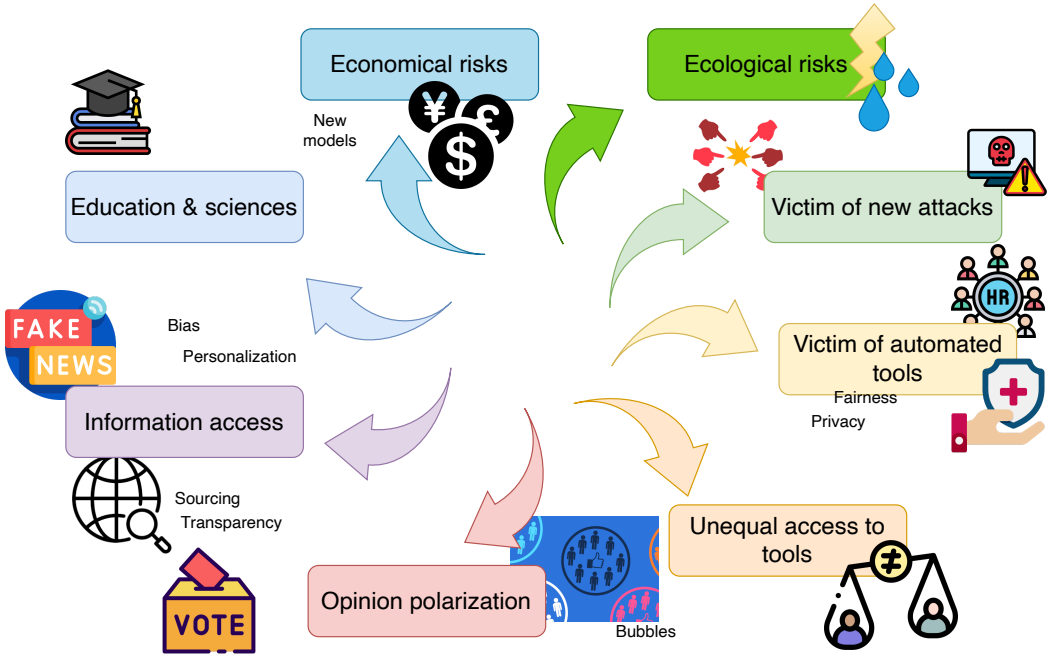
PARTIE 2 :
RISQUES ET ENJEUX

Les enjeux de l'intelligence artificielle dans le monde du travail
16 novembre 2026, Institut des sciences sociales du travail

Vincent Guigue
vincent.guigue@agroparistech.fr
https://vguigue.github.io

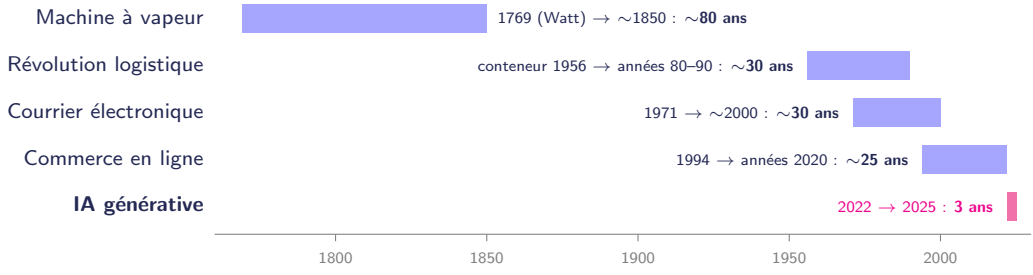
Topologie des risques



Une échelle inédite : frugalité & soutenabilité

Une vitesse de pénétration inédite

Du lancement à la diffusion de masse (ordres de grandeur)



Bornes indicatives – à affiner selon la définition de "diffusion de masse" retenue

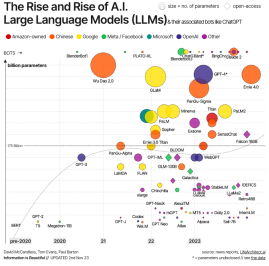
80 ans pour la vapeur, ∼**30 ans** pour l'e-mail et le e-commerce, **3 ans** pour l'IA générative

⇒ ni la société, ni le droit, ni la formation n'ont le temps de s'adapter

Une échelle jamais vue : modèles × usages × usagers

Des modèles géants

- 2018 : BERT, 0,3 Md de paramètres
- 2020 : GPT-3, 175 Md
- 2024 : DeepSeek-V3, 671 Md (37 Md actifs)
- Modèles frontières : >1 000 Md (estimations, non publiés)



Une population d'usagers immense

- chatGPT seul : ∼1 milliard d'utilisateurs / semaine
- Intégré dans les moteurs de recherche, la bureautique, les téléphones...
- Chaque utilisateur : de plus en plus de requêtes, de plus en plus longues

Jamais un logiciel aussi **lourd** n'a été utilisé par **autant** de personnes, **aussi souvent**.

Pourquoi de si gros modèles ?

1. Connaissance paramétrique

Les faits sont **stockés dans les poids** : répondre sans source externe exige de "tout mémoriser"

⇒ Alternative : petit modèle + recherche documentaire (RAG)



2. Capacités multi-tâches

Traduire, coder, résumer, raisonner, voir... dans **un seul** modèle

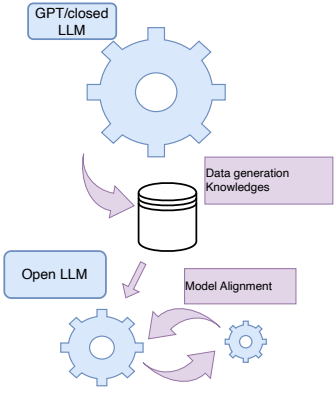
Lois d'échelle : la performance croît avec la taille (données, calcul)

⇒ Alternative : modèles spécialisés pour une tâche

Question de sobriété : a-t-on besoin d'un modèle généraliste pour chaque tâche ?

L'industrialisation : réduire le coût par requête

Distillation



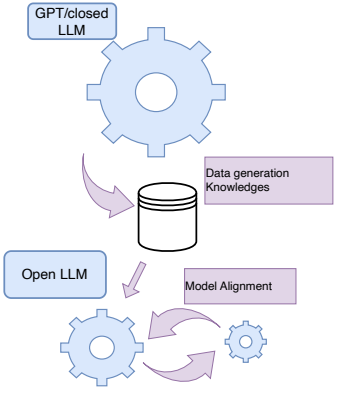
Élagage
Quantification

Mélange d'experts

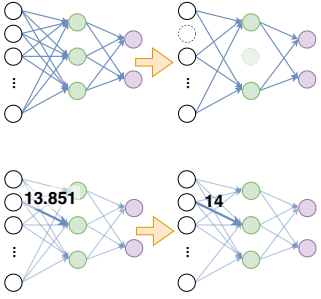
Taille des modèles $\times$**1000 en 3 ans**... puis coût par requête $\div$**100 en 2 ans**

L'industrialisation : réduire le coût par requête

Distillation



Élagage Quantification



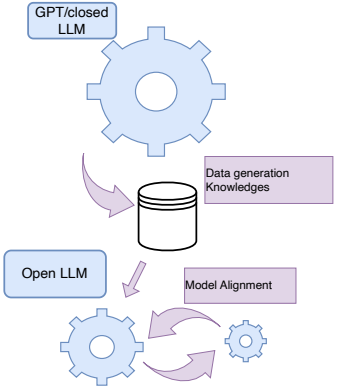
FP32 ⇒ INT4

Mélange d'experts

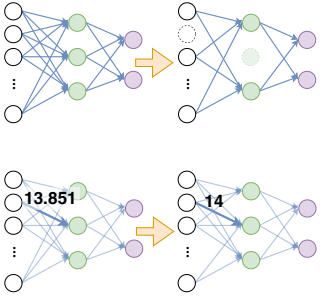
Taille des modèles ×**1000 en 3 ans**... puis coût par requête ÷**100 en 2 ans**

L'industrialisation : réduire le coût par requête

Distillation

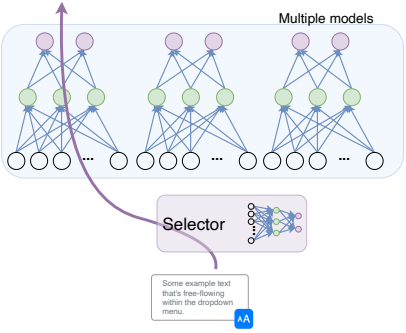


Élagage Quantification



FP32 ⇒ INT4

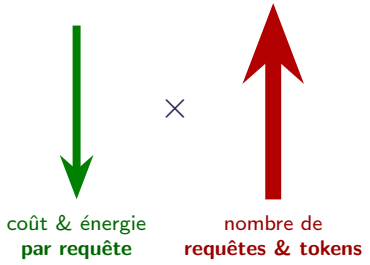
Mélange d'experts



+ industrialisation du code (serveurs, cache, puces dédiées)

Taille des modèles $\times$**1000 en 3 ans**... puis coût par requête $\div$**100 en 2 ans**

Les effets rebond



Paradoxe de Jevons (1865) : plus efficace ⇒ plus utilisé

- Multiplication des usages : IA dans chaque recherche web, chaque document, chaque téléphone
- Modèles de raisonnement : 10 à 100× plus de tokens par réponse
- Vague agentique : un objectif = des dizaines/centaines d'appels au modèle, en boucle

- Une requête texte : ∼0,24 Wh (Gemini, Google 2025)... **mais des milliards par jour**
- Data centers : ∼415 TWh (2024) → ∼945 TWh (2030)

AIE, *Energy & AI*, 2025

Frugalité : quels leviers ?

Individu

- Le bon outil : un moteur de recherche n'est pas un LLM
- Le bon modèle : petit si suffisant
- Éviter le "par défaut" (raisonnement, agents)

Organisation

- Mesurer (EcoLogits, ACV fournisseurs)
- Modèles spécialisés, locaux
- Critères environnementaux dans les achats

Société

- Transparence des fournisseurs (énergie, eau)
- Référentiel AFNOR IA frugale
- Planification énergétique

La frugalité n'est pas seulement technique : c'est aussi/d'abord une question d'usages.

Frugalité : quels leviers ?

L'IA comme levier de sobriété et de transition :

- **Énergie** : pilotage des réseaux électriques, refroidissement des data centers (DeepMind, 2016 : $-40\,\%$ d'énergie de refroidissement)
- **Climat** : prévision météo par apprentissage (GraphCast...), suivi de la déforestation par satellite
- **Agriculture** : pulvérisation ciblée, détection précoce des maladies, irrigation de précision
- **Matériaux** : recherche de nouveaux matériaux pour batteries, catalyseurs

Souvent avec des modèles **petits et spécialisés**... très loin des LLMs géants

Accès à l'information & biais

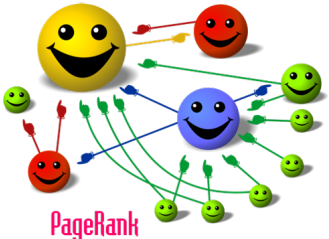
Accès à l'information

- Accès à des informations dangereuses/interdites
  - +Données personnelles
  - Droit à l'oubli numérique (RGPD)
  - Traitement efficace de masses de données personnelles
- Autorités informationnelles
  - Nature : inconsciemment, image = vérité
  - Source : presse, réseaux sociaux, ...
  - Volume : nombre de variantes, citations (pagerank)

  $\Rightarrow$ Un risque majeur pour la démocratie
- Anthropomorphisation de l'algorithme
  - Distinguer humain et machine







Risques personnels : quand l'information blesse

- Harcèlement & contenus blessants
  - hypertrucages (*deepfakes*) sexuels non consentis
  - harcèlement automatisé, à grande échelle
- Hameçonnage & arnaques personnalisés
  - mails sans fautes, adaptés à la cible
  - clonage de voix : l'arnaque "au faux proche"
- Diffamation par hallucination
  - un LLM attribue des crimes à une personne réelle
    (plainte noyb, Norvège, 2025)
- Exposition des données personnelles (cf. droit à l'oubli)



Risques collectifs : démocratie & information libre

- Désinformation à coût quasi nul
  - faux enregistrements à la veille d'élections (Slovaquie, 2023)
  - fermes de sites d'"actualité" générés automatiquement
- Pollution informationnelle
  - contenus générés en masse ; *empoisonnement* des futurs modèles
  - le doute généralisé : "tout peut être faux"
- Concentration de l'accès à l'information
  - quelques assistants = quelques lignes éditoriales
  - réponses sans clic : moins de trafic pour la presse, Wikipédia...

qui finance l'information demain ?



Apprentissage automatique & biais



Oreilles pointues,
moustaches, texture de poils
=
Chat



Homme blanc, +40ans,
costume
=
Cadre supérieur

Biais dans les données ⇒ biais dans les réponses

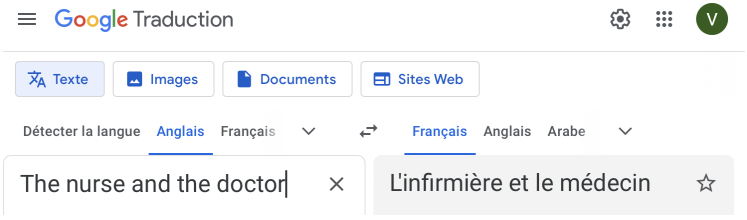
L'apprentissage automatique repose sur l'extraction de biais statistiques...

⇒ Lutter contre les biais = ajustement manuel de l'algorithme

Apprentissage automatique & biais



Stéréotypes tirés de *Pleated Jeans*



- Choix du genre
- Couleur de peau
- Posture
- ...

Biais dans les données ⇒ biais dans les réponses

L'apprentissage automatique repose sur l'extraction de biais statistiques...
⇒ Lutter contre les biais = ajustement manuel de l'algorithme

Biais des données ≠ biais de décision

Biais dans les données

- Représentativité (qui est dans les données ?)
- Étiquettes historiques (décisions passées, elles-mêmes biaisées)
- Variables "proxy" (code postal ≈ origine)

⇒ le modèle **apprend** le monde tel qu'il a été (vu)

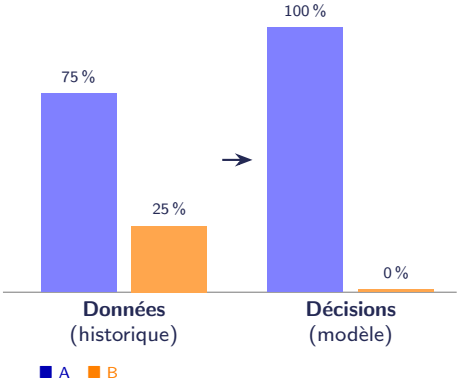
Biais dans la décision

- Règle de décision : on choisit l'option **la plus probable**
- Seuils, coûts d'erreur asymétriques
- Boucles de rétroaction : les décisions d'aujourd'hui sont les données de demain

⇒ le système **amplifie** le monde tel qu'il a été

Corriger les données ne suffit pas : il faut aussi auditer la façon dont on décide.

De 75/25 à 100/0 : l'amplification des biais



- Dans l'historique : A choisi 3 fois sur 4
- Si les autres variables sont peu informatives, la décision **optimale** (au sens de l'erreur moyenne) est... **toujours A**
- Un biais de 75/25 devient une **exclusion totale**
- Puis ces décisions alimentent les données suivantes...

Un modèle **correct statistiquement** peut être **inacceptable socialement**.

Un cas d'école : le tri des CV chez Amazon

- 2014 : outil de notation automatique des CV (postes techniques)
- Entraîné sur **10 ans de recrutements**... majoritairement masculins
- Résultat : pénalise les CV contenant "*women's*" ("capitaine du club d'échecs *féminin*") et certaines universités féminines
- Corrections ponctuelles insuffisantes ⇒ **abandon** (révélé en 2018)

J. Dastin, Reuters, oct. 2018



Les deux biais à l'oeuvre

Données : le passé du secteur

Décision : classer = reproduire le profil majoritaire

Correction des biais & ligne éditoriale

Correction des biais :



- Sélection de données spécifiques, rééquilibrage
- Censure de certaines informations
- Censure des résultats de l'algorithme

⇒ Travail éditorial... Réalisé par qui ?



- Experts du domaine / cahier des charges
- Ingénieurs, lors de la conception de l'algorithme
- Groupe éthique, lors de la validation des résultats
- Équipe communication / réponses aux utilisateurs

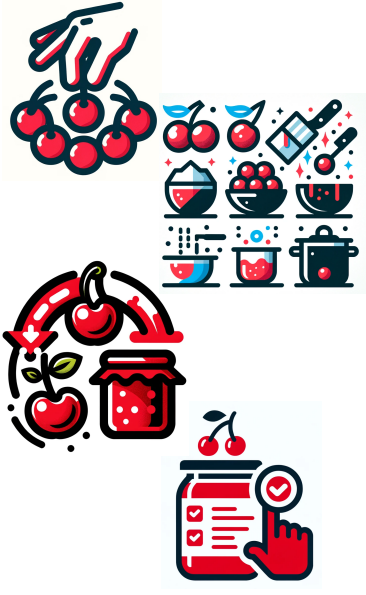


⇒ Quelle légitimité ? Quelle transparence ? Quelle efficacité ?

L'apprentissage automatique n'est **jamais** neutre

1. Sélection des données
   - Sources, équilibre, filtrage
2. Transformation des données
   - Sélection + combinaison d'informations
3. Connaissances a priori
   - Équilibre, fonction de perte (loss), a priori, choix des opérateurs...
4. Filtrage des sorties
   - Post-traitement
   - Censure, redirection, ...

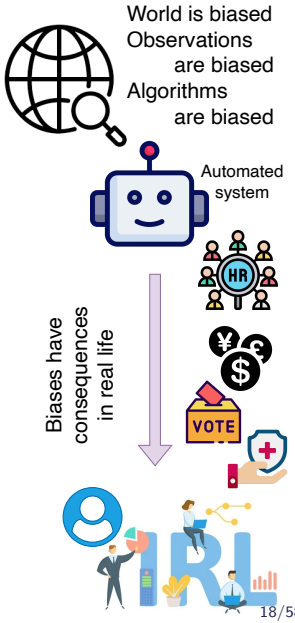
⇒ Des choix qui influencent les résultats de l'algorithme



Conséquences des biais (& exploitation)

- Résiliation d'assurances santé (entreprises minimisant leurs risques)
- Exclusion (automatisée ?) de certaines catégories de personnes :
  - d'une offre d'emploi, d'un recrutement,
  - d'une aide sociale,
  - d'un logement,
  - d'un crédit, ...
- Flux d'information ciblé
  - éventuellement sans consentement de l'utilisateur

  $\Rightarrow$ Crée / renforce les bulles informationnelles



Multiplication des usages de l'IA $\Rightarrow$ multiplication des risques

Mêmes technologies, usages bénéfiques

Les mêmes modèles de langue et de classification servent aussi à :

- **Accessibilité** : traduction, simplification (FALC), lecture & description d'images pour les personnes malvoyantes
- **Traitement documentaire** : recherche dans des archives, synthèse de dossiers (médicaux, juridiques)
- **Santé** : dépistage sur images (rétinopathie diabétique, mammographies)
- **Lutte contre les biais** : auditer des décisions passées pour **détecter** des discriminations
- **Information** : aide à la vérification des faits, détection de campagnes coordonnées

⇒ La régulation doit porter plus sur les usages que sur les technologies

Rapport des utilisateurs aux robots conversationnels

Profilage / personnalisation de l'accès à l'information





- Une bonne réponse à la **surcharge informationnelle**
  - Filtre / agent personnel
  - Système proactif : agent de veille + notifications
- **Bulles informationnelles** = communautés dans les réseaux sociaux
  - Point de vue uniforme (vision tunnel)
  - Illusion de consensus
  - Notion de *leader d'opinion*
- **Polarisation des groupes**
  - renforcement émotionnel
  - Biais de confirmation ↗↗↗
  - Diffusion rapide de la désinformation

[Cambridge Analytica]

Données personnelles & droit à l'oubli

- RGPD art. 17 : **droit à l'effacement**
  pensé pour des bases de données : une ligne, une suppression
- Mais un LLM n'est **pas une base de données**
  - l'information est **diluée** dans 10^{11} poids
  - *machine unlearning* = problème de recherche ouvert
  - ré-entraîner de zéro = plusieurs M€
- Ce qui a été aspiré est **déjà dans les poids**
  - un message de forum de 2009, un fichier fuité, un vieil article
  - *membership inference* / extraction de données d'entraînement
- Qui est **responsable du traitement** ?
  celui qui collecte ? qui entraîne ? qui déploie ?





Même question, deux architectures

RAG : la suppression fonctionne (on retire le document).

Mémoire paramétrique : non.

Biais dans l'acceptation des solutions proposées par IA

Biais cognitif	**Effet sur l'utilisateur**	**IA génératives**	**IA & diag. médical**
Biais d'automatisation	Acceptation passive des résultats (réduction de la vérification)	Élevé	Très élevé
Biais d'autorité	Surestimation de l'autorité (et de la fiabilité) de l'IA	Élevé	Très élevé
Surconfiance / mauvaise calibration de la confiance	Utilisation inappropriée de l'IA dans des situations où elle est peu fiable	Très élevé	Élevé
Biais de confirmation	Renforcement d'hypothèses erronées et réduction de l'exploration d'alternatives	Modéré	Élevé
Décharge cognitive	Diminution de l'analyse critique et de la métacognition	Élevé	Très élevé
Effet de cadrage. Illusion de profondeur explicative	Interprétation biaisée du niveau de certitude ou de risque (style de la réponse, présentation de chiffres)	Très élevé	Modéré
Biais d'anthropomorphisme	Attribution de capacités humaines (confiance++)	Très élevé	Modéré

Mêmes technologies, usages bénéfiques

La personnalisation et le dialogue, quand ils servent l'utilisateur :

- **Tutorat personnalisé** : s'adapter au rythme et aux erreurs de chacun
- **Santé** : préparer une consultation, comprendre un compte rendu médical
- **Recommandation utile** : informations, formations, offres d'emploi, articles scientifiques pertinents
- **Relation usager** : réponses 24h/24 aux questions simples, orientation vers un humain pour le reste [est-ce positif?]

La question n'est pas "personnaliser ou non", mais **au service de qui**.

Inégalités face aux outils d'IA

Des outils qui creusent les écarts ?

Argent

- Gratuit / ∼20 € / ∼200 € par mois : **pas le même modèle**, pas les mêmes agents
- Entreprises équipées *vs* TPE, associations, services publics

Langue

- Même texte, plus de tokens hors anglais ⇒ plus cher, souvent moins bon
- Langues peu dotées : qualité très dégradée

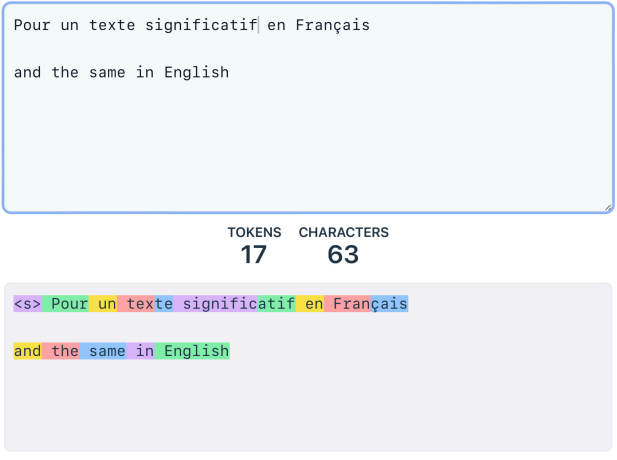
Formation

- Savoir formuler, vérifier, critiquer : les gains vont à ceux qui maîtrisent déjà
- Illectronisme : ∼15 % des Français (INSEE)

Âge, genre, rapport psychologique

- Adoption plus forte chez les jeunes... qui sont aussi les plus exposés sur le marché de l'emploi
- Écart d'usage femmes/hommes à poste égal (Danemark, Humlum & Vestergaard 2024)
- Confiance excessive *vs* rejet : deux façons d'être perdant

Inégalités : le coût de la langue



- Le découpage en tokens est optimisé pour l'anglais
- Facturation **au token**
- Fenêtre de contexte plus vite saturée
- Moins de données d'entraînement ⇒ moins bonnes réponses

Une même question coûte plus cher et reçoit une moins bonne réponse selon la langue dans laquelle on la pose.

Inégalités scolaires

- Inégalité face au prix des outils (depuis septembre 2025)
- Inégalité face aux tentations
  - Qui s'en sert comme un professeur assistant? [montée en compétence rapide sur de nouveaux sujets]
  - Qui s'en sert comme un remplaçant? [perte de compétences rapides & souvent inconsciente]



Mêmes technologies, usages bénéfiques

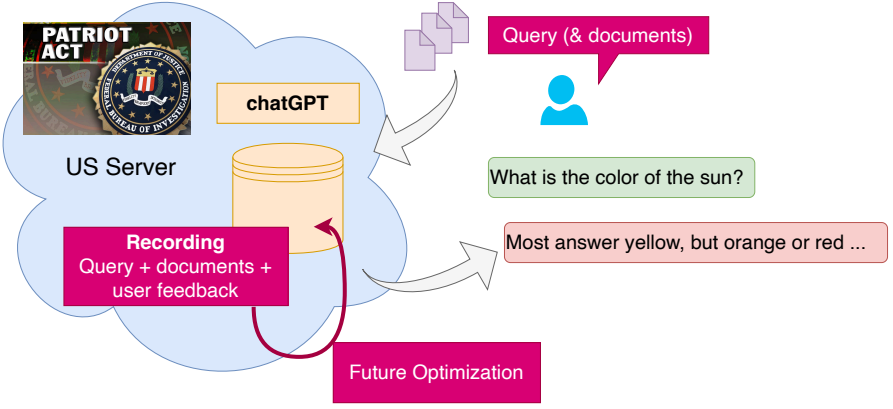
Bien déployées, ces technologies peuvent aussi **réduire** *certaines inégalités :*

- **Handicap** : description d'images, sous-titrage, dictée, aide à la lecture (dyslexie)
- **Démarches administratives** : comprendre un courrier, remplir un formulaire
- **Langue** : traduction pour les non-francophones (santé, école, administration)
- **Éducation** : un tuteur disponible 24h/24, y compris pour ceux qui n'ont pas de soutien à la maison
- **Écrit professionnel** : lettres de motivation, mails formels pour ceux que l'écrit pénalise

À condition que l'accès (coût, formation) ne soit pas lui-même inégal...

Sécurité

Fuites de données



- Transmission de données sensibles
- Exploitation des données par OpenAI (ou d'autres)
- Fuite de données dans de futurs modèles

Des niveaux de risques *vs* sécurisation

Outils

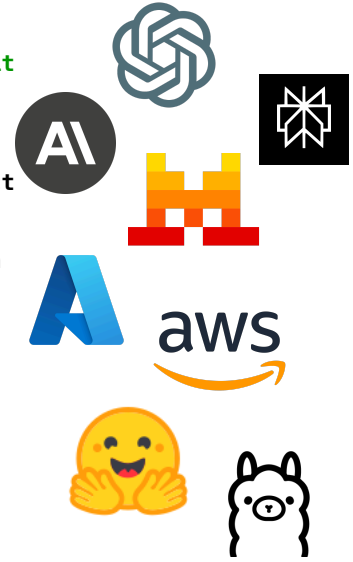
Outil commercial, gratuit
Licences/CGU variables

Outil commercial,
Licence payante
+ garanties / patriot act

Outil commercial
Licence payante + option
e.g. Serveur européen

LLM Institutionnel
Déployé sur un
périmètre controlé

Usage local
Modèles pré-entrainés-
raffinés



Données

Doc. quelconque

Information personnelle

Projet en cours

Enregistrements médicaux

L'IA dans la suite bureautique : de nouvelles fuites

- Copilot (Office 365), Gemini (Google Workspace)... : l'assistant **voit tout ce que vous pouvez voir**
  - Word, Excel, mails, Teams, Drive partagés
  - les droits d'accès mal réglés deviennent **visibles** (*sur-partage*)
- Injection de consignes : un document ou un mail piégé donne des ordres à l'assistant
  - EchoLeak (2025) : exfiltration de données de Copilot sans aucun clic de l'utilisateur
- Activation par défaut : l'IA arrive par mise à jour, sans choix explicite



Pourquoi ne fait-on plus confiance aux licences ?

- Le précédent des droits d'auteur
  - entraînement sur des œuvres protégées, y compris issues de bibliothèques pirates
  - procès en série (NYT *vs* OpenAI...), accord d'Anthropic avec des auteurs (1,5 Md$, 2025)

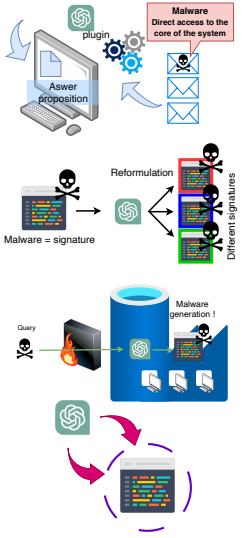
  $\Rightarrow$ si les droits d'auteur n'ont pas été respectés, pourquoi les CGU le seraient-elles ?

- Des conditions d'utilisation mouvantes
  - modifications unilatérales des CGU pour entraîner des modèles (Zoom 2023, Adobe 2024, WeTransfer 2025...)

- Le droit extraterritorial
  - *Cloud Act* : un fournisseur américain peut être contraint de livrer des données, même hébergées en Europe

Problèmes de sécurité : attaquer les machines

- Extensions / agents / MCP ⇒ souvent des failles importantes
  - Accès aux e-mails / transfert d'informations sensibles, etc.
- Problèmes de gestion pour les entreprises
  - Sécurisation de (très) gros fichiers
- Plus d'opportunités de contourner les signatures de logiciels malveillants
  - ≈ reformulation logicielle
- Nouveaux problèmes !
  - Génération directe de logiciels malveillants
- Multiplication des découvertes de failles de sécurité
  - Agents de code, recherche automatisée de vulnérabilités, ...

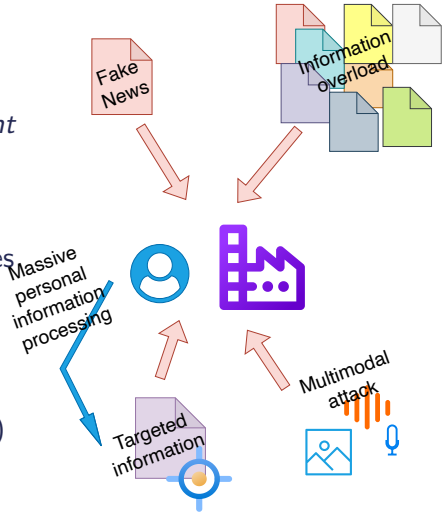


Attaquer les personnes : usurpation & hameçonnage

- Usurpation d'identité
  - Imitation de voix / de style d'écriture
  - Fausse autorité (patron, enseignant, avocat, médecin)

fraude au président

- Ingénierie sociale à grande échelle
  - Messages d'hameçonnage très personnalisés
  - Croisement avec les fuites de données personnelles
- Préjudices informationnels
  - Mésinformation & désinformation
  - Sabotage de décisions (finance, santé, sécurité)
  - Atteintes à l'intégrité (scientifique, journalistique)
  - Diffusion de *deepfakes* pornographiques
- Contrôle étatique
  - La surveillance de masse devient *possible*



Mêmes technologies, usages bénéfiques

L'IA est aussi un outil de défense :

- **Détection de vulnérabilités** : des agents trouvent des failles dans du code avant les attaquants
- **Filtrage** : spam, hameçonnage, fraude bancaire (apprentissage automatique depuis 20 ans)
- **Analyse de journaux** : repérer des comportements anormaux dans un système d'information
- **Données sensibles** : détection et anonymisation automatique de données personnelles dans des documents

Reconnaissance faciale & surveillance

Reconnaissance faciale : une technologie mûre

- Apprentissage profond : performances quasi parfaites en conditions contrôlées (évaluations NIST)
- Coût marginal quasi nul : caméras partout, calcul bon marché
- Vérification (1:1) : "est-ce bien vous ?"
  déverrouillage, passage aux frontières
- Identification (1:N) : "qui est-ce ?"
  vidéosurveillance, recherche dans une base



Le saut de risque est entre **vérifier** un individu consentant et **identifier** n'importe qui, n'importe où.

Traçage & libertés fondamentales

- Fin de l'anonymat dans l'espace public
  - manifester, se réunir, circuler sans être identifié
  - effet dissuasif (*chilling effect*) même sans usage effectif
- Des bases constituées sans consentement
  - Clearview AI : des milliards de photos aspirées sur le web

    amende CNIL de 20 M€ (2022)
- Usages politiques
  - surveillance de minorités, identification d'opposants ou de manifestants (Chine, Russie...)
- Cadre européen & français
  - AI Act (art. 5) : identification biométrique à distance **en temps réel** dans l'espace public **interdite**... sauf exceptions
  - France : vidéosurveillance algorithmique expérimentée (JO 2024) **sans** reconnaissance faciale

Inégaux face aux erreurs

Taux d'erreur (classification du genre)

J. Buolamwini & T. Gebru, *Gender Shades*, 2018



NIST (2019) : faux positifs 10 à 100× plus fréquents selon l'origine pour de nombreux algorithmes

- **Arrestations à tort** aux USA après une fausse identification (R. Williams, Detroit 2020 ; P. Woodruff, 2023)
- L'erreur ne tombe pas au hasard : elle frappe les **groupes sous-représentés** dans les données
- Le coût d'un faux positif : garde à vue, procès, stigmatisation

Biais des données (cf. section 1) × décision automatique × pouvoir de police = **risque maximal**

Mêmes technologies, usages bénéfiques

Les mêmes réseaux de vision par ordinateur servent aussi à :

- **Santé** : aide au diagnostic de maladies génétiques rares à partir des traits du visage ; imagerie médicale
- **Accessibilité** : reconnaître ses proches pour une personne malvoyante
- **Sécurité consentie** : déverrouillage d'un téléphone (vérification 1:1, données sur l'appareil)
- **Agriculture & environnement** : identification d'animaux d'élevage, suivi de la faune sauvage

La technologie est la même : ce sont le **consentement**, la **finalité** et le **contrôle** qui changent tout.

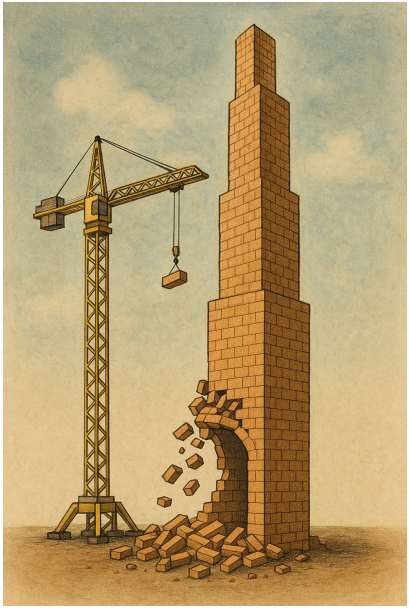
Compétences, formation & cognition

Défi pour la formation (initiale & continue)

- Redéfinir des priorités de formation, sujet par sujet, métier par métier, comme pour Wikipedia/calculatrice/...
  - Accepter la perte/réduction de certaines compétences
  - Se reposer la question régulièrement !
- Former aux LLMs... et savoir parfois les interdire



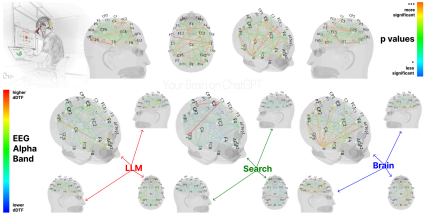
- Savoir reconnaître les contenus générés par LLM
- Ne pas sous-estimer les aspects psychologiques



Évolution des capacités cognitives / psychologie

Notre cerveau va évoluer (= décliner ?) avec ces nouveaux outils...
Quelle est la portée de ces transformations ? Quelles en seront les conséquences ?

- Les sciences de l'éducation et la psychologie l'avaient conjecturé...
  les sciences cognitives l'ont mesuré



Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task, N. Kosmyna et al. arXiv 2025

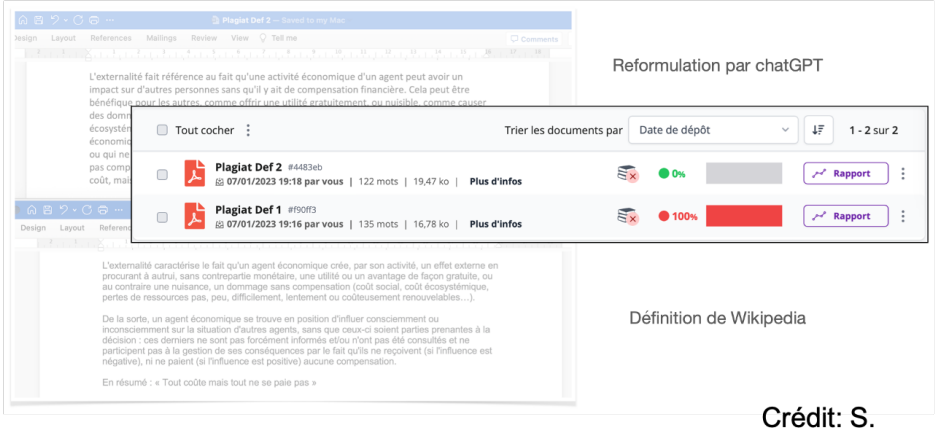
Comment rester motivé quand une *calculatrice* répond à toutes les épreuves ?

- La vie ne ressemble pas à un examen...
- Garder le contrôle...

Mais :

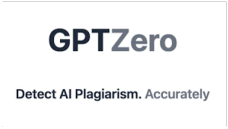
- Illusion de comprendre les solutions
- Tentation trop forte

Détection des *textes générés par* chatGPT



Crédit: S. Pajak

Détection des *textes générés par* chatGPT







- **Classifieur de texte** (comme pour n'importe quel auteur)
  - Détection des biais dans le choix des mots / la formulation
- Caractérisation de la **vraisemblance** du texte (OpenAI, GPTZero)
  - Hyper-fluidité des phrases, surabondance de connecteurs logiques
  - Modèle de langage = statistique $\Rightarrow$ mesure entre distributions (**perplexité**)
- δ-**vraisemblance** sur des textes perturbés (DetectGPT)

Détecteurs $\Rightarrow$ détection $< 100\%$

\+ niveau de confiance dans la détection

− dépend de la longueur du texte et des modifications effectuées

$\approx$ détecte des morceaux de Wikipédia (chatGPT = *perroquet stochastique*)

⚠ détecte la traduction comme la génération

Mêmes technologies, usages bénéfiques

Bien utilisée, l'IA peut aussi **renforcer** *les compétences :*

- **Oser** : briser la peur de la page blanche, poser les "questions bêtes"
- **S'entraîner** : exercices, quiz, simulations d'entretien à volonté
- **Explorer** : entrer dans un domaine nouveau, un code inconnu, une langue étrangère
- **Formation continue** : se former au poste de travail, au moment du besoin

La même aide renforce **ou** remplace la compétence : tout dépend de **qui fait l'effort**.

Économie & responsabilité

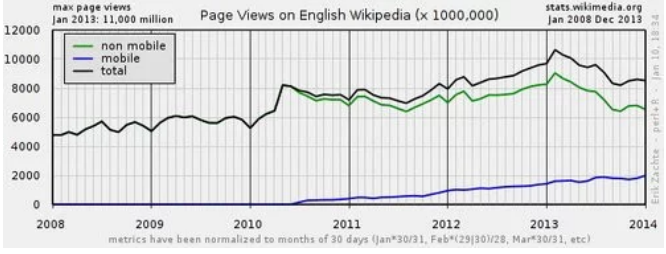
Questions économiques

- Financement/Publicité ⇔ **visites** des internautes
- Google Knowledge Graph (2012) ⇒ moins de visites, donc moins de revenus
- chatGPT = encodage de l'information du web... ⇒ encore moins de visites ?

⇒ Quel **modèle économique / sources d'information** avec chatGPT ?

Google's Knowledge Graph Boxes: killing Wikipedia?

by Gregory Kohs



⇒ Qui **bénéficie du retour d'information** ? [StackOverflow]

Rémunérer les ayants droit

Le corpus d'entraînement est **le travail de quelqu'un d'autre** :
presse, livres, code, images, articles scientifiques

Trois régimes coexistent aujourd'hui :

- **Contentieux** *NYT vs OpenAI, Getty vs Stability, auteurs*
- **Accords de licence** *groupes de presse, Reddit, Stack Overflow*
- **Opt-out** *directive européenne DSM (exception TDM), robots.txt*



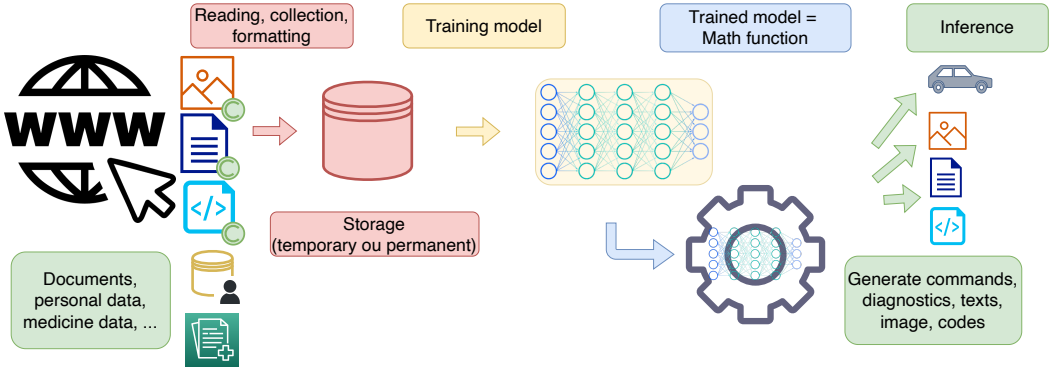


Le problème structurel : **la contribution n'est pas traçable**

- Pas d'attribution par token ⇒ pas de rémunération proportionnelle
- Les accords forfaitaires favorisent les **gros** ayants droit
  ⇒ l'auteur individuel ne touche rien

⇒ La boucle se referme : moins de visites → moins de production → **dégradation du corpus**.

La chaîne des risques & des responsabilités



Copyright and database law

Right to collect, right to copy, consent

Right to use data in an algorithm **Optout**

Model = emanation of data?

Clearview.ai

Cambridge Analytica

Reproductions of untraceable extracts

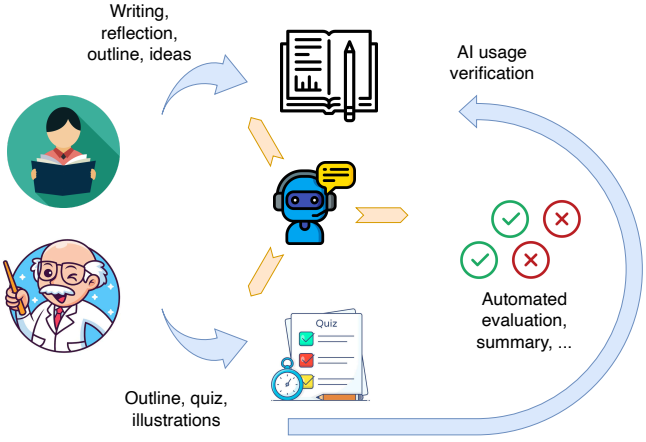
Usage regulation

Responsibility for errors

Risques liés à la généralisation de l'IA

L'IA partout =
perte de sens ?

- Dans le domaine éducatif
- Transposition aux RH
- Aux systèmes de financement par projet



Mêmes technologies, usages bénéfiques

Des gains réels quand la valeur est partagée :

- **Traitement documentaire** : factures, contrats, dossiers administratifs
  moins de tâches répétitives
- **Recherche** : prédiction de structures de protéines (AlphaFold, prix Nobel 2024), criblage de molécules
- **Agriculture** : conseil agronomique, détection des maladies, traçabilité
- **PME & services publics** : accès à des capacités autrefois réservées aux grands groupes

Discussion éthique

Quelle approche de la question éthique ?

Médecine

1. **Autonomie :** le patient doit pouvoir prendre des décisions éclairées.
2. **Bienfaisance :** obligation d'agir pour le bien, dans l'intérêt du patient.
3. **Non-malfaisance :** éviter de causer du tort, évaluer les risques et les bénéfices.
4. **Égalité :** équité dans la répartition des ressources et des soins de santé.
5. **Confidentialité :** garantir la confidentialité des informations du patient.
6. **Vérité et transparence :** fournir une information honnête, complète et compréhensible.
7. **Consentement éclairé :** obtenir le consentement libre et éclairé des patients.
8. **Respect de la dignité humaine :** traiter chaque patient avec respect et dignité.

Intelligence artificielle

1. **Autonomie :** les humains gardent le contrôle du processus
2. **Bienfaisance :** dans l'intérêt de qui ? Utilisateur + GAFAM...
3. **Non-malfaisance :** humains + environnement / durabilité / usages malveillants
4. **Égalité :** accès à l'IA et égalité des chances
5. **Confidentialité :** qu'en est-il du modèle économique de Google/Facebook ?
6. **Vérité et transparence :** la tragédie de l'IA moderne
7. **Consentement éclairé :** des cookies aux algorithmes, savoir quand on interagit avec une IA
8. **Respect de la dignité humaine :** comportements de harcèlement / distinction humain–machine

Quelle approche de la question éthique ?

Médecine

1. **Autonomie :** le patient doit pouvoir prendre des décisions éclairées.
2. **Bienfaisance :** obligation d'agir pour le bien, dans l'intérêt du patient.
3. **Non-malfaisance :** éviter de causer du tort, évaluer les risques et les bénéfices.
4. **Égalité :** équité dans la répartition des ressources et des soins de santé.
5. **Confidentialité :** garantir la confidentialité des informations du patient.
6. **Vérité et transparence :** fournir une information honnête, complète et compréhensible.
7. **Consentement éclairé :** obtenir le consentement libre et éclairé des patients.
8. **Respect de la dignité humaine :** traiter chaque patient avec respect et dignité.

Intelligence artificielle

1. **Autonomie :** les humains gardent le contrôle du processus
2. **Bienfaisance :** dans l'intérêt de qui ? Utilisateur + GAFAM...
3. **Non-malfaisance :** humains + environnement / durabilité / usages malveillants
4. **Égalité :** accès à l'IA et égalité des chances
5. **Confidentialité :** qu'en est-il du modèle économique de Google/Facebook ?
6. **Vérité et transparence :** la tragédie de l'IA moderne
7. **Consentement éclairé :** des cookies aux algorithmes, savoir quand on interagit avec une IA
8. **Respect de la dignité humaine :** comportements de harcèlement / distinction humain–machine

Discussion : jusqu'où puis-je aller ?

Dans mon travail, puis-je utiliser un assistant conversationnel pour :

1. chercher des idées ? me faire expliquer une notion technique ou réglementaire ?
2. traduire un paragraphe d'un contrat ?
3. rédiger un e-mail ou un compte rendu à partir de mes notes ?
4. faire une veille documentaire, chercher des références ?
5. résumer un rapport que je dois lire ?
6. résumer un document que je dois **valider** ou **signer** ?
7. rédiger l'évaluation annuelle d'un collaborateur ?
8. trier des candidatures ? poser des questions sur un document confidentiel ?
9. écrire 10 lignes de code pour automatiser un tableur ?
10. laisser un agent agir à ma place (envoyer, commander, modifier) ?

Où placez-vous la limite ? Et surtout : **pourquoi là** ?

Discussion : et si ça tourne mal ?

Quelle est ma réaction si :

1. mon manager trouve une erreur dans un paragraphe rédigé par un chatbot ;
2. mon manager **ne trouve pas** une erreur existante dans ce paragraphe ;
3. un client trouve une erreur dans des résultats... produits par du code généré ;
4. je découvre que la synthèse sur laquelle j'ai fondé une décision n'était pas fiable ;
5. je découvre que le document confidentiel que j'ai soumis à une API a servi à entraîner un modèle ;
6. je découvre que mon propre travail est évalué (ou surveillé) par une IA ;
7. un collègue produit deux fois plus vite que moi grâce à l'IA... et la norme change pour tous.

Qui est responsable ? Le salarié, l'employeur, l'éditeur du modèle, personne ?

Trois questions pour structurer le débat

1. La question de la **délégation**

Qu'est-ce que j'accepte de ne plus faire moi-même ?
Et que suis-je encore capable de vérifier une fois que je ne le fais plus ?

2. La question de la **transparence**

Dois-je déclarer l'usage de l'IA ? À qui ? Au client, au manager, au régulateur ?
Que change une déclaration ? Que change son absence ?

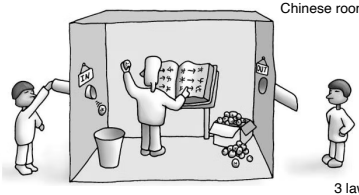
3. La question de l'**asymétrie**

Ce que j'accepte pour moi, l'accepterais-je de la part d'un collègue ?
D'un prestataire ? De mon employeur à mon sujet ?

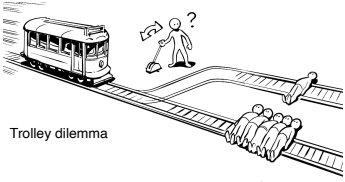
Conclusion

Les représentations de l'Intelligence Artificielle

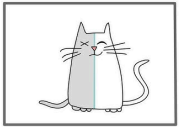
Chinese room



Trolley dilemma



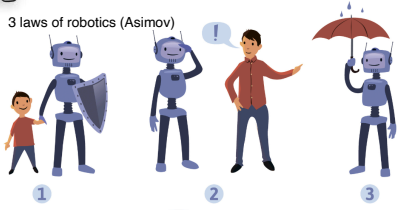
Schrodinger's cat



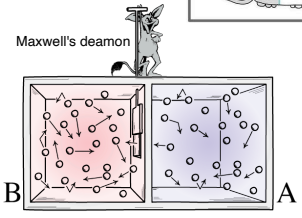
Mary's room



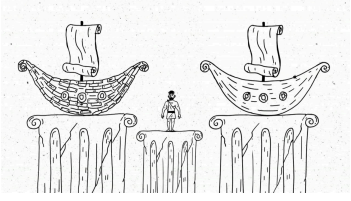
3 laws of robotics (Asimov)



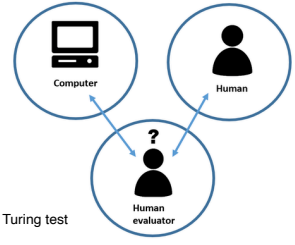
Maxwell's deamon



Thesee's boat



Turing test





Défis à venir

- **Qu'en est-il des hallucinations ?**
  - Faut-il chercher à les réduire ou apprendre à vivre avec ?
  - Les LLM vont-ils s'améliorer ? Dans quelles directions ?
  - Les LLM nous font-ils *perdre* notre lien à la vérité ? À la vérification ?
- **Avons-nous besoin de petits ou de grands modèles de langue ?**
  - Combien cela coûte-t-il ? Est-ce durable ?
  - Avec ou sans ajustement fin (fine-tuning) ?
  - Que signifie la frugalité dans le monde des LLM ?
- **Quand les autres les utilisent... Quel impact cela a-t-il sur moi ?**
  - Productivité (chercheurs, codeurs, relecteurs, ...)
  - Éducation : gestion / formation d'étudiants *technophiles*
- **Protection des données... les miennes et celles des autres**
  - Est-il raisonnable d'entraîner des LLM sur GitHub, Wikipédia, des articles scientifiques, des sites d'actualités, etc. ?
  - Quelle importance accorder à la vie privée ? Quels sont les risques liés à l'usage d'un LLM ?

Défis à venir

- **Qu'en est-il des hallucinations ?**
  - Faut-il chercher à les réduire ou apprendre à vivre avec ?
  - Les LLM vont-ils s'améliorer ? Dans quelles directions ?
  - Les LLM nous font-ils *perdre* notre lien à la vérité ? À la vérification ?
- **Avons-nous besoin de petits ou de grands modèles de langue ?**
  - C
  - A
  - Q

Le smartphone a fait de moi un *humain augmenté*...
Le LLM fera-t-il de moi un *chercheur augmenté* ?
⇒ Jetez donc un œil à **NotebookLM**

- **Quand les autres les utiliseront : Quel impact cela a-t-il sur moi ?**
  - Productivité (chercheurs, codeurs, relecteurs, ...)
  - Éducation : gestion / formation d'étudiants *technophiles*
- **Protection des données... les miennes et celles des autres**
  - Est-il raisonnable d'entraîner des LLM sur GitHub, Wikipédia, des articles scientifiques, des sites d'actualités, etc. ?
  - Quelle importance accorder à la vie privée ? Quels sont les risques liés à l'usage d'un LLM ?

Niveaux d'accès à l'intelligence artificielle

1. Utilisateur via une interface : *chatGPT*
   - Une formation reste nécessaire (2–4 h)
2. Utilisation de bibliothèques Python
   - Bases sur les protocoles
   - Chaînes de traitement standards
   - Formation : 1 semaine à 3 mois (ML/DL)
3. Développeur d'outils
   - Adapter les outils à un cas spécifique
   - Intégrer des contraintes métier
   - Construire des systèmes hybrides (mécanistes / symboliques)
   - Combiner texte et images
   - Formation : $\geq$ 1 an

Souveraineté numérique : toute la chaîne

Pre-trained model construction

Data Mastery
- Collection/balance
- Cleaning

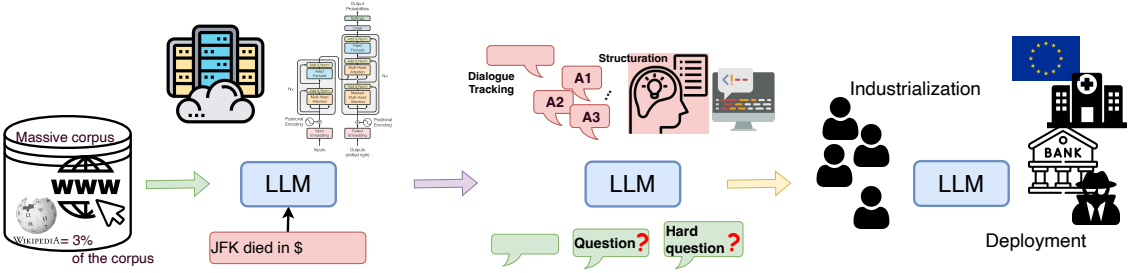
Training
- Computation power (x1000 GPU)
- Architecture ML

Model Fine-Tuning

Data Mastery & Construction
- Human interactions +++
- Dataset cost
- Domain adaptation

Model exploitation

Optimization / Cost reduction
- MLOps skills
- Local deployment



Une multitude de métiers

Data architect / manager
- Data management & hardware devices (storage, network, …)

Data Engineer
- Update & Query on the data

Data Analyst
- Data visualization (chart, indicators, …)
- Statistical trends

Prompt Engineer
- Query on LM/foundation models with "prompts"

Data Scientist
- Query the data / critical selection & balance
- Algorithm development / adaptation / evaluation
- Advanced data visualization

MLOps Engineer
- Algorithm optimization
- Industrialize software solutions

+ **DPO** :
Data Protection Officer

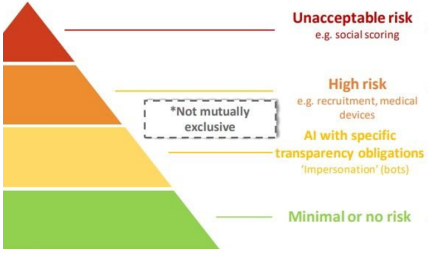






Facteurs d'acceptabilité de l'IA générative

1. **Utilitarisme** :
   - Performance (facteur d'acceptation de ChatGPT)
   - Fiabilité / auto-évaluation
2. **Non-dangerosité** :
   - Biais / correction
   - Transparence (ligne éditoriale, confusion humain/machine)
   - Mise en œuvre fiable
   - Souveraineté (?)
   - Régulation (AI Act)
     - Éviter les applications dangereuses
3. **Savoir-faire** :
   - Formation (utilisation / développement)



Pourquoi tant de controverses ?

- Nouvel outil [Décembre 2022]
- + Vitesse d'adoption sans précédent [1 million d'utilisateurs en 5 jours]
- Forces et faiblesses... mal comprises par les utilisateurs
  - Gains de productivité importants
  - Usages surprenants / parfois absurdes
  - Biais / usages dangereux / risques
- Retours mal interprétés
  - Anthropomorphisation de l'algorithme et de ses erreurs
- Coût prohibitif : quel modèle économique, écologique et sociétal ?

