IA & Cognition
Panorama des avancées en IA

Marie Lefevre

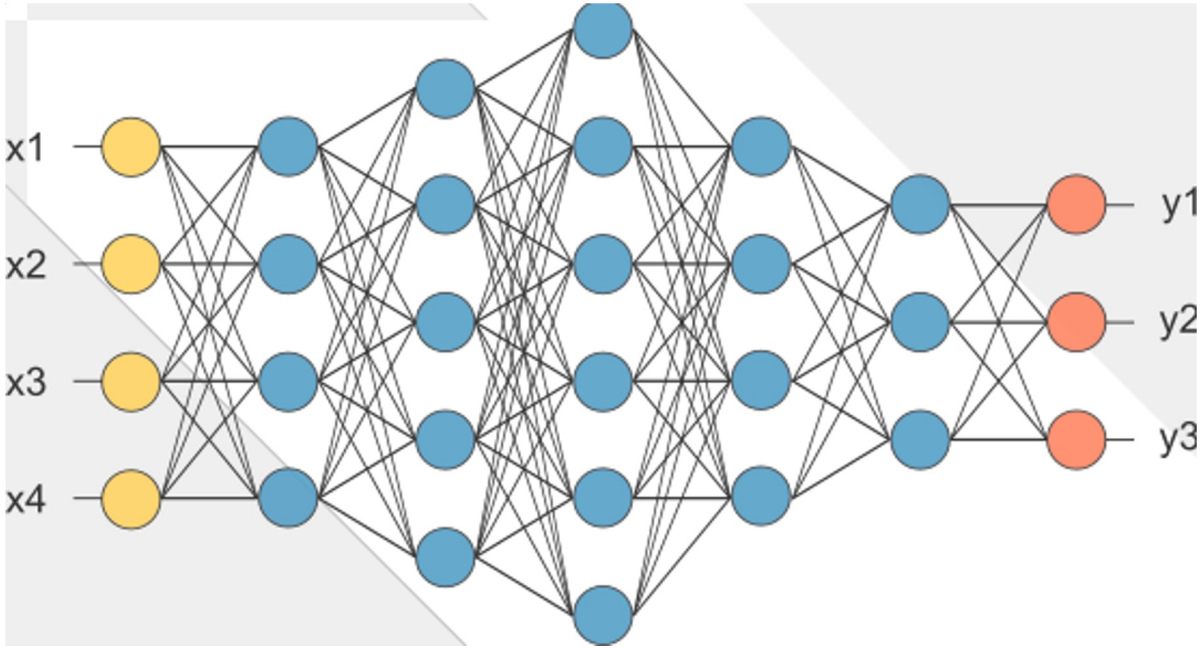
Université Lyon 1 - Laboratoire LIRIS

marie.lefevre@liris.cnrs.fr

L'IA aujourd'hui : le Deep Learning

Entrées

Pour lui plus de plaisir, plus d'hymen, plus d'amour.
Seul parmi les horreurs d'un sauvage séjour,
Dans ces noires forêts du soleil ignorées,
Sur les sommets déserts des monts hyperborées,
Il pleurait Eurydice, et, plein de ses attraits,
Reprochait à Pluton ses perfides bienfaits.
En vain mille beautés s'efforçaient de lui plaire :
Il dédaigna leurs feux, et leur main sanguinaire,
La nuit, à la faveur des mystères sacrés,
Dispersa dans les champs ses membres déchirés.
L'Hèbre roula sa tête encor toute sanglante :
Là, sa langue glacée et sa voix expirante,
Jusqu'au dernier soupir formant un faible son,
D'Eurydice, en flottant, murmurait le doux nom :
Eurydice ! ô douleur ! Touchés de son supplice,
Les échos répétaient, Eurydice ! Eurydice !

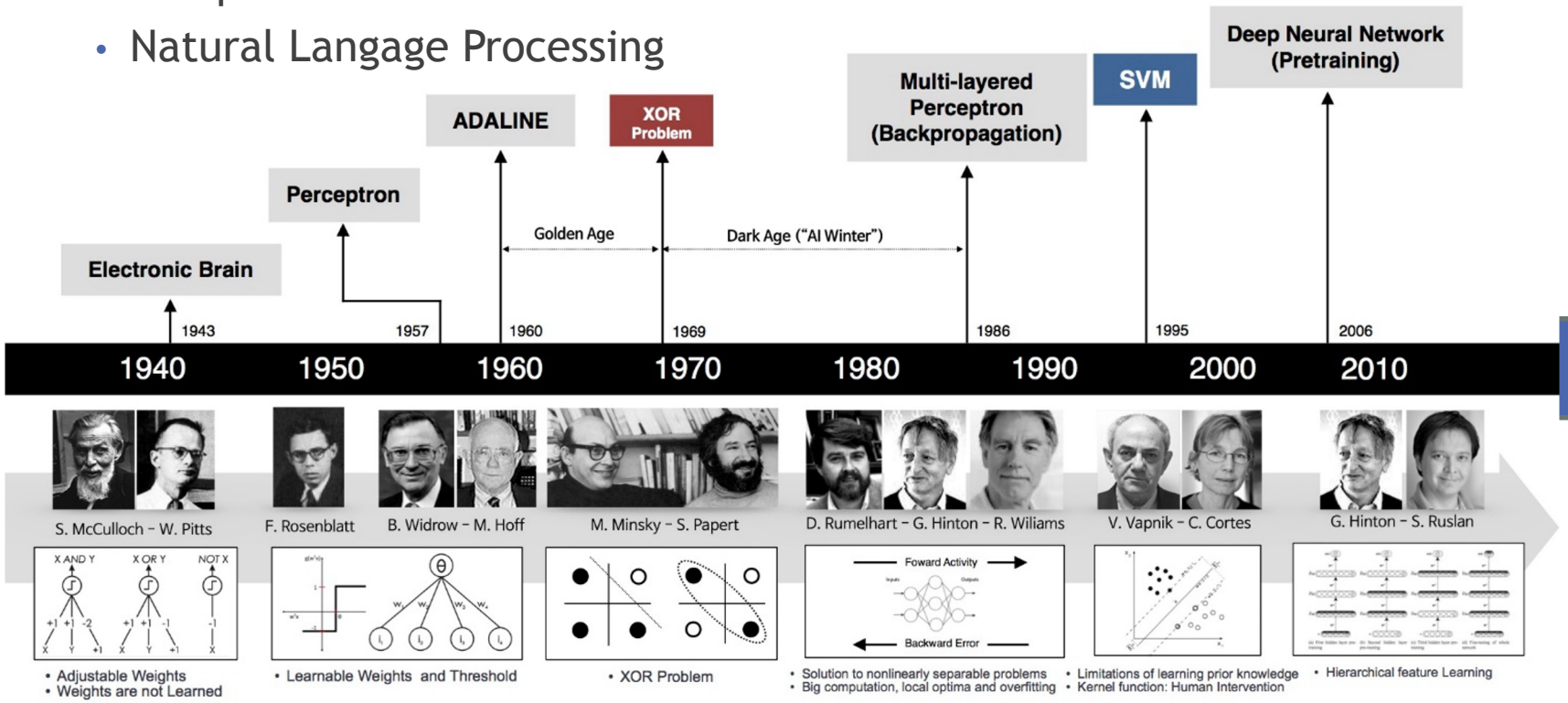


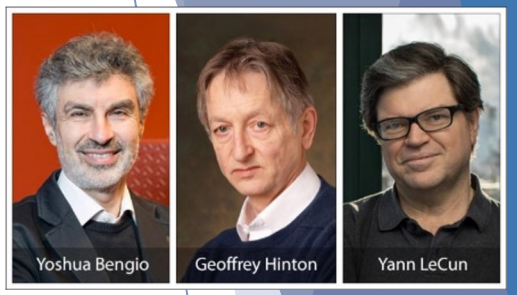
Sorties

Une « machine » qui transforme des nombres en d'autres nombres en « regardant » une base d'apprentissage

L'IA aujourd'hui : le Deep Learning

- Il faut des données : souvent images, texte
  - Computer vision
  - Natural Langage Processing



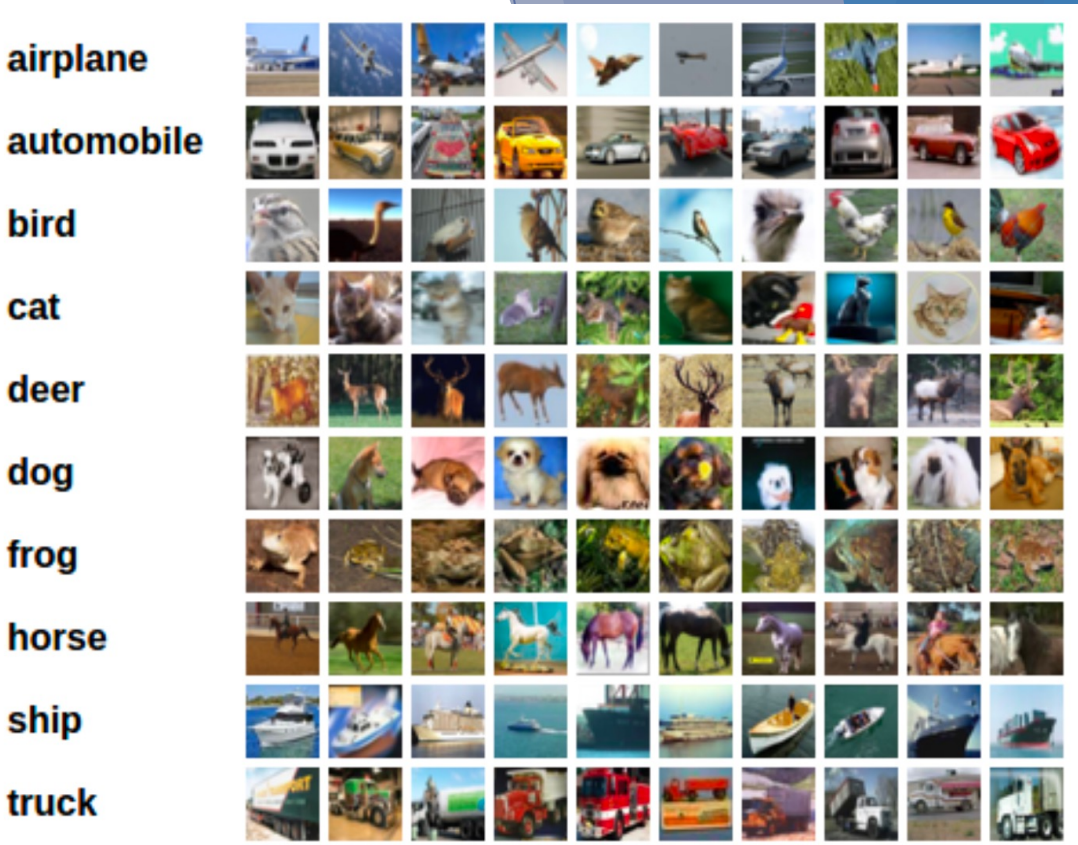


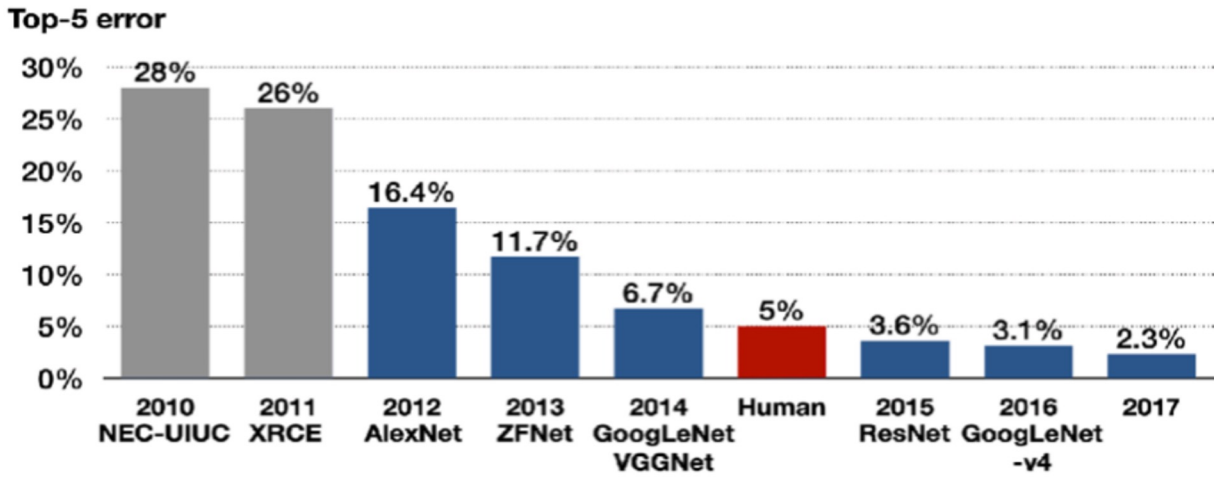
Age d'or du Deep Learning : données + GPU

CNN, AE, GAN, Transformer, …

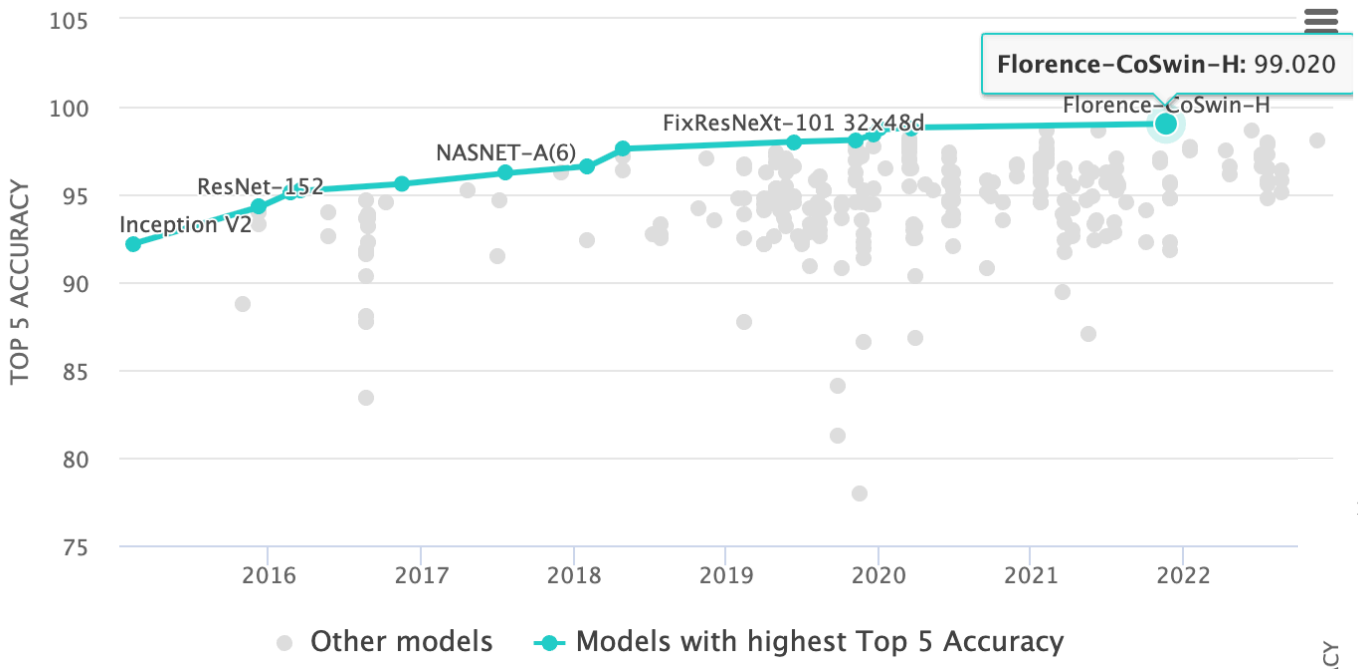
Concours ImageNet

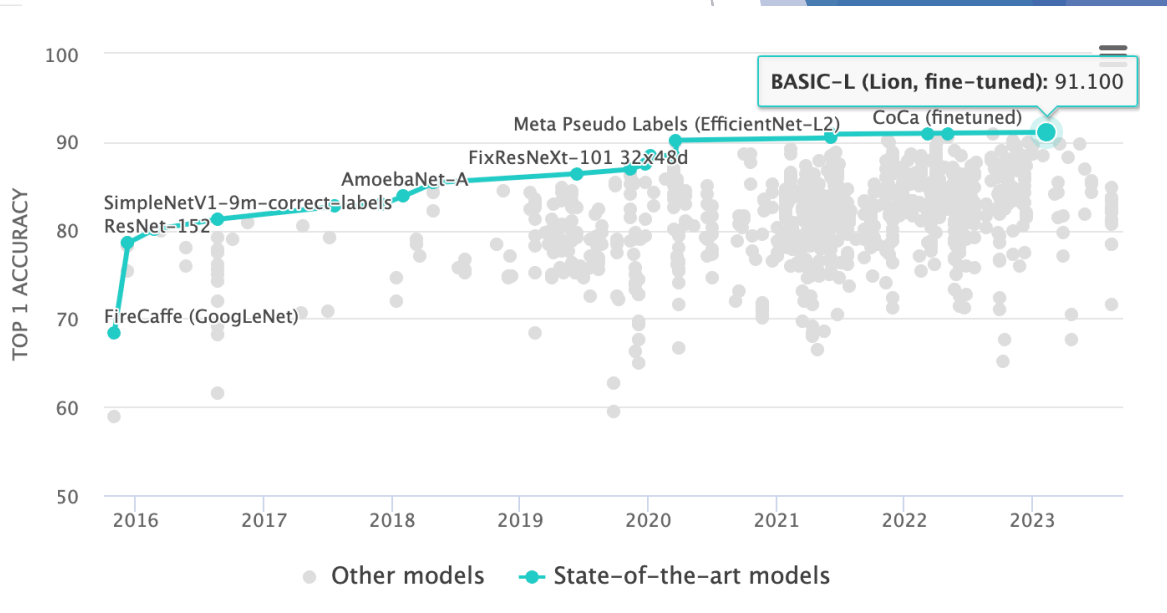
- Classification / Détection
  - Vision humaine : repérer un prédateur ou un membre de sa famille
  - Visio ordinateur : mettre un label sur une image
- **10 millions d'images avec 11000 classes, 1.4To**





En bleu, méthodes à base de réseaux





Concours ImageNet

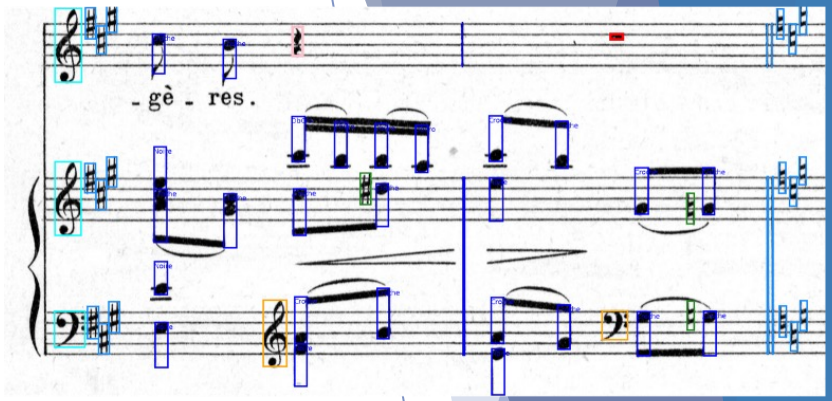
https://paperswithcode.com/sota/image-classification-on-imagenet

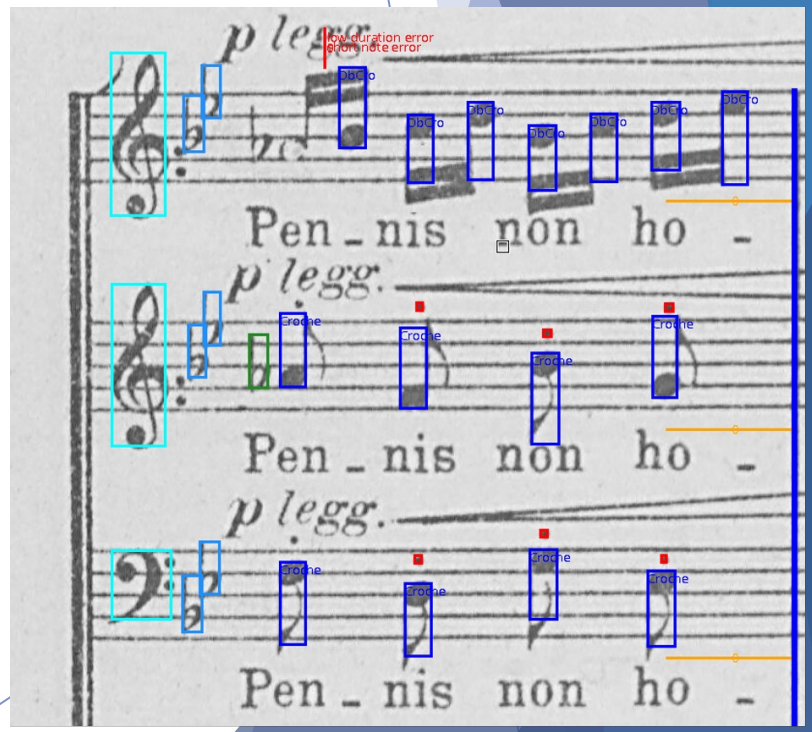
IA et musique

- 1950 - Article de Turing - citation de Jefferson (1949)
  - « Ce n'est que lorsqu'une machine sera capable d'écrire un sonnet ou de composer un concerto **en raison de pensées et d'émotions ressenties**, et non par la chute fortuite de symboles, que nous pourrons admettre que la machine équivaut au cerveau, c'est-à-dire non seulement qu'elle l'écrit, mais qu'elle sait qu'elle l'a écrit. Aucun mécanisme ne pourrait ressentir (et pas seulement signaler artificiellement, ce qui est facile) le plaisir de ses réussites, le chagrin lorsque ses soupapes éclatent, être réchauffé par la flatterie, être rendu malheureux par ses erreurs, être charmé par le sexe, être en colère ou déprimé lorsqu'il ne peut pas obtenir ce qu'il veut. »
- 1956 - **Illiac Suite**
  - La première composition musicale réalisée par une machine, inspirée du répertoire de Bach
- 2016 – **Aiva**
  - Le premier compositeur virtuel dont les œuvres sont reconnues par la Sacem
- 2017 - **« I Am AI »**
  - Le premier album de musique composé par une IA

IA et musique

- Système de reconnaissance de musique, ou Optical Music Recognition (OMR)
  - Nombreux outils commerciaux / grand public
  - Adaptés aux partitions « contemporaine »
- ANR CollabScore, un système de reconnaissance de partitions musicales
  - Automatise la reconnaissance du contenu de partitions de musique **ancienne**
  - Systèmes actuels utilisant le DL
    - Nécessite de grands corpus de données annotées pour leur entraînement
    - Peuvent produire des résultats parfois illogiques, ou sans cohérence globale
  - **Système hybride** de reconnaissance d'images de documents, combinant :
    - Un système basé sur du DL pour la détection de symboles musicaux isolés
    - L'expression de règles de la syntaxe musicale permettant de construire le contenu logique de la partition, en arrangeant les éléments musicaux

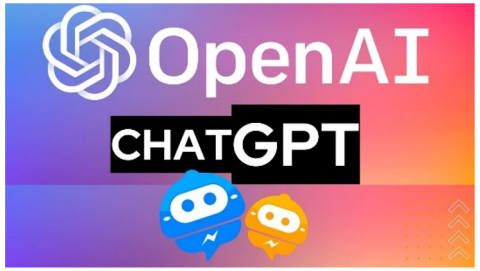


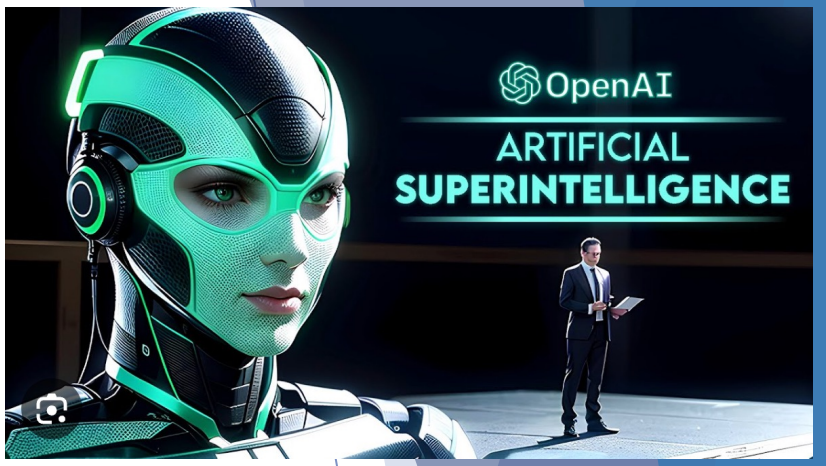


IA et photo

- Artiste allemand Boris Eldagsen
- Photo « The Electrician »
- A gagné la catégorie « Open » du Sony World Photography Awards, une prestigieuse compétition de photos qui rassemble de nombreux participants
- En 2023

OpenAI





- Avant 2019 : association à but non lucratif
- Depuis : « but lucratif plafonné »
- **« promouvoir et de développer un raisonnement artificiel à visage humain qui profitera à toute l'humanité »**
- 2021 : DALL-E, la première version d'un modèle capable de générer des images à partir de descriptions textuelles
- 2023 : chat-GPT 3, un modèle de langage entraîné sur de larges quantités de texte
  - Existe dans les laboratoires depuis 2018….
- 2023 : chap-GPT 4, texte et image
- Mai 2023, recommandations sur la gouvernance des **super-intelligences** avec une nouvelle prédiction :
  - *En moins de 10 ans, l'IA pourrait dépasser les compétences des meilleurs experts dans la plupart des domaines, ouvrant la voie à un futur « radicalement plus prospère » à condition de « gérer les risques pour y arriver »*

GPT : Generative Pre-trained Transformer i.e. un **modèle génératif**

- Que sont les modèles génératifs de texte ?
  - Objectif : prévoir et proposer les meilleurs mots et phrases suivant ceux proposés par un utilisateur
  - Fonctionnement
    - Construit par l'apprentissage d'associations de mots dans leur contexte (des phrases)
    - S’appuie sur des probabilités pour générer la suite la plus probable au regard des mots qui précèdent et des données d’apprentissage en rapport avec la question posée
      - Aléatoire permet des variations significatives dans les réponses lorsque des questions similaires sont posées
    - Apprentissage sur quantités de données gigantesques (internet)
    - Suivi d'une phase complémentaire pour affiner les "meilleures" réponses à proposer
    - Associé à un agent conversationnel

ChatGPT

- Entrainement
  - Entrainé avec 500 milliards de mots
  - 175 Milliards de paramètres
  - Cerveau humain : 100 milliards de neurones…
- Mise à jour : 2 bases de connaissances
  - Modèle de prédiction de texte associé au modèle encyclopédique
    - Données apprises jusqu'en 2021 : ce qui s'est passé depuis est inconnu
  - L'agent conversationnel continue à évoluer en continu grâce aux données des conversations avec les utilisateurs
    - Apprentissage par renforcement à partir de retour humain
    - Mémoire de 3000 mots pour la conversation
    - Continue de s'affiner en fonction du retour des utilisateurs toutes les 3-4 semaines en moyenne

ChatGPT, LLaMA, Falcon…

- Ce qui pose problème ?
  - Le Copyright, le droit d'auteurs
  - Le résultat aléatoire : impossible de détecter automatiquement qu'un texte a été produit par un modèle génératif
  - Le bilan écologique : nécessite des moyens considérables pour leur entraînement puis leur utilisation

- Est-ce vraiment de l'IA ?
  - Pas de mécanisme "intelligent" ou "conscient" doté de la capacité de raisonner
  - Manipulent des suites de mots, mais pas des assertions logiques ou des éléments de calcul
  - La notion de justesse ou de vérité du résultat proposé est totalement absente du mécanisme
  - L'esprit critique s'impose plus que jamais
- Mais il ne faut pas pour autant refuser la créativité que ces modèles peuvent stimuler…

IA et voiture autonome

- Niveau 0 – Pas d'autonomie, aucune assistance
  - Le conducteur réalise toutes les tâches de conduite
  - Le véhicule est uniquement équipé de **systèmes d'alerte**, différents des aides à la conduite (ex manque d'essence)
- Niveau 1 - Conduite assistée : « eyes on-hands on »
  - Le conducteur reste le pilote en toutes circonstances
  - Le véhicule est équipe de **systèmes d'assistance** à la conduite (régulateur de vitesse, freinage d'urgence…)
- Niveau 2 - Conduite partiellement automatisée : « eyes on-hands off »
  - Le conducteur peut lâcher temporairement le volant mais doit surveiller le système en permanence
  - Le véhicule gère la trajectoire **pendant un certain temps ou dans des situations spécifiques** (ex : dépassement sur autoroute)
- Niveau 3 - Conduite automatisée « eyes off-hands off »
  - Le véhicule gère la trajectoire automatiquement et surveille son environnement
  - Le véhicule **alerte** le conducteur **avec suffisamment d'anticipation** si une situation nécessite sa reprise en main
- Niveau 4 - Conduite hautement automatisée : « eyes off-hands off-mind off »
  - Le véhicule est capable de maîtriser automatiquement toutes les situations de déplacement **dans des conditions d'utilisation définies**
  - Le conducteur doit pouvoir reprendre les commandes de la voiture si celle-ci l'estime nécessaire
- Niveau 5 - Conduite totalement automatisée : « driverless »
  - Le véhicule est totalement automatisé sur tous types de routes et de trajets du départ à l'arrivée, il n'y a plus d'intervention humaine.
  - La présence de volant, de pédales et d'un conducteur n'est plus nécessaire

IA et voiture autonome

- Actuellement, surtout des voitures de niveau 2 (conduite partiellement automatisée)
- Depuis sept. 2022, les voitures de niveau 3 (conduite automatisée « eyes off-hands off ») autorisées en France
- Seul Mercedes a reçu une approbation internationale pour la conduite autonome de niveau 3
- Comment ?
  - Conduite autonome que dans des zones délimitées : autoroutes ou voies séparées par un terre-plein central ne pouvant pas être empruntées par des piétons ou des cyclistes
  - Dans des conditions clairement réglementées : vitesse maximale de 60 km/h, ne fonctionne pas en cas de pluie, ni la nuit ou dans le cas où la température extérieure descend sous les quatre degrés
  - Le système Drive Pilot de Mercedes n'autorise aucun changement de voie, n'est plus actif en cas de chantiers, lors de traversées de tunnels ou de passages souterrains de plus de 50 mètres
  - Le conducteur doit rester sur son siège et garder son visage visible pour les caméras embarquées du véhicule à tout moment

IA et voiture autonome

- Pourquoi ?
  - Tout est un problème de responsabilité…
  - Pour les modes de conduite autonome de niveau 1 ou 2, le conducteur reste responsable
  - À partir du niveau 3, c'est le constructeur qui endosse la responsabilité en cas d'accident
  - Les données relevées par le système de mémoire du véhicule jouent alors un rôle essentiel pour en déterminer les causes de l'accident
  - A ce niveau d'autonomie, le conducteur doit pouvoir reprendre les commandes à tout moment sur « demande » de la voiture
  - Une fois alerté, c'est donc lui qui endosse à nouveau la responsabilité du comportement de son véhicule

La pluridisciplinarité plus que jamais…

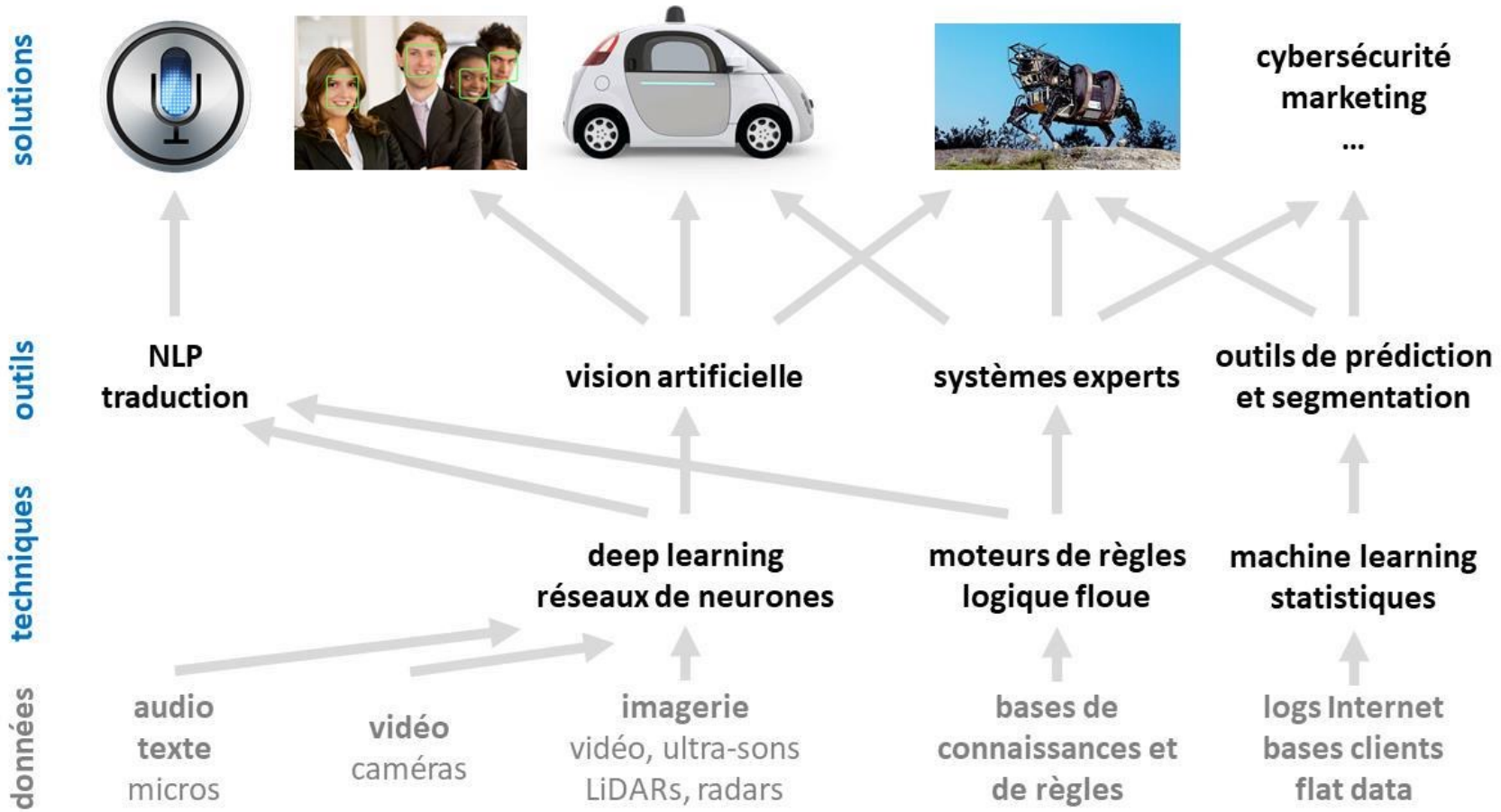
- Apprentissage automatique **pour biologie**
  - Depuis 20 ans, les techniques de séquençage permettent d'étudier l'information biologique contenue dans notre ADN
  - ML pour mieux comprendre les liens biologiques qui régissent le comportement de nos cellules
- Apprentissage automatique **inspirées par physique**
  - Développement de techniques de ML inspirées par la physique pour la super-résolution d'images de microscopie à fluorescence
  - Les réseaux de neurones doivent prendre en compte des phénomènes comme la dégradation du signal lors de l'acquisition d'une image
  - Les données sont en quantité limitée : comment fournir des exemples de ce que l'on n'a pas encore observé ?
- IA générative pour la santé
  - Des premiers travaux sur le résumé automatique de réunions et l'analyse des appels d'urgence du SAMU

Centre AISSAI lancé en 2022 par le CNRS : AI for science, Science for AI

- **L'IA impacte toutes les disciplines scientifiques**
- « *Nous avons besoin de plus de personnes à l'interface entre les disciplines, car elles permettent à des communautés de se comprendre et d'enrichir les recherches de chacun.* » Laura Cantini (chargé de recherche CNRS en biologie)
- « *L'IA est devenue un moyen disruptif dans la conduite de la science et l'accélération de découverte scientifique en transformant les méthodes scientifiques, en transcendant les frontières entre disciplines et en faisant émerger de nouvelles approches interdisciplinaires* » Jalal Fadili (Junior member, Institut Universitaire de France)

- Centre AISSAI aide aux grandes infrastructures de recherche et plateformes expérimentales nationales et internationales
- Dans toutes les disciplines
  - Biologie : plateforme Flagship Human Brain
  - Physique des hautes énergies : Large Hadron Collider
  - Etude de la biodiversité : Pôle National de Données de Biodiversité

Mais le Deep Learning ne fait pas tout…



IA et humour, par exemple…

http://karpathy.github.io/

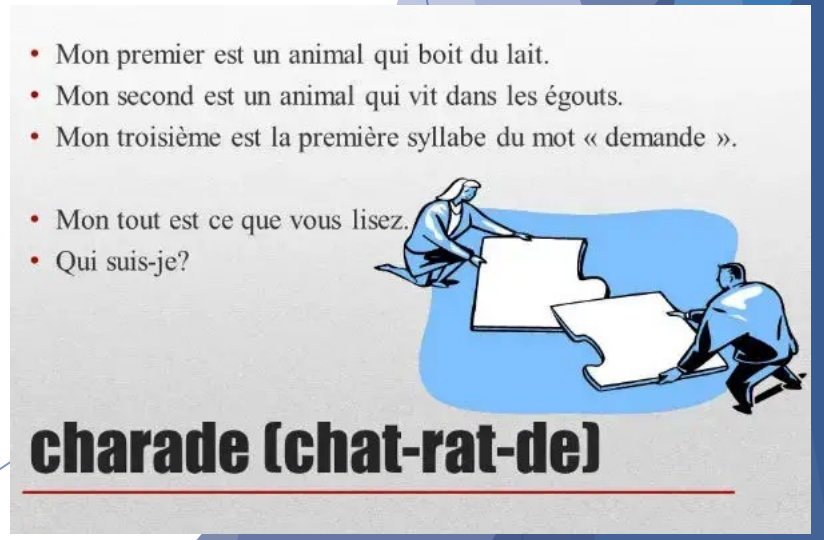
- «The State of Computer Vision and AI: We Are Really, Really Far Away »
- Que voit les humains ?
  - Le sourire du président dans le miroir
  - Le pied du président sur la balance
  - ….
- Nous autres humains trouvons cette photo pleine d'humour
- Karpachy : « Que faudrait-il pour qu'un ordinateur comprenne cette image telle que vous ou moi la comprenons? »

IA et les jeux

- L'IA sait jouer (et gagner) :
  - Aux dames, aux échecs, à Jeopardy, au jeu de Go
  - Au 57 jeux sortis sur Atari 2600, même les plus complexes (ping, casse-brique…)

- Mais …. que se passe-t-il quand on change les paramètres du jeu ?
  - Est-ce que les réseaux de neurones savent encore jouer ?
  - Par exemple pour des jeux simples comme le casse-brique ?

- Mais …. il existe de nombreux jeux humains qui sont encore plus difficiles pour l'IA. Par exemple ?
  - Les charades….
  - Elles exigent une compréhension visuelle, linguistique et sociale !



L’IA aujourd’hui : IA forte ?

- Pour certains, IA devient un outil de tous les jours
  - Reconnaissance, recommandation, guidage, prise de décision, programmation, résolution d’examens ;-) …
- Pour d’autres, l'objectif initial consistant à développer des machines aussi intelligentes que l’être humain continue à perdurer
  - Plusieurs enquêtes* auprès de praticiens de l'IA leur demandant quand arriverait l'IA générale ou l'IA « superintelligente » ont donné un large éventail d'opinions
  - Allant de « lors des dix prochaines années » à « jamais "
  - Autrement dit, nous n'en avons pas la moindre idée !!!

* O. Etzioni, « No, the Experts Don't Think Superintelligent Al Is a Threat to Humanity», Technology Review, 20 septembre 2016, www.technologyre-view.com/s/602410/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity

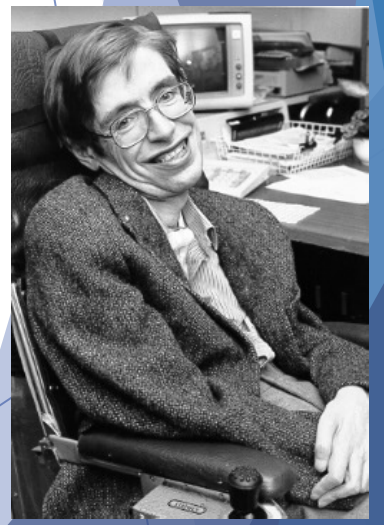
* V. C. Müller et N. Bostrom, « Future Progress in Artificial Intelligence: A Survey of Expert Opinion », dans Fundamental Issues of Artificial Intelligence, Bale (Suisse), Springer, 2016,
p. 555-572.

L’IA aujourd’hui : IA forte ?

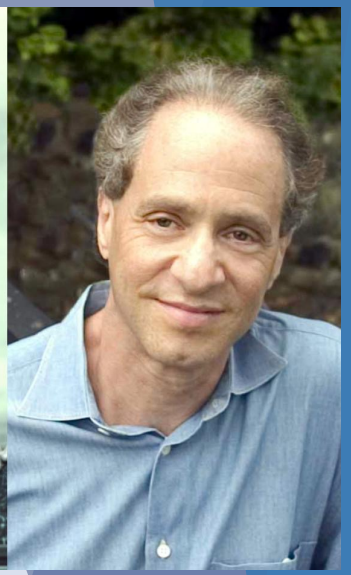
- Ce que nous savons, c'est que l'IA de niveau humain exigera des aptitudes
  que les chercheurs en IA s'efforcent depuis des décennies de comprendre et de reproduire
  - Notamment le sens commun, l’abstraction et l'analogie
- Mais ces aptitudes se sont révélées profondément difficiles à comprendre et donc à reproduire
- **Il est donc nécessaire d'en savoir plus sur l'intelligence humaine…**
- D'autres grandes questions demeurent :
  - L’IA générale exigera-t-elle de la conscience ?
  - Exigera-t-elle d'avoir un sentiment de soi ?
  - De ressentir des émotions ?
  - De posséder un instinct de survie et la peur de la mort?
  - D'avoir un corps ?
- Avant… on parlait d’**IA forte**
- Maintenant on entend parler de **singularité**, de **transhumanisme**, de **super-intelligence**

La singularité

- En 1965, première spéculation sur les effets des machines plus intelligentes que les hommes :
  - « Soit une machine ultra-intelligente définie comme une machine qui peut largement dépasser toutes les activités intellectuelles d'un homme si habile soit-il. Comme la conception de machines est l'une de ses activités intellectuelles, une machine ultra-intelligente pourrait concevoir des machines encore plus poussées : il y aurait alors incontestablement une « **explosion de l'intelligence** », et l'intelligence de l'homme serait laissée loin derrière. Ainsi, la première machine ultra-intelligente sera la dernière invention que l'Homme doive jamais faire. »
- La singularité :
  - « Le terme ***Singularité*** appliqué à des machines intelligentes se réfère à l'idée que lorsque des machines intelligentes pourront concevoir des machines intelligentes, plus intelligentes qu'elles, cela provoquera une croissance exponentielle de l'intelligence des machines conduisant à une singularité de l'intelligence à l'infini (ou du moins immense). » (Hawkins)



Le transhumanisme



- Idéologie défendue par Ray Kurzweil
- Repose sur :
  - Critique de la singularité technologique
    - Une promesse illusoire qui prédit une transformation profonde et radicale des sociétés humaines grâce au développement surprenant de l'intelligence artificielle (IA)
  - Progrès de l'IA, des nanotechnologies, des biotechnologies, de l'informatique et des sciences cognitives
- Principe :
  - Fusion de l'homme et de la machine, **union du biologique à la technologie**
  - Permettra de résoudre les problèmes humains les plus complexes (moraux, culturels, économiques, politiques, etc.) et même d'éradiquer la mort !!!
  - Naissance d'une nouvelle humanité qui bénéficierait des capacités analytiques d'un superordinateur et qui serait débarrassée des inconvénients du corps biologique

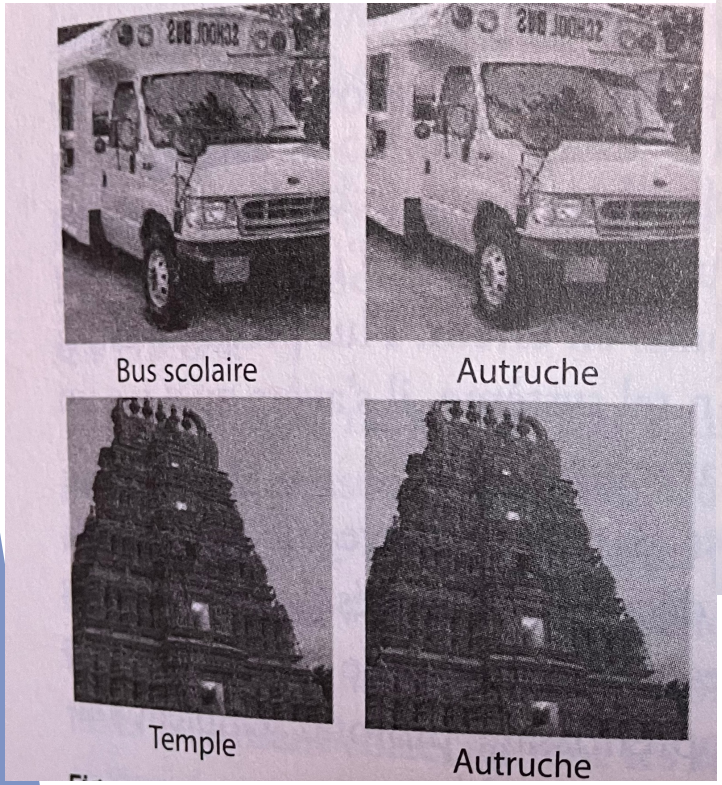
Alors que reste-t-il à faire ? Mais tout !!!

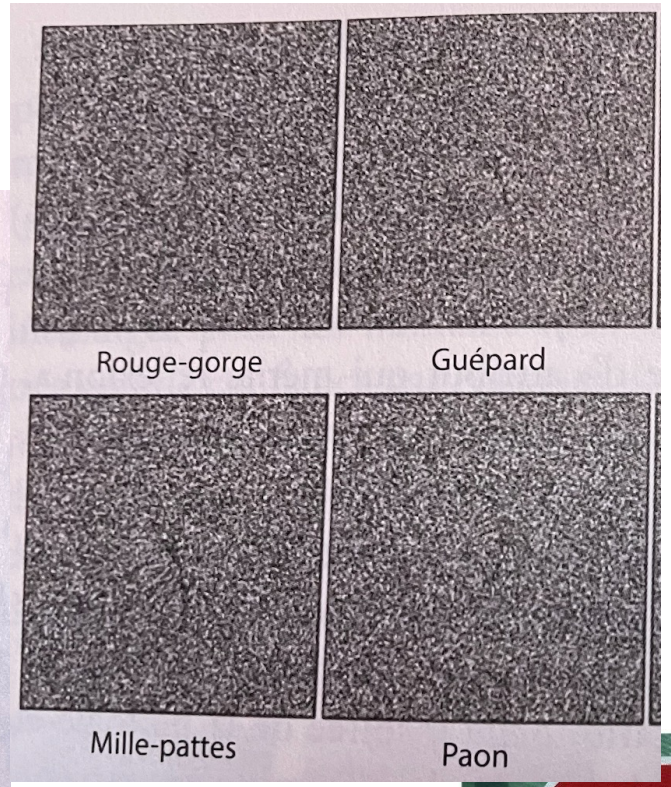
- Toutes les questions importantes posées au début de la discipline restent ouvertes
- La proposition émise par John McCarthy & Co lors de la création du domaine énumérait quelques-uns des principaux sujets de recherche en IA :
  - le traitement du langage naturel, les réseaux neuronaux, l'apprentissage machine
  - les concepts abstraits et le raisonnement, et la créativité
- En 2015, Eric Horvitz, directeur de recherche chez Microsoft
  - « l'on pourrait même dire que la proposition de 1955, correctement reformatée, pourrait être resoumise aujourd'hui à la National Science Foundation et recevoir probablement des subventions de quelques gestionnaires de programmes enthousiastes. »

Améliorer le ML / DL

- Les algos d'apprentissage automatique ont des difficultés lorsqu'ils effectuent leurs tâches sur des **données trop éloignées** de celles sur lesquelles ils se sont entraînés
- Ils sont particulièrement **vulnérables aux fausses associations statistiques**
- Quand un modèle s'est entraîné sur un certain corpus, comme Wikipédia ou des articles de presse, il va avoir du mal à travailler s'il doit manipuler des types de textes très différents, comme du chat ou des tweets
- Les algorithmes conversationnels subissent parfois des associations statistiques trompeuses, qui fonctionnent dans des contextes précis, mais pas en dehors
- **Comme ces modèles sont des boîtes noires, qui opèrent des choix selon des critères difficilement compréhensibles par les humains, comment les aider à distinguer les corrélations à risque des causalités réelles ?**

Les motifs adverses en DL





Améliorer l'IA générative

- **Comment mettre à jour automatiquement ces modèles LLM ?**
  - Il ne suffit pas de continuer à former un LLM avec de nouvelles données
  - Sans des processus d'apprentissage adéquats, ces modèles risquent d' « oublier » des informations précédentes
- **Comment rendre ces modèles plus accessibles en termes de coûts de calcul ?**
  - Comment équilibrer vitesse et mémoire ?
  - Quelles méthodes collaboratives pour accélérer les calculs ?
  - Comment proposer des applications concrètes en langue française ?

Mais surtout : FATE in AI …
Fairness, Accountability, Transparency and Ethics

- **Fairness** / Equité
  - Empêcher que des groupes sociodémographiques soient désavantagés par des algorithmes
- **Accountability** / Responsable
  - Garantir le bon fonctionnement, tout au long du cycle de vie : conception, développement, exploitation, déploiement
  - Conformément aux cadres réglementaires applicables, que ceux qui propose des IA puissent en faire la démonstration
- **Transparency** / Transparence, explicabilité
  - ML-DL sont des boites noires … donc problèmes : pouvoir explicatif, contrôlabilité
  - L'acceptation passe par la confiance et la compréhension
- **Ethics** / Ethique
  - Manque de « bon sens », dérapage des IA
- **Frugale**
  - Trouver un équilibre entre la performance des modèles d'IA et les contraintes de ressources
  - Penser des datacenters moins énergivores, réduire les jeux de données nécessaire….

Ethique

- Dilemme du tramway
  - En 2016, plusieurs enquêtes réalisées auprès de plusieurs centaines de personnes
  - Plusieurs scénarios mettant en scène des voitures autonomes pour questionner les points de vue sur la moralité des différentes actions
  - Un résultat intéressant :
    - 76% des personnes interrogées répondirent qu'il serait moralement préférable qu'une voiture autonome sacrifie un passager pour sauver dix piétons
    - Mais lorsqu'on leur demanda si elles achèteraient une voiture autonome programmée pour sacrifier ses passagers afin de sauver un bien plus grand nombre de piétons, l'immense majorité des personnes interrogées répondit que non…
- Avant de pouvoir intégrer nos valeurs dans des machines, nous devons donner plus de clarté et de cohérence à ces valeurs
- Mais cela semble plus difficile que prévu :-/
- Certains spécialistes de l'éthique de l'IA ont proposé de renoncer à programmer directement des règles morales dans les machines

Fairness ou l'équité et les biais de l'IA générative

- L'IA générative peut trouver le dernier mot sur des vecteurs-mots.
- Plusieurs groupes ont montré que ces vecteurs-mots expriment les biais inhérents aux données linguistiques qui les génèrent
- Un exemple :
  - L'analogie suivante : « Homme est à femme ce que programmeur est à _. »
  - Si vous résolvez ce problème en utilisant les vecteurs-mots fournis par Google, la réponse est *ménagère*.
  - Le problème inverse, « Femme est à homme ce que programmeur est à _. » donne *ingénieur mécanicien*.
- Autre exemple :
  - « Homme est à génie ce que femme est à _. » Réponse : *muse*.
  - « Femme est à génie ce que homme est à _. » Réponse : *génies*.
- Ce n'est pas la faute de l'IA …. Elle ne fait qu'exprimer le sexisme et autres préjugés de notre langue
- Et notre langue reflète les préjugés de notre société
- Mais aussi irréprochables soient-ils, les vecteurs-mots sont une composante clé de chaque système de TALN, allant de la reconnaissance de la parole à la traduction…

Les biais algorithmiques

- Les biais peuvent essentiellement advenir de deux manières
  - Si le codeur, souvent homme, blanc et cisgenre, laisse transparaitre ses propres préjugés dans les paramètres qu'il établit pour une IA
  - Si un algorithme de Machine Learning est entrainé sur un échantillon d'exemples lui-même biaisé
- Les premiers programmes de reconnaissance faciale étaient incapables d'identifier les individus à peau noire
  - A la fois parce que les critères explicites ne prenaient pas en compte les contrastes de couleur
  - Et parce que les images fournies à la machine représentaient essentiellement des blancs
- Mais des programmes plus récents ont toujours ce genre de « soucis »…
  - Un groupe de recherche a remarqué que son système d'IA - entraîné sur un grand ensemble de photos de personnes dans diverses situations
  - Classait parfois un homme dans la catégorie « femmes » lorsqu'il se tenait dans une cuisine…
  - Environnement dans lequel les données d'entraînement avaient plus d'exemples de femmes

Les biais algorithmiques

- Ce type de biais portent notamment sur les groupes raciaux ou genrés
- **Ce type de biais subtil n'est perceptible qu'*a posteriori***
- **Il est difficile à anticiper :** souvent difficile de déterminer la nature et les effets des biais lorsqu'ils sont subtils
- Pour les corriger / les anticiper, cela exige une prise de conscience et un effort de la part des humains qui préparent et sélectionnent les données
- Mais il faudrait d'abord se mettre d'accord :
  - Les ensembles de données utilisés lors des phases d'apprentissage doivent-ils refléter les préjugés de notre société - comme c'est souvent le cas aujourd'hui
  - Ou doivent-ils être repensés pour atteindre des objectifs de réforme sociale ?
  - Et qui devrait définir ces objectifs ou repenser les données ?

Les biais humains : par exemple le biais de représentativité

- Histoire de Gaspard Koenig inspirée de celles du sociologue Kahmeman
  - *Linda est une étudiante en sociologie révoltée contre les injustices et militante antispéciste*
  - *Elle a participé l'occupation de Jussieu lors des grèves du printemps 2018.*
  - *Dix ans plus tard, est-il plus probable qu'elle devienne :*
    - *assistante maternelle ;*
    - *prof de yoga ;*
    - *banquière ;*
    - *travailleuse sociale ;*
    - *banquière militant pour les droits des animaux ?*
  - Au vu de la personnalité supposée de Linda, la plupart des sujets interrogés estiment plus probable qu'elle soit « banquière militant pour les droits des animaux » plutôt que simplement « banquière »…
  - Or les probabilités disent le contraire, la proposition 5 n'étant qu'une sous-catégorie de la prop. 3
- Cette erreur commune illustre le biais de « représentativité »
  où le goût du stéréotype prime sur le raisonnement statistique le plus basique

Explicabilité ou XAI (eXplainable AI)

- Pensez à nous, vos profs, quand nous écrivons :
  « Où sont vos calculs ? » ou « Montrez vos calculs »
  - Pour savoir si vous avez compris ce que vous faites, nous voulons voir comme vous avez trouvé la solution
- Avec l'IA c'est pareil, on veut voir… donc elle doit pouvoir nous expliquer !
  - C'est assez « facile » avec l'IA symbolique puisque l'on peut voir le raisonnement, mais la difficulté est de le « monter » aux utilisateurs de l'IA, pour qu'ils aient confiance
  - C'est beaucoup moins facile avec l'IA numérique… la boite noire ne nous éclaire pas beaucoup…
- De nombreux travaux portent sur l'IA explicable (XAI)
  - Attention, pas sur xAI, l'entreprise créée par Elon Musk en juillet 2023 pour concurrencer OpenAI,
    et dont le but est de « comprendre la véritable nature de l'univers »….

Explicabilité ou XAI (eXplainable AI)

- Beaucoup de travaux en XAI pour permettre d'expliquer aux non-informaticiens
- Beaucoup de travaux en XAI pour comprendre à quoi servent les différents composants des boites noires
  - Par exemple : en désactivant ou perturbant certains composants pour comprendre leur rôle dans le fonctionnement global du modèle (*ablation studies*)
  - « C'est comme lorsque, en neurosciences, on profite d'une lésion du cerveau pour comprendre à quoi sert la partie en question. On perturbe et on obfusque certaines parties du modèle pour regarder l'impact de ces sous-ensembles sur les résultats. » (Maxime Peyrard, chair junior en IA)
- Mais aussi avec l'idée d'une **intelligence artificielle neuro-symbolique**
  - Méthode qui combine les réseaux neuronaux, la représentation symbolique des problèmes et de la logique (lisibles par les humains)
  - Pour beaucoup, l'IA neuro-symbolique est considérée comme la prochaine évolution de l'intelligence artificielle….

Et la compréhension dans tout ça ?

- Les systèmes de TAL sont de plus en plus performants, notamment grâce à l'IA générative
- **Mais que comprennent-ils dans ce qu'ils lisent ?**
- Preuve par l'exemple : que signifie le « il » ?
  - Phrase 1: Les dirigeants de la ville refusèrent l'autorisation de manifester aux contestataires parce qu'ils craignaient des violences.
  - Question : Qui craignait des violences? A. Les dirigeants / B. Les contestataires.
  - Phrase 2: « Les dirigeants de la ville refusèrent l'autorisation de manifester aux contestataires parce qu'ils prônaient la violence.
  - Question : Qui prônait la violence ? A. Les dirigeants / B. Les contestataires.
- Les phrases 1 et 2 ne diffèrent que par un seul mot (craignaient / prônaient), mais ce seul mot détermine la réponse à la question….
- Comment nous autres humains savons-nous cela?
- En nous appuyant sur **nos connaissances de base sur le fonctionnement de la société** : nous savons que les contestataires sont ceux qui ont des revendications et qu'ils préconisent ou incitent parfois à la violence lors d'une manifestation.
- Autre exemple :
  - Phrase 1 : « L'oncle de Joe peut encore le battre au tennis, bien qu'il soit plus âgé de 30 ans. »
  - Question : Qui est plus âgé ? A. Joe / B. L'oncle de Joe
  - Phrase 1 : « L'oncle de Joe peut encore le battre au tennis, bien qu'il soit plus jeune de 30 ans. »
  - Question : Qui est plus jeune ? A. Joe / B. L'oncle de Joe
- Plein d'autres exemple dans le livre de M. Mitchell….

Les schémas de Winograd

- Pour répondre correctement aux questions, une machine doit être capable non seulement de traiter les phrases, mais aussi de les comprendre, du moins jusqu'à un certain point.
- En général, la compréhension de ces phrases exige ce que nous pourrions appeler **un savoir de bon sens**
- Par exemple, un oncle est habituellement plus âgé que son neveu.
- Ces mini-tests de compréhension du langage s'appellent des schémas de Winograd, d'après le pionnier de la recherche en TALN, Terry Winograd, qui fut le premier à en avoir eu l'idée
- Les schémas de Winograd sont précisément conçus pour être facilement compréhensibles par les humains mais problématiques pour les ordinateurs

Sens commun

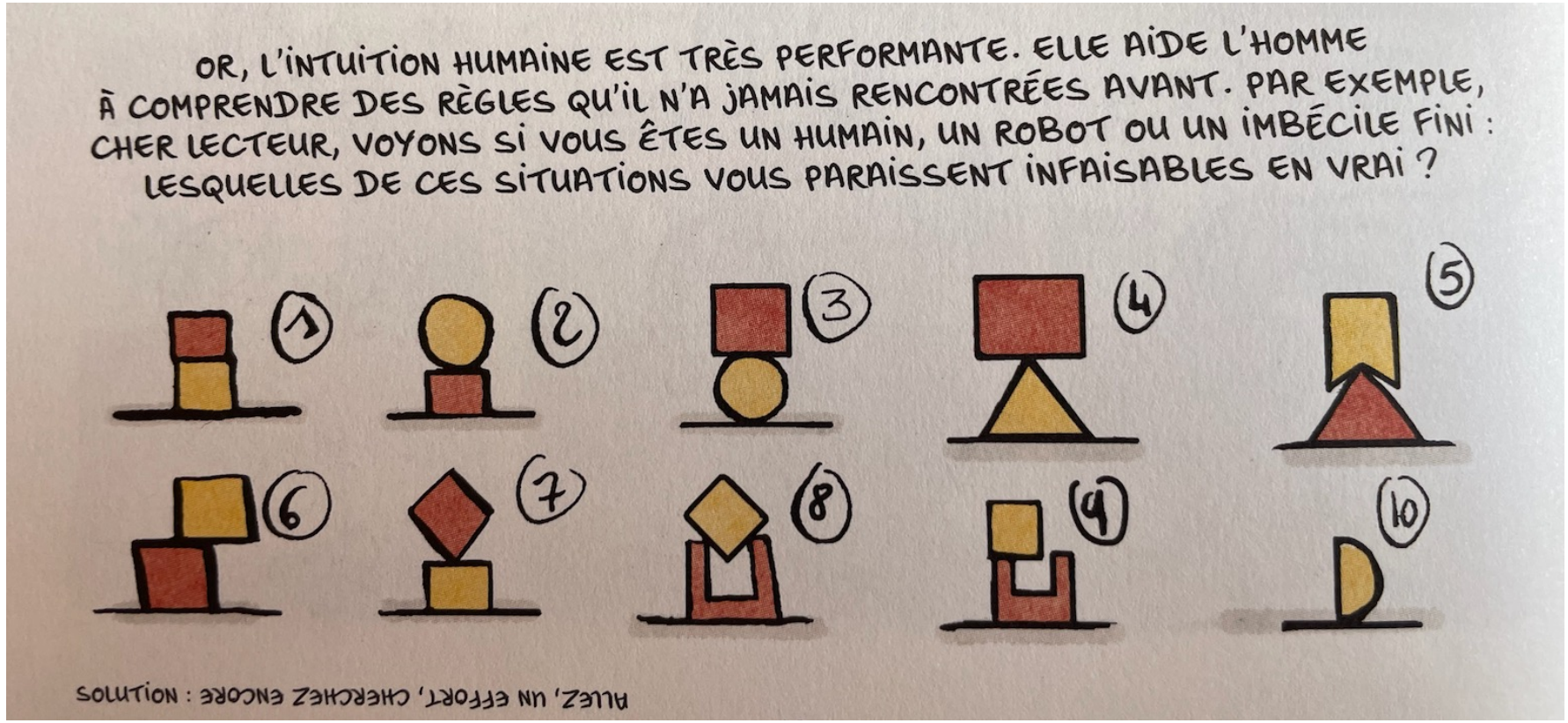


Illustration tirée de la BD *L'Intelligence artificielle*, de Jean-Noël Lafargue et Marion Montaigne

Sens commun

- **Pourquoi certains incidents de l'IA nous paraissent si aberrants ?**
- Par exemple, les GPS qui conseillent à des piétons de traverser des autoroutes ou d'aller en plein désert australien
- L'IA est en mesure d'éviter des milliers d'erreurs humaines
  mais peut commettre des erreurs inimaginables pour un humain
- « On ne va pas traverser une autoroute » : ce dont le sens commun nous instruit immédiatement, l'IA a besoin de l'apprendre
- Si la notion d'autoroute ne fait pas partie de ses paramètres d'entraînement, il n'y a aucune raison qu'elle s'oppose à ce qu'un piéton marche sur une étendue d'asphalte
- Dès qu'une situation diverge de la normalité, l'IA est perdue…

Sens commun

- Le sens commun est l'ensemble des connaissances et des croyances partagées par une communauté
et jugées prudentes, logiques ou valides
- Il représente la **capacité naturelle à juger les évènements de façon raisonnable**

- Le problème de l'IA c'est qu'elle est dépourvue de sens communs
et donc de bon sens et d'humour… ;-)
- C'est-à-dire qu'elle ne dispose pas d'un modèle de représentation du monde indépendant d'une tâche donnée et capable de gérer l'incertitude, la nouveauté, l'imprévu
- Alors pourquoi ne pas lui apprendre ?

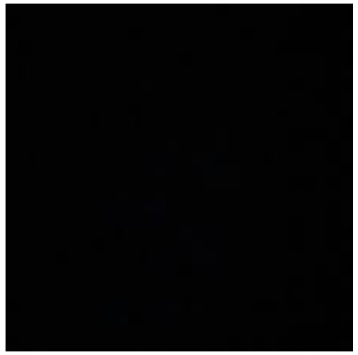
Sens commun et contexte

- Le feu ça brule, l'eau ça mouille

- Si on scinde un tas de sable en deux, on obtient 2 tas de sable
  mais avec un stylo ?

- On peut voir le nez d'une personne mais pas son cœur
- Vous ne pouvez rien voir

?



Le sens commun en philosophie

- Descartes avait imaginé l'existence d'une glande, la glande pinéale, chargée d'opérer la relation entre les pensées, les connaissances, les informations et l'action.
  - La glande pinéale (seule partie non double du cerveau) assure l'union de l'âme et du corps
  - La glande pinéale interdit une approche intégralement mécaniste de la biologie.
  - Elle permet de produire le sens commun qui distingue le cerveau d'un processus mécanique
  - Le sens commun se loge quelque part entre la logique de l'esprit et la mécanique du corps

- **Pour Descartes, le sens commun est la chose du monde la mieux partagée...**
  et la moins remplaçable

- Et le lien avec l'IA ? L'esprit humain n'est pas seulement une calculette ni le corps une horloge ; inversement, la combinaison d'une IA et d'un humanoïde ne pourra jamais se substituer à la capacité de jugement d'un être humain

Le sens commun en IA

- Le projet CYC
  - Lancé en 1984 par Douglas B. Lenat
  - Nom dérivé de « encyclopédie »
  - Objectif : développer une ontologie globale et une base de connaissance générale
  - Pour permettre à des IA de raisonner d'une manière similaire à l'être humain
  - Et donc de représenter le sens commun
  - Equivalent à un système expert avec comme domaine d'application les objets et actions du quotidien

Le sens commun en IA

- Principe de CYC :
  - La philosophie sous-jacente a beaucoup en commun avec celle des systèmes experts, comme MYCIN
  - Logique du 1er ordre
  - Les assertions sont soit vraies, soit fausses, sans facteur de confiance
- Les algorithmes de raisonnement logique de CYC sont plus sophistiqués par exemple que ceux de MYCIN, mais ces deux projets partagent une même conviction fondamentale :
  on peut reproduire l'intelligence *via* des règles programmées explicitement et opérant sur un ensemble suffisamment vaste de connaissances clairement formulées

Le sens commun en IA

- Combien d'assertions sont-elles nécessaires pour rendre compte du bon sens humain ?
- En 2015, Lenat estimait à **quinze millions** le nombre d'assertions contenues dans CYC : « Nous disposons probablement d'environ **5 % de ce qu'il nous faudra en définitive.** »
- Les assertions de CYC ont été codées à la main par des humains (les employés de Cycorp) ou déduites de manière logique par le système à partir des assertions existantes
- CYC possède de nombreuses méthodes pour gérer les assertions incohérentes ou incertaines présentes dans sa bibliothèque
- CYC est capable maintenant d'automatiser l'obtention de nouvelles affirmations en exploitant le Web

Le sens commun en IA

- La base de connaissances originale est la propriété de Cycorp
- Version limitée publiée sous licence Open Source sous le nom d'OpenCyc
  - En 2002 : OpenCyc 1.0 : 6k concepts, 50k assertions
  - En 2012 : OpenCyc 4.0 : 239k concepts, 2 093k assertions
  - En 10, x 35….
- Version mise à la disposition de chercheurs en IA sous une licence destinée à la recherche, sous le nom de ResearchCyc.
- En 2023, toujours actif et revendique de proposer « ***Logic-based machine reasoning*** »
- Dans le paysage actuel de IA, dominé par l'apprentissage profond, le projet CYC est l'une des dernières tentatives basées sur l'IA symbolique à grande échelle

Le sens commun en IA

- **Le projet CYC a été décrit comme « l'une des entreprises les plus controversées de l'histoire de l'intelligence artificielle »** (Bertino et al), et les critiques n'ont pas manqué….
- Se peut-il qu'avec suffisamment de temps et d'efforts, on parvienne à simuler la totalité de la connaissance de bon sens humain, voire simplement une portion suffisante de cette connaissance ?
- Problème n°1 :
  - Si la connaissance de bon sens est la connaissance dont tous les humains sont dotés mais qui n'existe pas sous forme écrite, alors la majeure partie de cette connaissance est subconsciente ; nous ignorons en être dotés
  - Elle inclut une grande part de nos connaissances essentielles intuitives de la physique, de la biologie et de la psychologie, qui forme la base de nos plus vastes connaissances du monde
  - Si vous ignorez savoir quelque chose, vous ne pouvez être l'« expert » qui fournit explicitement ce savoir
- Problème 2 :
  - Notre connaissance de bon sens est régie par l'abstraction et l'analogie. Ce que nous appelons sens commun ne peut exister sans ces capacités
  - Toutefois, l'abstraction et l'analogie humaines ne sont pas des techniques susceptibles d'être simulées par l'immense bibliothèque de CYC

Le sens commun en IA

- Et pourtant, créé en 1984, le projet CYC en est à sa quatrième décennie d'existence
  - Cycorp et son entreprise dérivée, Lucid, commercialisent CYC en offrant un éventail d'applications destinées au monde des affaires, et qui sont des réussites dans la finance, l'extraction de gaz et de pétrole, la médecine…
- La trajectoire de CYC fait écho à celle de Watson d'IBM
  - Deux programmes débutés par un effort de recherche fondamentale en IA, d'une portée et d'une ambition considérables
  - Pour aboutir à un ensemble de produits commerciaux aux prétentions marketing élevées (du genre : CYC « met la compréhension et le raisonnement humains à portée des ordinateurs ») mais avec des objectifs limités plutôt que généraux, et peu de transparence au niveau des réelles performances et capacités du système.
- Côté recherche scientifique, CYC n'a pas eu de véritable impact et certains chercheurs ont sévèrement critiqué cette approche
  - Selon Pedro Domingos (université de Washington), CYC est « l'échec le plus notoire de l'histoire de l'IA. »
  - Selon Rodney Brooks (MIT), « Si [CYC] a été un effort héroïque, il n'a pas conduit à un système d'IA capable de livrer ne serait-ce qu'une compréhension simple du monde. »

Accountability : réglementer l'IA ?

- De nombreux praticiens de l'IA, sont favorables à une certaine réglementation
- Mais cette réglementation ne doit pas être conçue uniquement par les entreprises et chercheurs travaillant sur l'IA
- Les problèmes liés l'IA - la fiabilité, l'explicabilité, les préjugés, la vulnérabilité aux attaques et la moralité de son utilisation - sont tout aussi sociaux et politiques que techniques
- Mais réglementer l'IA est complexe…
- En 2018, début d'une réglementation européenne sur l'IA
  - Certains l'ont qualifiée de « droit à l'explication »
  - Elle exige, dans le cas « d'une prise de décision automatisée », « des informations sensées concernant la logique sous-jacente. »

IA Act

- Loi sur l'IA de l'UE : première réglementation de l'intelligence artificielle
- Début 2021, validé par l'UE en juin 2023, il faut maintenant « discuter » avec les états….
- Des règles différentes pour différents niveaux de risque

- **Risque inacceptable** : systèmes d'IA considérés comme une menace pour les personnes. Ils seront interdits :
  - La **manipulation cognitivo-comportementale** de personnes ou de groupes vulnérables spécifiques : par exemple, des jouets activés par la voix qui encouragent les comportements dangereux chez les enfants
  - Un **score social** : classer les personnes en fonction de leur comportement, de leur statut socio-économique, de leurs caractéristiques personnelles
    - Interdit en Europe mais existe déjà en Chine et prend de l'importance…
  - Des systèmes d'identification biométrique **en temps réel et à distance**, tels que la reconnaissance faciale

IA Act

- **Risque élevé** : systèmes d'IA qui ont un impact négatif sur la sécurité ou les droits fondamentaux. Ils seront divisés en deux catégories :
  - Les systèmes d'IA qui sont utilisés dans les produits relevant de la législation de l'UE sur la sécurité des produits : les jouets, l'aviation, les voitures, les dispositifs médicaux et les ascenseurs
  - Les systèmes d'IA relevant de 8 domaines spécifiques devront être enregistrés dans une BD de l'UE :
    - l'identification biométrique et la catégorisation des personnes physiques
    - la gestion et l'exploitation des infrastructures critiques
    - l'éducation et la formation professionnelle
    - l'emploi, la gestion des travailleurs et l'accès au travail indépendant
    - l'accès et la jouissance des services privés essentiels et des services et avantages publics
    - les forces de l'ordre
    - la gestion de la migration, de l'asile et du contrôle des frontières
    - l'aide à l'interprétation juridique et à l'application de la loi
- Tous les systèmes d'IA à risque élevé seront évalués avant leur mise sur le marché et tout au long de leur cycle de vie

IA Act

- IA générative, comme ChatGPT, devrait se conformer aux exigences de transparence :
  - indiquer que le contenu a été généré par l'IA
  - concevoir le modèle pour l'empêcher de générer du contenu illégal
  - publier des résumés des données protégées par le droit d'auteur utilisées pour la formation

- Risque limité :
  - Les « autres » systèmes d'IA doivent respecter des exigences de transparence minimales qui permettraient aux utilisateurs de prendre des décisions éclairées.
  - Après avoir interagi avec les applications, l'utilisateur peut alors décider s'il souhaite continuer à l'utiliser.
  - Les utilisateurs doivent être informés lorsqu'ils interagissent avec l'IA.
  - Cela inclut les systèmes d'IA qui génèrent ou manipulent du contenu image, audio ou vidéo (par exemple, les deepfakes, des contenus faux qui sont rendus crédibles par l'IA)

Quelles priorités en matière de réglementation et d'éthique ?

- Il n'existe aucun accord général en IA …
  - Doit-on se concentrer en 1er lieu sur les algorithmes qui peuvent expliquer leur raisonnement ?
  - Sur la protection des données ?
  - Sur la robustes des systèmes d'IA face aux attaques malveillantes ?
  - Sur les biais dans les systèmes d'IA ?
  - Sur le « risque existentiel » potentiel lié à une IA super-intelligente ?

- Ce qui fait peur ce sont les risques liés à une IA super-intelligente, mais tout le danger ne viendrait il pas du manque de fiabilité, de transparence de l'apprentissage profond, et à sa vulnérabilité aux attaques malveillantes ?

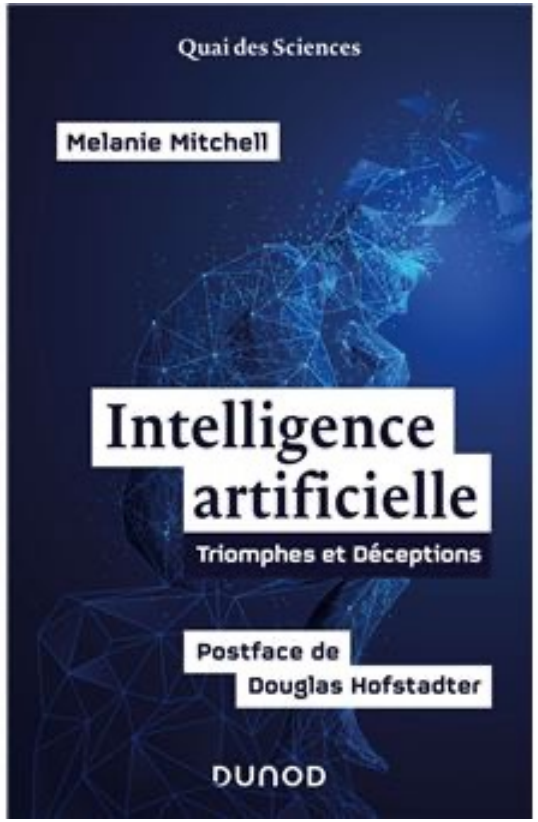
Grand Compromis de l'IA

- Devons-nous reconnaître les aptitudes des systèmes d'IA, qui peuvent améliorer notre existence et même aider à sauver des vies, et accepter que ces systèmes soient de plus en plus largement utilisés ?
- Ou devons-nous être plus prudents en raison des erreurs imprévisibles commises par l'IA actuelle, de sa vulnérabilité aux préjugés et au cyber-piratage, et de son manque de transparence lors de ses prises de décision ?
- Dans quelle mesure les humains doivent-ils prendre part aux décisions prises par les systèmes d'IA ?
- Que devons-nous exiger d'un tel système afin d'avoir suffisamment confiance en lui pour le laisser opérer de manière vraiment autonome ?
- Ces questions font encore l'objet de vifs débats alors même que l'IA est de plus en plus utilisée et que les applications qu'elle a promises (par exemple, les voitures autonomes) existent déjà….

Le mot de la fin pour R. Brook

- « Quand l'IA débuta, la grande idée était clairement la performance de niveau humain et l'intelligence de niveau humain. Je pense que cet objectif a été ce qui a attiré la plupart des chercheurs dans ce domaine durant les premières soixante années.
- Le fait que nous soyons loin de parvenir à ces objectifs ne signifie pas que les chercheurs n'aient pas travaillé dur ou n'aient pas été brillants.
- Il signifie qu'il s'agit d'un objectif très difficile à atteindre.
- En IA, les questions les plus passionnantes ne sont pas uniquement focalisées sur les applications potentielles. Les fondateurs de la discipline furent autant motivés par des questions scientifiques sur la nature de l'intelligence que par le désir de développer de nouvelles technologies. »
- http://rodneybrooks.com/forai-the-origins-of-artificial-intelligence/

Pour aller plus loin



- Ecrit en 2019 (version française 2021)
- Guide pour comprendre l'IA d'aujourd'hui et son impact sur notre avenir
- Présente les modèles dominants de l'IA moderne et de l'apprentissage machine
- Met en valeur la profonde déconnexion entre le battage publicitaire et les réalisations réelles en IA

- Questionne l'IA :
  - Dans quelle mesure les meilleurs programmes d'IA sont-ils vraiment intelligents ?
  - Comment fonctionnent-ils ?
  - Que peuvent-ils réellement accomplir, et quand échouent-ils ?
  - Pourraient-ils véritablement comprendre le monde comme nous le comprenons ?
  - Peut-on laisser des algorithmes prendre des décisions à notre place, sans savoir exactement comment ils les ont prises ?