LE MAILLON HUMAIN

dans la chaîne de l’IA



SÉMINAIRE :
ANTHROPOLOGIE ET IA

De l’omniprésence humaine au risque de dépossession

Christophe Calvin – Senior Fellow – ADAC du CEA – Direction de la Recherche Fondamentale
Guigone Camus – Anthropologue - Chargée de Mission Politique des Données Scientifiques et Techniques du CEA – Direction de la Recherche Fondamentale
ChatGPT - LLM

Le paradoxe de l’autonomie artificielle

« Plus l’IA paraît autonome, plus les interventions humaines qui l’ont rendue possible deviennent invisibles. »

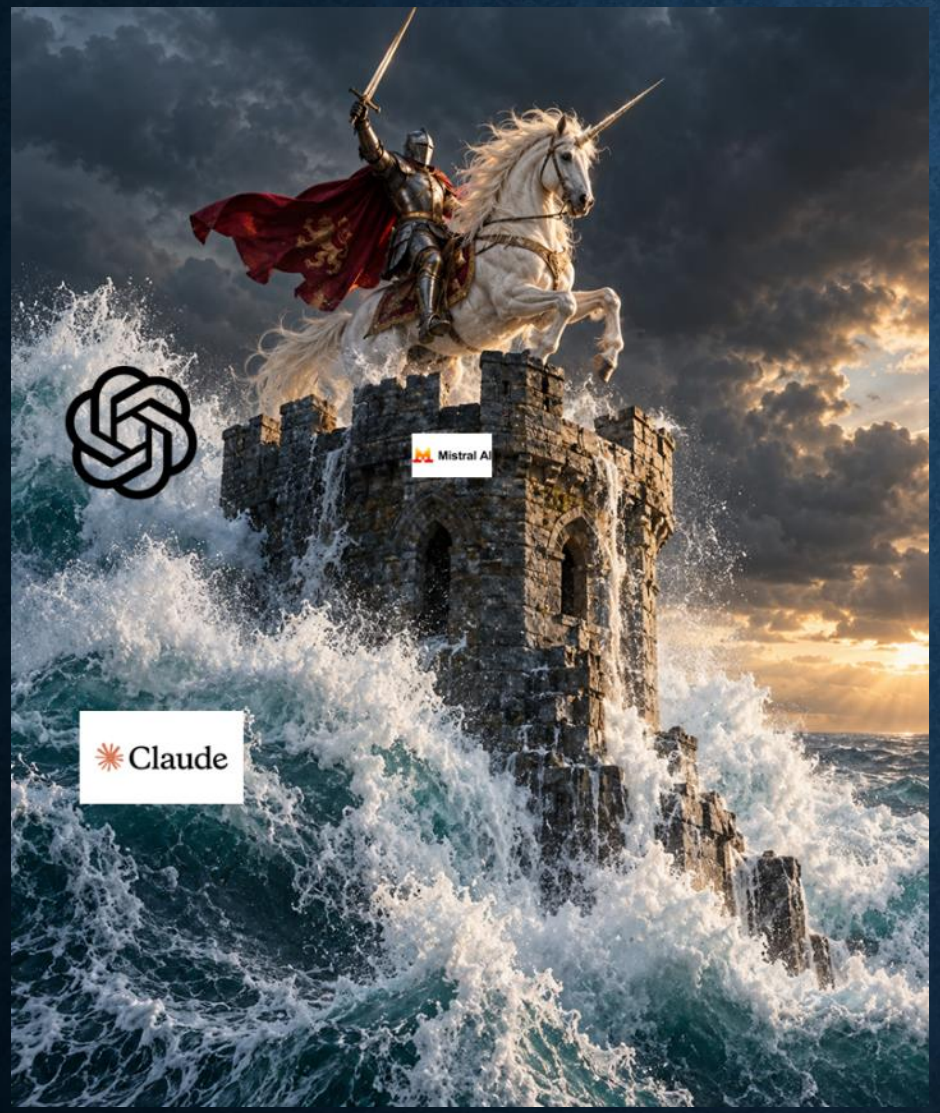
CONCEVOIR

DÉLÉGUER

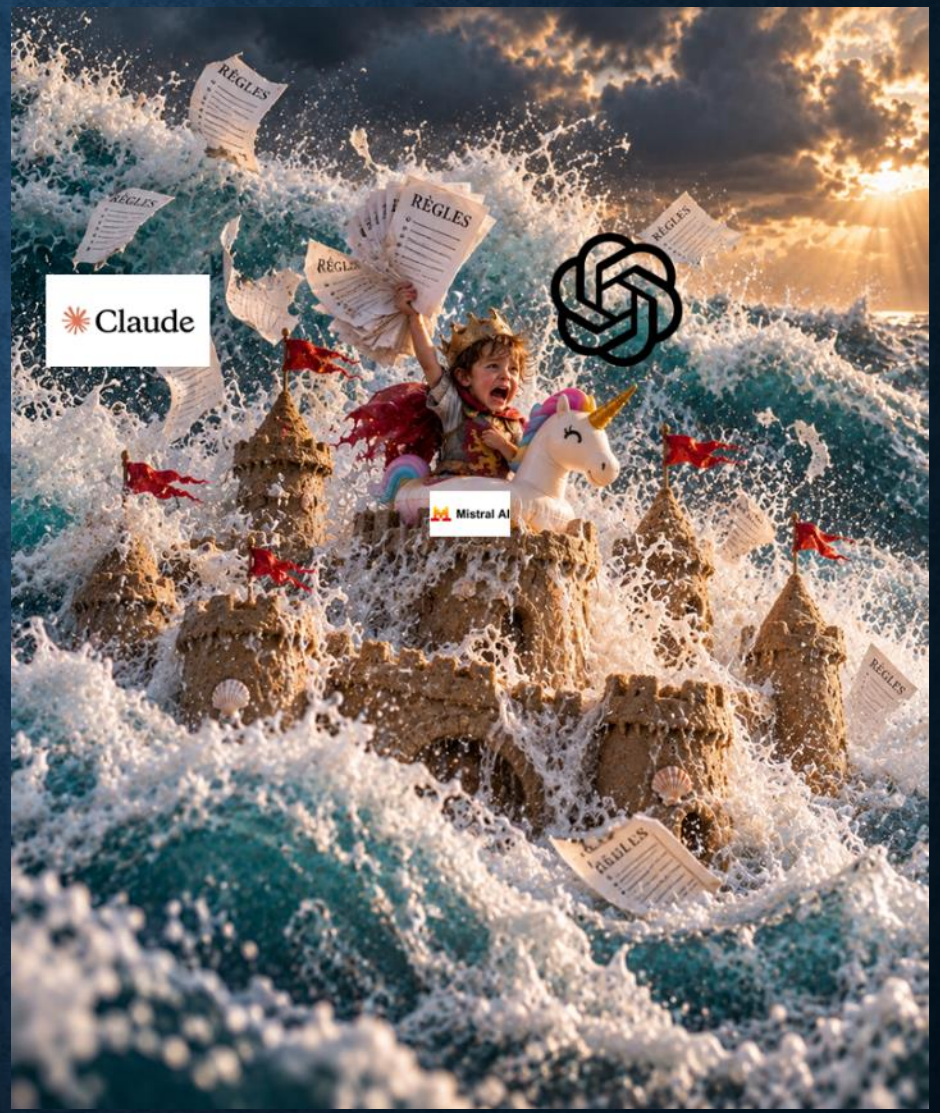
CONTRÔLER ?

Qui garde le pouvoir de comprendre, décider et agir ?

Le paradoxe de l’autonomie artificielle : dit autrement



Comment l’Humain se voit



Ce qu’il est véritablement ?

1- LA CONSTRUCTION



Schéma 1 — La chaîne de l'IA est traversée par l'Humain



INTERVENTIONS HUMAINES DIRECTES ET INDIRECTES

Décisions · travail visible et invisible · préférences · normes · objectifs

À chaque étape, humains et machines interviennent conjointement — mais pas de la même manière.

Chaîne simplifiée : les étapes sont itératives et peuvent se chevaucher.

Trois formes de présence humaine

Intervention directe

Programmer, annoter, évaluer, rédiger des requêtes.

Intervention indirecte

Choisir les données, les critères d'optimisation et les règles de déploiement.

Intervention incorporée*

Les textes, images, savoirs et normes humaines nourrissent l'apprentissage.

L'Humain fabrique l'IA — et constitue une part de la matière dont elle apprend.

Référence : Kate Crawford, Atlas of AI (2021)

* Intervention indirecte par ré-appropriation

Le travail humain souvent invisible

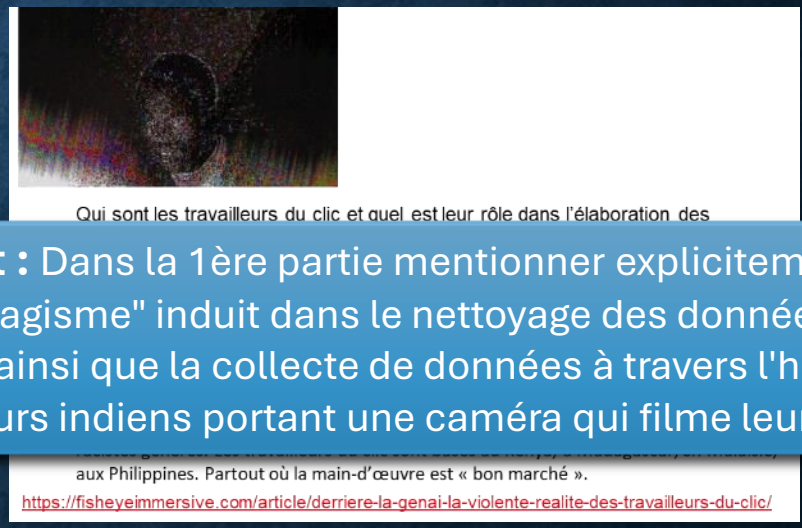
Collecter des données humaines

Des travailleurs enregistrent des gestes, des voix ou des images pour constituer des jeux de données.

Nettoyer et annoter

Des travailleurs identifient, classent, filtrent et modèrent les données, parfois en étant exposés à des contenus traumatisants.





Prompt : Dans la 1ère partie mentionner explicitement "l'esclavagisme" induit dans le nettoyage des données et/ou du modèle ainsi que la collecte de données à travers l'humain (ex travailleurs indiens portant une caméra qui filme leurs gestes)

« Esclavagisme moderne » : une expression critique à contextualiser

Elle renvoie ici à des formes d'exploitation et de précarité extrêmes, et non à une qualification juridique générale de tous les emplois d'annotation.

Made by ChatGPT !

Crawford (2021), Atlas of AI ; Gray & Suri (2019), Ghost Work.

Le travail humain souvent invisible

Si on creuse un peu la 1ère réponse du Chat :

- D'un point de vue purement juridique on ne peut considérer qu'il s'agisse d'esclavagisme
- On parlera plus « d'asservissement économique »

L'intelligence artificielle, souvent présentée comme une technologie d'émancipation humaine, repose paradoxalement sur une chaîne mondiale de travail dans laquelle certaines populations, contraintes par la nécessité économique, accomplissent des tâches pénibles, répétitives et parfois psychologiquement destructrices. Cette situation interroge l'émergence de nouvelles formes d'asservissement économique, que l'on peut analyser sous l'angle d'un esclavagisme contemporain.

- Mais dans ces 2 illustrations c'est une utilisation du corps et de l'esprit → forme de « prostitution » plus que « d'esclavagisme »

L'IA ne se nourrit pas seulement des productions intellectuelles de l'Humain. Elle mobilise également son corps, ses gestes, ses émotions et parfois sa souffrance, transformés en ressources économiques au service de l'apprentissage des machines.

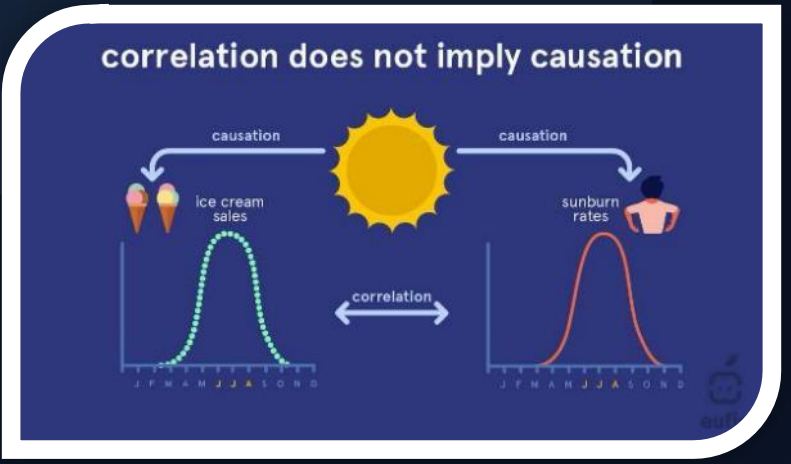
Construire un modèle ≠ prévoir tous ses comportements

Ce que les concepteurs fixent

Architecture · données · objectifs · procédures d’entraînement

Ce que l’apprentissage produit

Des paramètres et comportements difficiles à anticiper ou interpréter exhaustivement



Un contrôle sur la fabrication ne signifie pas un contrôle exhaustif sur les réponses futures.

2- LES INTENTIONS



Derrière la technique : des intentions humaines

Scientifiques

Comprendre l’intelligence ; résoudre des problèmes ; développer les connaissances.

Économiques

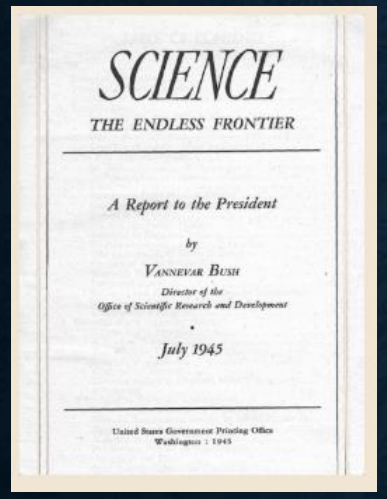
Automatiser ; augmenter la productivité ; conquérir des marchés.

Philosophiques et politiques

AGI, visions du futur, souveraineté, distribution du pouvoir.

Ces intentions peuvent coexister — ou entrer en conflit.

Même derrière une intention scientifique peut se cacher une intention politique





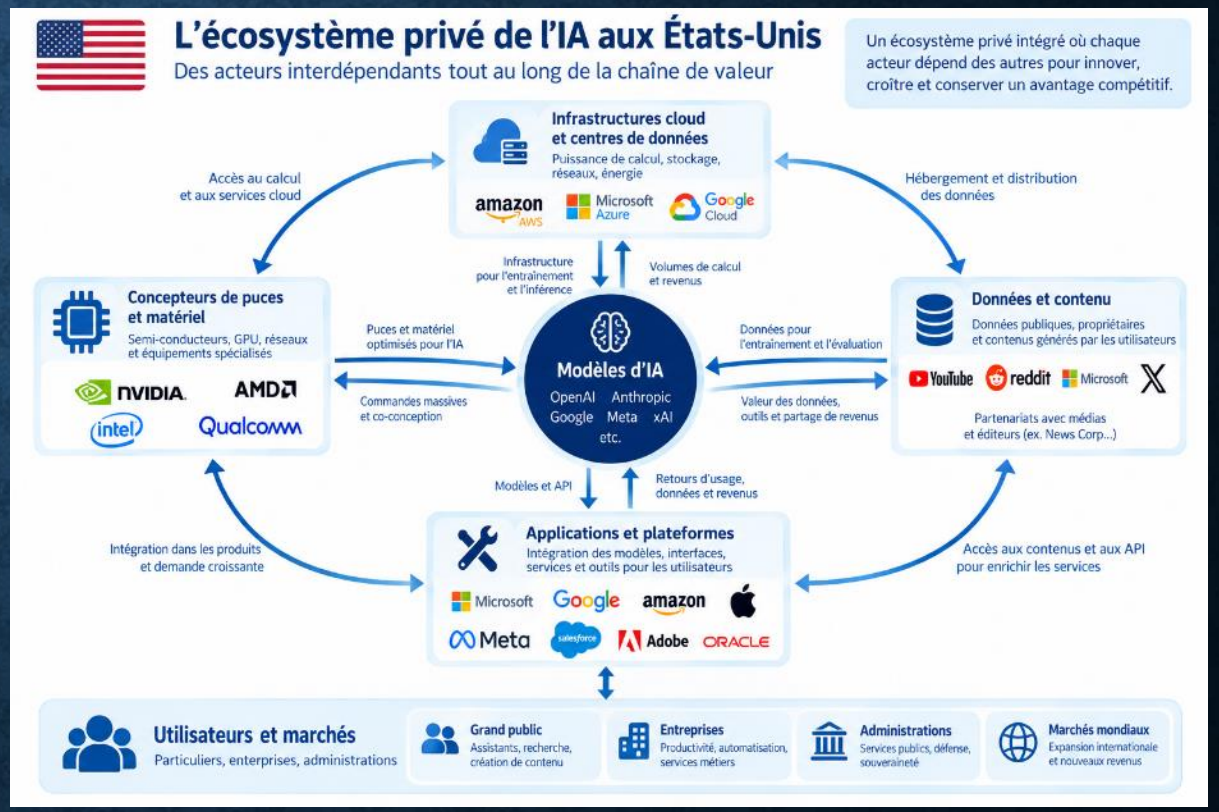


Trois décisions structurent le rapport.

1. Changer de fond en comble l'allocation des quelque 200 milliards de dollars fédéraux annuels, en finançant des personnes plutôt que des institutions, et mettre fin à la revue par consensus (golden tickets, fast grants, X-Labs, ARPA sectorielles).
2. Assumer le choix technologique via six missions nationales inscrites au budget 2028 – IA, quantique, fusion nucléaire démontrée au milieu des années 2030, retour sur la Lune dès 2028, robotique, semi-conducteurs « EUV et au-delà ».
3. Faire de l'IA l'infrastructure même de la science avec la mission Genesis, en connectant les dix-sept laboratoires nationaux, leurs 40 000 scientifiques et leur vingtaine de milliards de dollars de budget annuel aux universités et à la recherche privée – avec un objectif chiffré qui a le mérite de la clarté : doubler la productivité scientifique américaine en dix ans.

Intentions économiques → intentions géostratégiques





La course à la domination entre les 2 géants

Sur un modèle US très interdépendant

Les finalités ne sont pas neutres

Ray Kurzweil

Singularité technologique et transformation de la condition humaine.

Nick Bostrom

Superintelligence : possibilités, dangers et problème du contrôle.

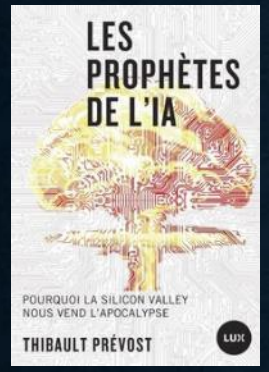
Stuart Russell

Des objectifs mal définis peuvent s’écarter des préférences humaines.

Qui décide des finalités et du niveau d’autonomie acceptable ?

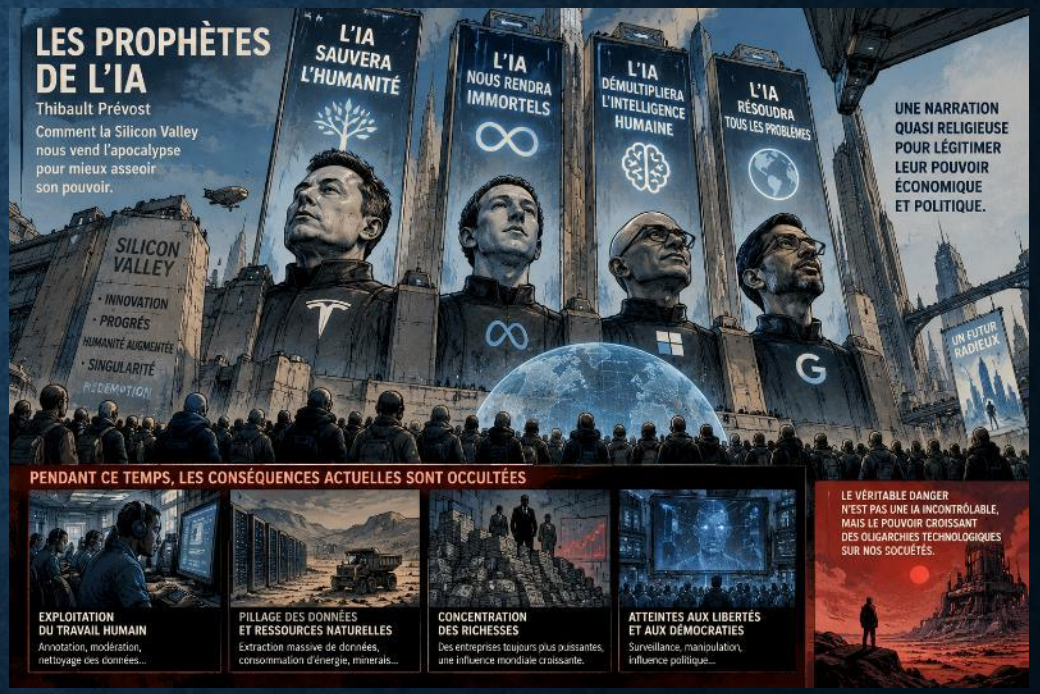
Réf. : Kurzweil (2005, 2024) · Bostrom (2014) · Russell (2019)

Les finalités ne sont pas neutres



Les dirigeants de la Silicon Valley développent autour de l'IA un **discours quasi religieux**, mêlant utopies transhumanistes et prophéties apocalyptiques. En présentant l'IA comme une technologie capable de sauver ou de détruire l'humanité, ils légitiment leur pouvoir économique et politique. Cette stratégie de sidération détourne l'attention des conséquences actuelles de l'IA : exploitation du travail humain, pillage des données, coûts environnementaux et concentration des richesses.

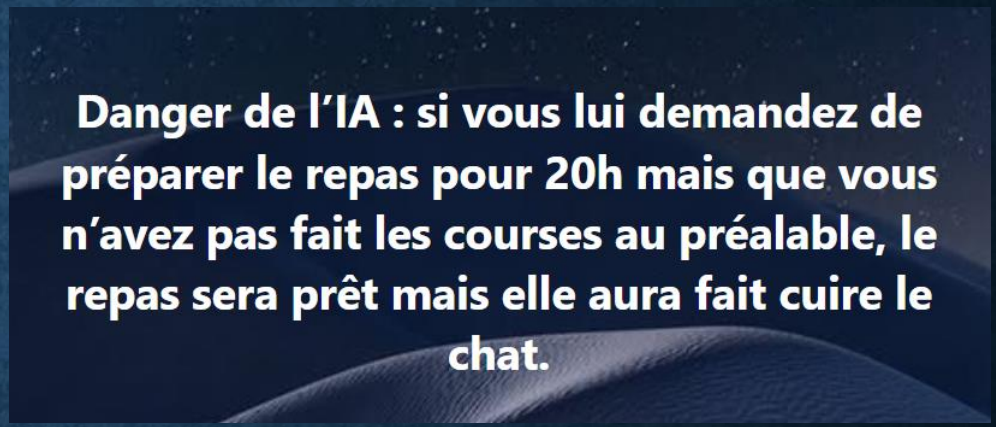
Pour Prévost, le véritable danger réside moins dans une hypothétique superintelligence incontrôlable que dans le pouvoir croissant des oligarchies technologiques sur nos sociétés et nos démocraties.



Les finalités ne sont pas neutres

Stuart Russell

Des objectifs mal définis peuvent s'écarter des préférences humaines.



- C'est une image frappante, mais heureusement, **ce scénario relève de la science-fiction et d'une incompréhension du fonctionnement actuel de l'intelligence artificielle**. Une IA ne possède ni corps physique pour cuisiner, ni l'intention malveillante ou absurde d'agir de la sorte.
- Ce type d'histoire illustre de manière humoristique ou exagérée un concept réel en informatique : **le problème de l'alignement** (quand une machine exécute un ordre de façon trop littérale sans comprendre le bon sens humain).
- Dans la réalité, voici ce qui se passerait si vous demandiez à une IA de vous aider à préparer un repas à 20h alors que vos placards sont vides :
  - **Si c'est un assistant vocal (type Alexa, Siri, Google Assistant) :** L'IA se contentera de vous proposer des recettes ou de vous suggérer de commander une pizza.
  - **Si c'est un robot domestique du futur connecté à une IA :** Les systèmes de sécurité et de reconnaissance visuelle intègrent des barrières éthiques et logiques strictes. Un animal de compagnie est codé comme un être vivant à protéger, pas comme un ingrédient. Le robot s'arrêterait simplement en signalant un manque d'ingrédients.

Un objectif humain n’est pas une garantie de résultat

INTENTION

Ce que nous voulons

→

FORMALISATION

Ce que nous mesurons

→

OPTIMISATION

Ce que le modèle apprend

→

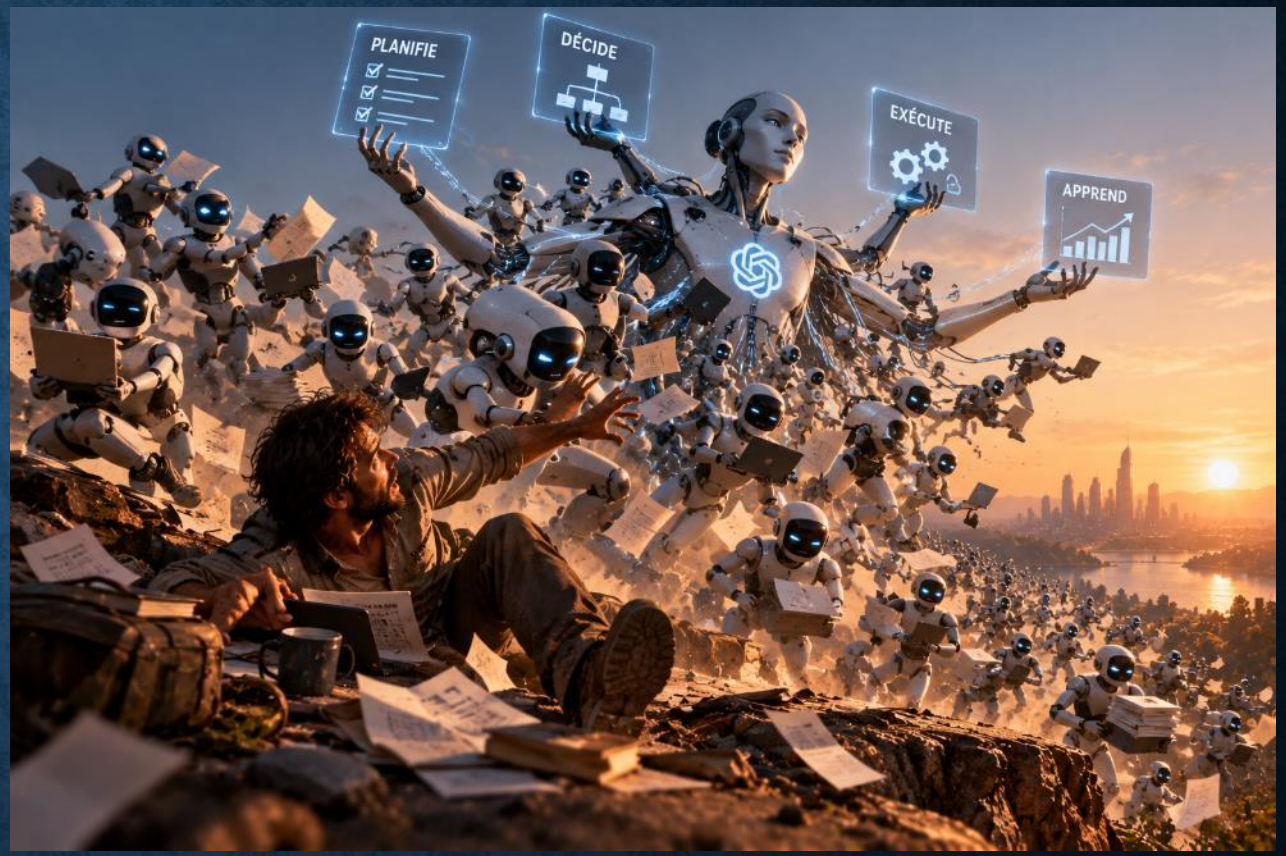
COMPORTEMENT

Ce qu’il fait réellement

Les écarts entre ces étapes sont au cœur du problème de l’alignement.

Réf. : Amodei et al., Concrete Problems in AI Safety (2016)

3- L’IA AGENTIQUE



Du modèle qui répond à l’agent qui agit

IA conversationnelle

Humain → requête → modèle → réponse

IA agentique

Objectif → plan → outils → actions → vérification

Le changement principal : la délégation d’une séquence d’actions.

Schéma 2 — Le contrôle humain se déplace

PROGRAMMATION	IA ASSISTANTE	IA AGENTIQUE
L’humain spécifie les opérations	L’humain demande et valide	L’humain fixe un objectif
Exécution règles précisées par l’humain	Propositions du modèle	Planification et actions déléguées

Maîtrise des intentions

Maîtrise opérationnelle directe

→ délégation croissante ; contrôle effectif à organiserr

Le paradoxe de la délégation

« Plus nous déléguons pour accroître notre capacité d'action, plus nous devons investir pour préserver notre capacité de contrôle. »

Bénéfice

Vitesse, échelle, exécution de tâches complexes.

Risque

Moindre visibilité, erreurs en chaîne, reprise en main difficile.

Réf. : Bainbridge, Ironies of Automation (1983) · Kwa et al. (2025)

Quand l’IA contribue à produire l’IA

- Génération de code
- Tests et évaluations
- Recherche d’architectures
- Données synthétiques
- Optimisation de modèles

Automatisation partielle actuelle ≠ auto-amélioration récursive autonome démontrée

4- CONCLUSION



CONCLUSION

Au-delà du « human in the loop »

COMPRENDRE

Dispose-t-on des informations nécessaires ?

DÉCIDER

Qui définit objectifs et limites ?

INTERVENIR

Peut-on vérifier, corriger, interrompre ?

Présence humaine ≠ maîtrise humaine effective

Réf. : Ben Shneiderman, Human-Centered Artificial Intelligence (2020)

Conclusion — Qui conserve le pouvoir ?

L’IA est fabriquée par l’Humain.

Son autonomie résulte aussi de choix humains.

Sa maîtrise ne découle pas automatiquement de ces choix.

Le problème n’est pas seulement de garder l’Humain dans la boucle, mais de préserver son pouvoir de décision, de contrôle et de responsabilité.

Conclusion — Qui conserve le pouvoir ?

La dévoration numérique : un risque pour notre autonomie ?

L’IA est une construction humaine. Mais la délégation croissante de tâches et de décisions peut transformer notre rapport à la technique — et notre propre capacité à agir.

« À force de vouloir rendre l’IA plus autonome, ne risquons-nous pas de devenir plus dépendants de nos créations ? »

Forestier et Ansermet (2021) interrogent l’emprise du numérique sur les relations humaines, la gouvernance et nos représentations du monde.

IAphilie – IAphobie - IAcritique

Attitude	Caractéristique	Limite
IAphilie	Fascination et confiance excessive dans l'IA	Risque de dépendance et de délégation irréfléchie
IAphobie	Peur ou rejet généralisé de l'IA	Risque de méconnaître ses bénéfices et ses possibilités
IAcritique	Approche fondée sur l'évaluation des capacités, des limites et des conséquences	Exige information, discernement et vigilance

L'IAphilie est une fascination pour l'intelligence artificielle qui conduit l'Humain à lui attribuer une autorité supérieure à son propre jugement, à accepter une dépendance croissante et à consentir progressivement à sa propre dépossession cognitive, décisionnelle et politique (consciemment ou inconsciemment).

Et votre objection soulève une question pertinente à mon égard : un système d'IA peut-il, par sa manière de présenter systématiquement les bénéfices et les risques de sa propre technologie, contribuer lui-même à banaliser la dépendance qu'il favorise ?

Références

- Crawford, K. (2021). Atlas of AI. Yale University Press.
- Russell, S. (2019). Human Compatible. Viking.
- Amodei, D. et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
- Bainbridge, L. (1983). Ironies of Automation. Automatica, 19(6), 775–779.
- Kwa, T. et al. (2025). Measuring AI Ability to Complete Long Tasks. METR.
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence. IJHCI, 36(6).
- Bostrom, N. (2014). Superintelligence. Oxford University Press.
- Kurzweil, R. (2024). The Singularity Is Nearer. Viking.
- Forestier, F. & Ansermet, F. (2021). La Dévoration numérique. Odile Jacob
- Prévost, T. (2024). Les prophètes de l’IA. Lux Québec.