans l'industrie de l'IA, comme dans beaucoup d'autres, c'est la structure oligopolistique qui domine. Je rappelle qu'un oligopole est une structure de marché où il n'y a que quelques offreurs, dont la taille peut être variable et dont le fonctionnement peut éventuellement se coupler avec les stratégies qu'adoptent les entreprises dans une autre structure de marché imparfait, celle de la concurrence monopolistique, à savoir des stratégies de différenciation des produits. Dans un oligopole, une autre espèce de stratégie est nettement plus importante, elle concerne les barrières à l'entrée que les oligopoleurs érigent pour protéger leur place sur le marché de l'intrusion de tout concurrent supplémentaire éventuel.

L'industrie de l'IA comporte essentiellement trois étages, et chacun est caractérisé par un fort degré de concentration et doté de barrières à l'entrée spécifiques. D'où au total une industrie très concentrée avec quelques entreprises ultra dominantes au point que certains, comme la journaliste spécialiste de ces questions, Nastasia Hadjadgi, n'hésitent pas à dire que « l'IA est un outil pensé par les capitalistes pour les capitalistes ».

Ces 3 niveaux sont à la base les infrastructures, puis il y a celui des LLM et le niveau du dessus est celui des applications.

Alors qu'au niveau des applications, qui est le plus visible avec les chatbots, les outils de productivité, les assistants spécialisés et les logiciels professionnels enrichis par l'IA mais aussi avec des barrières à l'entrée relativement faibles puisque toute entreprise peut lancer un produit en s'appuyant sur un LLM existant en en payant seulement l'utilisation, bien que cette ouverture soit toute relative puisque ce niveau supérieur de la pyramide est conditionné par les deux niveaux du dessous : au niveau des intrastructures, on a les puces électroniques spécialisées, les centres de données qui les hébergent et les services cloud qui permettent d'y accéder à distance. C'est couche la plus intensive en capital physique et la plus concentrée. Sans accès à cette infrastructure, il est tout simplement impossible d'entraîner ou de faire tourner un LLM. C'est ici que réside la première barrière, la plus physique et la plus concrète. Le niveau intermédiaire est celui des modèles de langage ; on a les grands systèmes d'intelligence artificielle comme ChatGPT chez OpenAl, Gemini chez Google, Claude chez Anthropic, Lama chez Meta ou Mistral Al en France. Ces deux niveaux inférieurs, le niveau des infrastructures et celui des LLM, ont tous deux une très redoutable barrière à l'entrée : une barrière financière, puisqu'il faut être en mesure de supporter des coûts extrêmement élevés dans la mesure où cette industrie se caractérise fondamentalement par une forte intensité capitalistique, non seulement capital physique mais aussi capital humain (importance au niveau des LLM des talents des personnes recrutées au niveau mondial).

Mais, si le marché de l'IA est à coup sûr très imparfait, puisque fortement oligopolistique, il n'en est pas pour autant complètement figé parce s'y exercent certaines forces de contestabilité, pour faire référence à la théorie des marchés contestables de William Baumol, John Panzar et Robert Willig (1982).

La première source de contestabilité est celle de gains possibles d'efficience algorithmique : en effet, alors que les acteurs présents sur le marché se font concurrence le plus souvent par l'amélioration des performances de leurs produits, avec plus de données, plus de paramètres, plus de puissance de calcul, il est possible, comme l'a fait DeepSeek, d'entrer sur le marché en réalisant un produit moins cher mais aussi performant que les autres.

Une seconde source de contestabilité est celle des modèles « open weight ». Précisons qu'il ne faut pas confondre les modèles « open weight » et les modèles « open source » : les modèles « open weights » sont les modèles dont les paramètres sont rendus publics, donc téléchargeables, ce qui permet en particulier un déploiement local, sans nécessairement dévoiler les données d'entraînement ou le code source. On en voit donc les résultats mais sans savoir comme ils sont produits. Par contre les modèles « open source » sont entièrement transparents : les codes, les poids, les données d'entraînement sont accessibles au public ; ce qui permet de les améliorer ou même créer de nouveaux modèles.

Longtemps, l'IA générative a été dominée par les modèles « closed weights », c'est-à-dire des modèles propriétaires accessibles uniquement via API (Application Programming Interface », interface logicielle qui permet de connecter un logiciel à un autre) mais, dès que les modèles « open weigts » ont rattrapé le retard technologique qu'ils avaient, ils sont devenus la règle, y compris chez les géants comme OpenAI et Anthropic. D'où le fait que la performance n'est plus le critère différenciant, surtout que, comme pour beaucoup de matériels, notamment en Hi-Fi, pour 90% des usages, les écarts de qualité sont devenus difficilement perceptibles. Quatre

autres critères jouent maintenant un rôle décisif : le prix, la souveraineté des données, la spécialisation-métier et la capacité d'intégration. Les armes de la guerre qu'utilisent les oligopoleurs les uns contre les autres ont donc changé et certains sont mieux placés que d'autres : Google et Alibaba sont peut-être mieux placés qu'OpenAI et Anthropic parce que l'IA n'est pas leur seule source de revenus.

II Les scénarii possibles pour le proche avenir

Si j'en crois les différentes études sur le sujet des conséquences économiques et sociales de l'IA, et tout spécialement la dernière en date, écrite par un groupe d'économistes américains de plusieurs universités (Virginie, Pennsylvanie, Stanford, …) et parue en mars dernier, la conclusion est la même : « la principale source de désaccord parmi les économistes ne porte probablement pas sur la question de savoir si les capacités de l'IA progresseront de manière significative — la majorité attribue une probabilité non négligeable au scénario modéré ou rapide — mais sur la rapidité avec laquelle l'économie peut absorber ces capacités potentiellement transformatrices, et sur ce que cette absorption produira en termes d'impacts économiques. Plus précisément, les économistes qui partagent des vues similaires sur la probabilité d'un progrès rapide de l'IA divergent néanmoins sur le taux de diffusion probable, la mesure dans laquelle la création de nouveaux emplois compensera le déplacement, le degré auquel des décalages surviendront entre l'adoption de l'IA et les gains de productivité [c'est ce que l'on appelle en économie le paradoxe de Solow ; 1987)], et la manière dont les réponses institutionnelles et réglementaires façonneront la transition. (…) Même dans le scénario rapide, où les systèmes d'IA surpassent les performances humaines dans la plupart des tâches cognitives et physiques d'ici 2030, les experts ne prévoient pas de résultats économiques en dehors de la plage de l'expérience historique. Au contraire, leurs justifications mentionnent à plusieurs reprises les retards de diffusion, les goulets d'étranglement infrastructurels, l'instabilité politique et les vents contraires démographiques comme mécanismes susceptibles d'empêcher même une IA très capable de produire des résultats économiques sans précédent à court terme ».
Ce dissensus se retrouve dans le très récent débat entre Yann Le Cun (ancien directeur scientifique de l'IA de Meta, Prix Turing en 2018) et Stéphane Mallard (conférencier international et prospectiviste) : pour Le Cun, l'IA ne va pas remplacer les humains, au contraire, elle va ouvrir de nouvelles possibilités et créer des métiers ; pour Mallard au contraire, l'IA « commoditise » l'expertise et va entraîner un net déplacement au sein des facteurs de production, du travail vers le capital, tout spécialement dans les domaines d'expertise.

pour tout ce qui concerne les prévisions sur les évolutions macroéconomiques, je fais assez confiance à l'avis des économistes qui travaillent dans les agences de notation. il se trouve que ceux de l'une d'elles - Moody's, Mark Zandi, Cristian Deritis, Marisa Dinatale, Dante Deantonio, Matt Colyar, Shandor Whitcher, Justin Begley, Ilir hHsa and Gwen Semmens, 25 février 2026 - ont fait connaître leurs scénarii, en les probabilisant, pour la période qui vient.

À l'occasion de la présentation de leurs scénarii, les auteurs font des remarques utiles sur les différences d'approche qu'il y a entre les économistes et les technologues car, selon l'angle de vue adopté, *les scénarii peuvent être très différents, en particulier entre les économistes et les technologues* :
« *Les économistes* s'accordent généralement à dire que l'intelligence artificielle stimulera considérablement la croissance de la productivité dans les années, voire les décennies à venir. Cependant, ces gains escomptés restent modestes comparés à ceux anticipés par *les technologues* qui développent l'IA et dirigent les entreprises spécialisées.
Sans vouloir généraliser, ces deux camps se parlent sans s'entendre pour diverses raisons.
Premièrement, économistes et technologues utilisent des définitions différentes de la productivité. Les économistes la mesurent en termes de production réelle par heure travaillée, tandis que les technologues la mesurent en termes de débit de tâches, comme le nombre de lignes de code, de tickets clients, de documents juridiques ou d'examens radiologiques.
Deuxièmement, les économistes observent des délais d'adoption des nouvelles technologies que les technologues sous-estiment systématiquement. La courbe en J de la productivité est un concept central pour les économistes. Elle suggère que, pour qu'une technologie à usage général comme l'IA soit adoptée, les entreprises doivent investir massivement dans du capital immatériel complémentaire – notamment en réorganisant leurs flux de travail, en formant à nouveau leurs employés, en repensant leurs processus et en renégociant leurs contrats – avant de constater des gains de productivité significatifs. Ces investissements peuvent être importants, non

comptabilisés dans le PIB et longs à rentabiliser. Les technologues, quant à eux, évoluent au sein d'organisations généralement capables de s'adapter rapidement.

Troisièmement, économistes et technologues ont des conceptions fondamentalement différentes quant au rythme d'amélioration des capacités de l'IA. Se basant sur l'histoire des nouvelles technologies, la plupart des économistes estiment que les capacités technologiques progressent par étapes et subissent des rendements décroissants. On observe des améliorations impressionnantes entre GPT-3, GPT-4 et GPT-5, par exemple, mais les gains marginaux ralentissent sur divers indicateurs, tandis que les coûts de formation augmentent. À l'inverse, les technologues perçoivent des capacités émergentes et des axes de recherche qui, selon eux, permettront des progrès extrêmement rapides dans le domaine de l'IA.
Quatrièmement, le point de vue des économistes s'appuie sur l'observation de Robert Solow en 1987 : « On constate l'avènement de l'ère informatique partout, sauf dans les statistiques de productivité. » Autrement dit, les gains issus des nouvelles technologies seront moindres et plus lents que ne le laisse entendre le battage médiatique. Les technologues, quant à eux, affirment qu'il ne s'agit pas d'un simple effet de mode et que le décalage perçu entre l'impact économique de la technologie et sa représentation dans les données n'est pas un véritable décalage, mais plutôt une conséquence naturelle du temps nécessaire aux mesures pour rattraper la réalité.
Enfin, cinquièmement, économistes et technologues défendent sans doute chacun leurs propres intérêts. Les économistes ont intérêt à la prudence, car ils mettent leur crédibilité en péril s'ils adhèrent à la prédiction des technologues selon laquelle « cette fois-ci, c'est différent » et que l'IA révolutionnera l'économie. Mais si l'IA s'avère être un facteur de changement économique imprévu par les économistes, ils s'exposeront à moins de sanctions professionnelles. À l'inverse, les technologues ont tout intérêt à être audacieux, car ils doivent convaincre les investisseurs de fournir des capitaux importants pour développer l'infrastructure de l'IA. Par conséquent, leur succès repose sur l'idée que l'IA sera transformatrice ».

Voici les 4 scénarii :

1- **Avec une probabilité de 40%, l'économie serait dopée par l'IA ; mais pas tellement l'emploi.**

Ce scénario se fonde sur ce qui s'est passé l'an dernier : en 2025, l'IA a ajouté environ un demi-point de pourcentage à la croissance du PIB réel.
Environ un tiers de cette contribution provient de l'essor des centres de données et de la construction d'infrastructures électriques, ainsi que d'autres investissements dans les équipements et les logiciels.
Les deux tiers restants proviennent de l'augmentation des dépenses de consommation, stimulée par les effets positifs sur le patrimoine induits par la forte hausse de la valeur des actions des entreprises d'IA.
Dans ce scénario, la situation a des chances d'évoluer rapidement, la croissance de la productivité s'accélérant progressivement à mesure que l'adoption de l'IA par les entreprises se poursuit à un rythme comparable à celui d'Internet ; si bien que les gains de productivité anticipés entraînent certes un ralentissement de la croissance de l'emploi et une hausse du chômage structurel à long terme mais ils soutiennent la hausse des revenus des ménages et des bénéfices des entreprises. Ils soutiennent également la hausse des cours boursiers, de l'immobilier et des autres actifs, et contribuent à atténuer les difficultés budgétaires de l'État. Ce scénario offre donc des perspectives macroéconomiques optimistes. Notons que celles des « technologues » le sont encore davantage.

2- **Avec une probabilité de 25%, il y aurait une crise de l'IA ; peut-être même un krack.**

Ce scénario, pessimiste au contraire du scénario de référence, est motivé par la flambée extraordinaire des cours boursiers des entreprises d'IA. Ces cours ont grimpé et grimpent toujours si haut et si vite que le marché boursier semble surévalué, voire spéculatif, et potentiellement en situation de bulle. Dans ce scénario, les investisseurs ont gravement sous-estimé les taux d'adoption de l'IA et leur impact sur les processus métier, et par conséquent, les gains futurs de productivité et de rentabilité des entreprises. Le marché boursier peut subir une forte correction à court terme, dont les répercussions frappent durement le système financier et l'économie, précipitant une récession.
Les ingrédients d'une bulle boursière sont sans doute réunis.
Depuis la première publication de l'indice Dow Jones Industrial Average — composé des 30 plus grandes entreprises cotées en bourse — à la fin du XIXe siècle, on constate que seules trois autres décennies ont connu une telle hausse des prix.

Les Années folles, bien sûr, se sont terminées brutalement avec le krach boursier de 1929, qui a engendré la Grande Dépression des années 1930. Il s'agissait clairement d'une bulle spéculative.
Dans les années 1950, la croissance boursière a été alimentée par la domination des entreprises américaines sur l'économie mondiale au lendemain de la Seconde Guerre mondiale. Parmi ces entreprises figuraient General Electric, AT&T, General Motors, U.S. Steel et DuPont. Ce n'était pas une bulle spéculative.
Puis il y a eu l'engouement pour Internet dans les années 1990, qui s'est achevé peu après le passage à l'an 2000 par une chute spectaculaire des cours boursiers. Il ne fait aucun doute qu'il s'agissait d'une bulle spéculative. Internet était une technologie révolutionnaire qui a permis d'énormes gains de productivité et, en fin de compte, de générer des profits considérables.
Cependant, les investisseurs ont sous-estimé tous ces éléments, et bien d'autres encore. La valorisation boursière – le prix des actions par rapport aux bénéfices des entreprises – a explosé.
Un indicateur fiable de valorisation est le ratio entre la valeur de toutes les actions cotées en bourse, mesurée par l'indice Wilshire 5000, et les bénéfices des entreprises à l'échelle de l'économie. Au cours des 75 années sur lesquelles cet indicateur peut être calculé, le cours des actions a représenté *en moyenne* 12 fois les bénéfices des entreprises. Actuellement, les cours sont plus de 20 fois supérieurs aux bénéfices. La seule autre période où les valorisations ont été plus élevées remonte au plus fort de la bulle Internet de l'an 2000, lorsque ce ratio a brièvement atteint 24 fois.
Si la tendance actuelle du marché se maintient, la bourse formera bientôt une bulle. Le dernier ingrédient nécessaire à la formation d'une bulle est la capitulation de la quasi-totalité des sceptiques. À force de dénoncer cette bulle, ils ne sont plus crédibles. Tout scepticisme est balayé d'un revers de main et la bulle grossit encore.
La flambée des marchés boursiers est un puissant moteur pour l'économie. Les plus riches, qui détiennent la majeure partie des actions, sont désormais bien plus fortunés et dépensent en conséquence. Cette richesse nouvellement acquise soutient des dépenses importantes, qui, à leur tour, soutiennent un nombre substantiel d'emplois. Ceci met progressivement mais sûrement en lumière une menace considérable pour l'économie : si la bourse est effectivement une bulle et qu'elle éclate, la consommation subira un choc brutal, déclenchant une récession et chômage. C'est précisément ce qui s'est produit après l'éclatement de la bulle du bug de l'an 2000, et c'est le fondement de ce scénario.

3- Avec une probabilité de 20%, gonflement du chômage

Ce scénario est encore plus pessimiste.
L'enchaînement est le suivant : les investisseurs ayant fait grimper le cours des actions des entreprises d'IA ont vu juste, les entreprises adoptent rapidement l'IA, en constante évolution, dans leurs processus. Les gains de productivité qui en résultent sont si importants et si rapides que l'économie subit des pertes d'emplois considérables et un chômage beaucoup plus élevé.
Les grandes technologies d'amélioration de la productivité *du passé* n'ont pas entraîné ce type de pertes d'emplois. Leur adoption a été suffisamment lente pour permettre aux entreprises et à l'économie de s'adapter.
On disait comme Alfred Sauvy et d'autres que la machine ne crée pas de chômage, qu'il y a un « déversement » de l'emploi de certains secteurs vers d'autres, et comme Schumpeter que le progrès passe nécessairement par une « destruction créatrice ».
Mais la situation est différente aujourd'hui. En effet, le marché du travail est déjà en difficulté avant même que les gains de productivité liés à l'IA ne se concrétisent. La croissance de l'emploi est au point mort, car une politique d'immigration restrictive a réduit l'offre de main-d'œuvre, et la guerre commerciale mondiale ainsi que l'incertitude économique générale pèsent sur la demande.
Bien que certains signes indiquent que l'IA stimule la productivité et affaiblit la demande de main-d'œuvre, ils restent encore timides. Cependant, la situation peut évoluer rapidement dans le contexte de bouleversements du marché du travail, à mesure que l'adoption de l'IA s'accélère et que les gains de productivité qu'elle génère se concrétisent. Le taux historiquement élevé de créations d'entreprises depuis la pandémie favorise cette adoption. Ces créations sont importantes dans la plupart des secteurs et sur l'ensemble du territoire. Nombre de ces nouvelles entreprises devront pleinement intégrer l'IA pour rester compétitives.
Contrairement aux périodes précédentes de mutations technologiques rapides, les bénéfices économiques de l'IA ne sont pas équitablement répartis dans ce contexte. À ce jour, cela semble se confirmer, un petit nombre de géants du cloud et leurs actionnaires ayant été les grands gagnants économiques. L'écart salarial entre les travailleurs hautement qualifiés et les travailleurs moyennement qualifiés se creuse à mesure que l'adoption de l'IA s'accélère. L'IA complète les compétences des travailleurs hautement qualifiés, créatifs et techniques, leur permettant d'obtenir des salaires plus élevés. La baisse de la demande et des salaires pour les emplois

moyennement qualifiés, tels que les postes administratifs et de bureau, contribue à la raréfaction du marché du travail. Les travailleurs peu qualifiés ne peuvent échapper aux conséquences de cette situation, car ils doivent désormais faire face à la concurrence de travailleurs moyennement qualifiés dont les compétences ont été réduites.
Le marché du travail se contracte donc et une récession s'ensuit. De plus, l'économie souffre d'hystérésis, de rémanence, car la longue période de chômage élevé et le décalage persistant entre les compétences des travailleurs et celles nécessaires à une utilisation efficace de l'IA entraînent un chômage structurel accru. Ce scénario est pessimiste, peut-être même trop, car il suppose une absence de réaction énergique des Banques centrales (spécialement de la Réserve fédérale) et des responsables de la politique budgétaire.
Remarque importante : la question de l'effet de l'IA sur l'emploi n'a pas seulement une dimension quantitative mais aussi qualitative.
Début mars dernier, Anthropic a publié une étude sur les effets de son IA sur les emplois : Les travailleurs les plus menacés ne sont pas ceux auxquels on s'attendait. Ils sont plus âgés, plus diplômés et gagnent 47 % de plus que la moyenne. De plus, ils sont près de quatre fois plus susceptibles d'être titulaires d'un diplôme d'études supérieures que les travailleurs non concernés par l'IA.
L'argument est simple. Anthropic a créé un nouvel indicateur appelé « exposition observée ». Il ne s'agit pas de ce que l'IA pourrait théoriquement faire, mais de ce qu'elle fait réellement actuellement en milieu professionnel, mesuré à partir de millions de conversations Claude réelles avec des utilisateurs en entreprise.
Pour les informaticiens et les mathématiciens, l'IA est théoriquement capable de prendre en charge 94 % de leurs tâches. Elle n'en prend actuellement en charge que 33 %. Pour les postes administratifs, la capacité théorique est de 90 %, tandis que l'utilisation observée actuelle est de 40 %. L'écart entre le potentiel de l'IA et son utilisation actuelle est considérable. Les chercheurs sont clairs sur ce qui va suivre : à mesure que ses capacités s'améliorent et que son adoption se généralise, la zone rouge s'étend et finit par recouvrir la zone bleue.
Ce constat démographique est ce qui rend l'étude troublante. Les travailleurs les plus exposés à l'IA gagnent en moyenne 47 % de plus que le groupe le moins exposé. Ils sont plus souvent des femmes et plus susceptibles d'avoir fait des études supérieures. Il ne s'agit pas ici des magasiniers ou des chauffeurs routiers, mais des avocats, des analystes financiers, des chargés d'études de marché et des développeurs de logiciels. Soit précisément le groupe dont la formation était censée les protéger.
Les programmeurs informatiques présentent le taux d'exposition à l'IA le plus élevé (74,5 %), suivis des conseillers clientèle (70,1 %), des opérateurs de saisie (67,1 %), des spécialistes des dossiers médicaux (66,7 %) et des analystes d'études de marché et spécialistes marketing (64,8 %). Ce ne sont pas des prédictions, mais des mesures du travail déjà effectué sur les plateformes d'IA.
Et puis, il y a cette problématique de la chaîne de valeur, dont on ne parle pas assez.
Les chercheurs d'Anthropic ont constaté une baisse de 14 % du taux d'insertion professionnelle des 22-25 ans occupant des emplois fortement exposés à l'IA depuis le lancement de ChatGPT. Aucun effet comparable n'a été observé chez les plus de 25 ans. Les postes de débutant n'ont jamais été de simples emplois. Ils constituaient le tremplin où les analystes juniors devenaient analystes seniors, où les jeunes avocats apprenaient à construire des arguments solides. Si ce processus disparaît, personne ne sait d'où viendra la prochaine génération de cadres supérieurs.

4- Avec une probabilité de 15%, Le marché du travail reste stable malgré les gains de productivité que génère l'IA

C'est, après tout, le cas du boom internet de la fin des années 1990. Entre 1996 et 2000, tandis que la productivité des entreprises non agricoles progressait de 3 % par an, l'emploi salarié augmentait de près de 2,5 millions de postes par an et le taux de chômage baissait régulièrement, passant de 5,5 % à un peu moins de 4 %. À la fin de cette période de forte croissance de la productivité, l'économie avait retrouvé le plein emploi. Bien sûr, des différences importantes existent entre aujourd'hui et cette époque, ce qui rend beaucoup moins probable que la productivité et la croissance de l'emploi soient aussi bonnes lorsque la vague d'IA déferlera sur l'économie. Le marché du travail actuel est fondamentalement différent. Il succède à une période de chômage historiquement bas, et même si le marché du travail s'est affaibli plus récemment, le chômage reste faible. La demande de main-d'œuvre est également faible, les entreprises étant peu enclines à embaucher dans un contexte d'incertitude politique accrue. De plus, l'offre de main-d'œuvre disponible demeure tendue en raison du renforcement des restrictions à l'immigration et du départ à la retraite de la génération du baby-boom. La vague de l'IA semble plus susceptible de remplacer et de supprimer des emplois existants que ne l'a jamais fait le boom d'Internet.

CONCLUSION :

L'industrie de l'IA a la structure d'un oligopole et, comme toujours dans une telle situation, les entreprises n'ont que trois possibilités pour se rapprocher encore davantage de la situation de monopole : la guerre tous azimuts (constitution d'un trust), l'entente totale (le cartel) ou un mix des deux, guerre dans certains domaines et entente tacite dans les autres. La guerre de l'IA a déjà commencé, surtout au niveau des empires. L'IA n'est plus vraiment gratuite et il faut peut-être s'attendre à une guerre de prix. La guerre que se livrent OpenAI et Anthropic va prendre une autre dimension à la Bourse quand ils vont entrer sur le marché (la fusion entre ChatGPT et Codex que Sam Altman (OpenAI) vient d'annoncer s'inscrit dans cette bataille).
Pour mieux comprendre les enjeux et les modalités de cette guerre, il faut se féliciter de la création au début de ce mois de juin, par le Gouvernement britannique, de l'Institut d'économie de l'IA (AIEI), le premier du genre dans le monde, financé par le gouvernement, et qui va être présidé par le professeur Simon Johnson, prix Nobel d'économie en 2024 et chef économiste du FMI.
Il faut aussi se féliciter de l'existence précieuse, dans le cadre de l'université d'Oxford, donc également au Royaume-Uni, de l'Institut d'éthique de l'IA créé en 2021. Dans la définition de ses missions, on lit notamment : « Chaque jour apporte son lot de défis éthiques posés par l'IA, de la reconnaissance faciale au profilage des électeurs, des interfaces cerveau-machine aux drones armés, en passant par l'avenir de l'emploi. Il y a 40 ans, les philosophes ont joué un rôle majeur dans l'élaboration de l'éthique médicale. L'Institut ambitionne aujourd'hui d'apporter une contribution similaire à l'élaboration de l'éthique de l'IA.
Notre mission est de mener des recherches interdisciplinaires de pointe – à la croisée de la philosophie, du droit, des sciences sociales et de l'informatique – sur l'éthique de l'IA, notamment sur les questions de sécurité et de gouvernance (…) ». Le hasard du calendrier veut que l'Observatoire mondial de l'éthique et de la gouvernance de l'IA de l'UNESCO ait fait la même année, en 2021, une recommandation en la matière, applicable pour les 194 membres de l'Organisation (les États-Unis l'ont quittée fin 2018).
Je laisserai le dernier mot à Daron Acemoglu, du MIT et Prix Nobel d'économie, qui a écrit hier : « Bien entendu, utilisée de la bonne manière, l'IA est un outil qui peut être bénéfique dans de nombreux domaines. La question est de savoir si nous pouvons développer des institutions, des normes et des pratiques pour soutenir son utilisation bénéfique et si la direction actuelle de la technologie dans la Silicon Valley nous permettra de le faire ».
Cette dernière phrase prend tout son sens quand on le livre qu'Alexander Karp, co-fondateur et PDG de Palantir a publié en février, « The Technological Republic », et quand on lit le commentaire qu'en fait Anders Rasmussen, un ancien secrétaire général de l'OTAN : « « L'appel de Karp à une "République technologique" expose clairement les conditions nécessaires au maintien de la prééminence du monde démocratique à l'ère de l'intelligence artificielle. Ingénieurs et technologues doivent mettre leurs talents au service de nos libertés démocratiques, afin que l'avenir numérique les renforce et ne les compromette pas. Ce livre est un signal d'alarme pour les entrepreneurs du secteur technologique, dans la Silicon Valley et au-delà. »