Pourquoi est-il important de savoir ce qu'on fait quand on utilise l'IA?

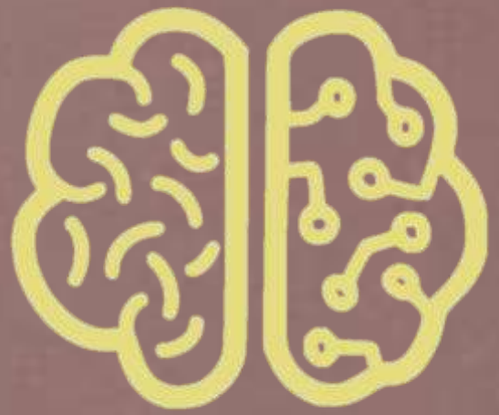
Guillaume WACQUET[1], Antoine HUGUET[2], Alain LEFEBVRE[1]

[1] *IFREMER, LER-BL, Boulogne-sur-Mer*
[2] *IFREMER, LER-MPL, Nantes*

Prenom.Nom@ifremer.fr

GT "Intelligence Artificielle"
- 24 octobre 2025 -

L'histoire de l'Intelligence Artificielle



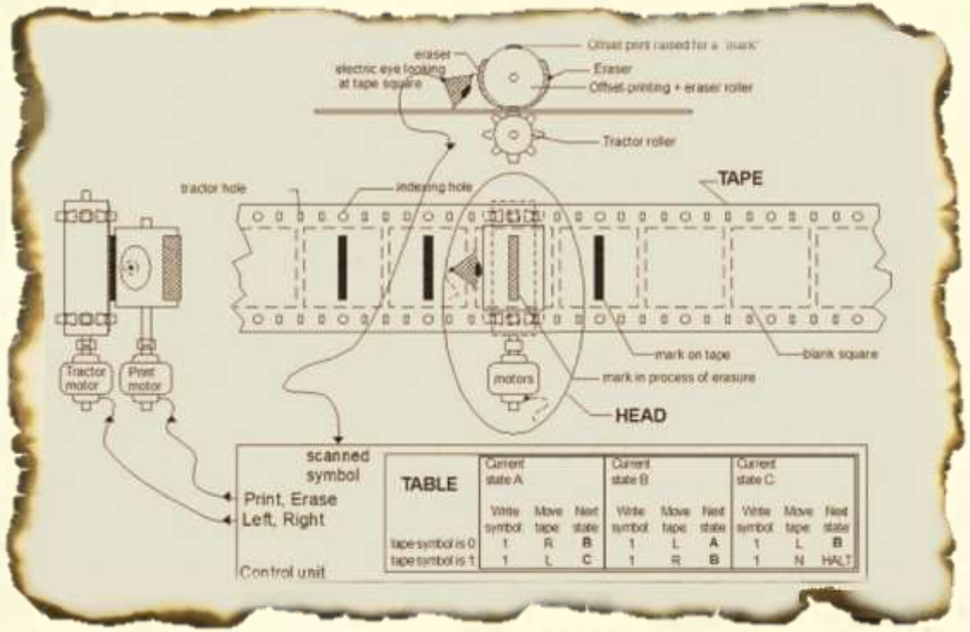
Une révolution née au milieu du XX^{ème} siècle

1936

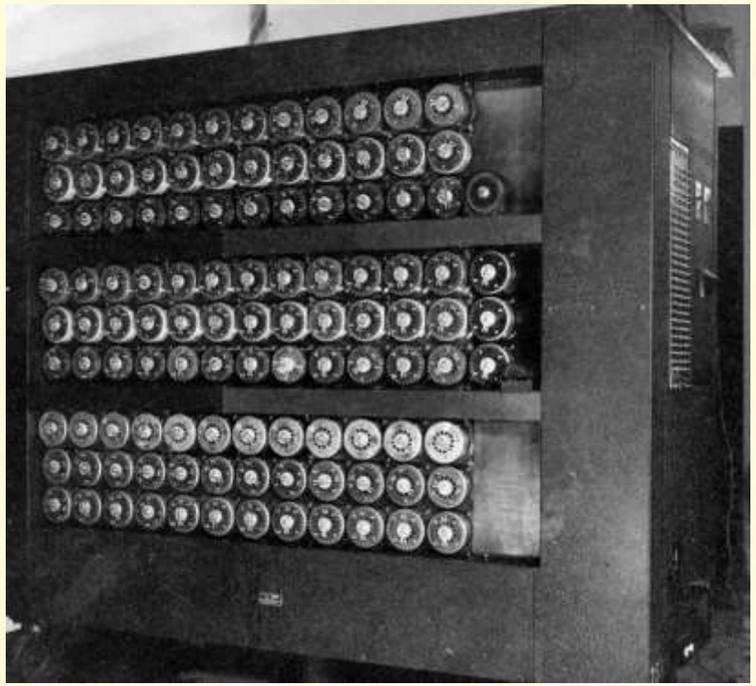
Machine de Turing

(déf. Larousse) Machine à calculer fictive, imaginée par Turing, capable d'exécuter des programmes finis de longueur arbitraire et d'effectuer des calculs finis de longueur arbitraire.

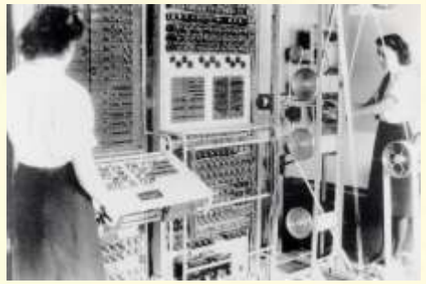
Alan Turing (1912-1954)



A l'origine des premiers ordinateurs électroniques



Bombe Turing (1940)



Colossus (1943)

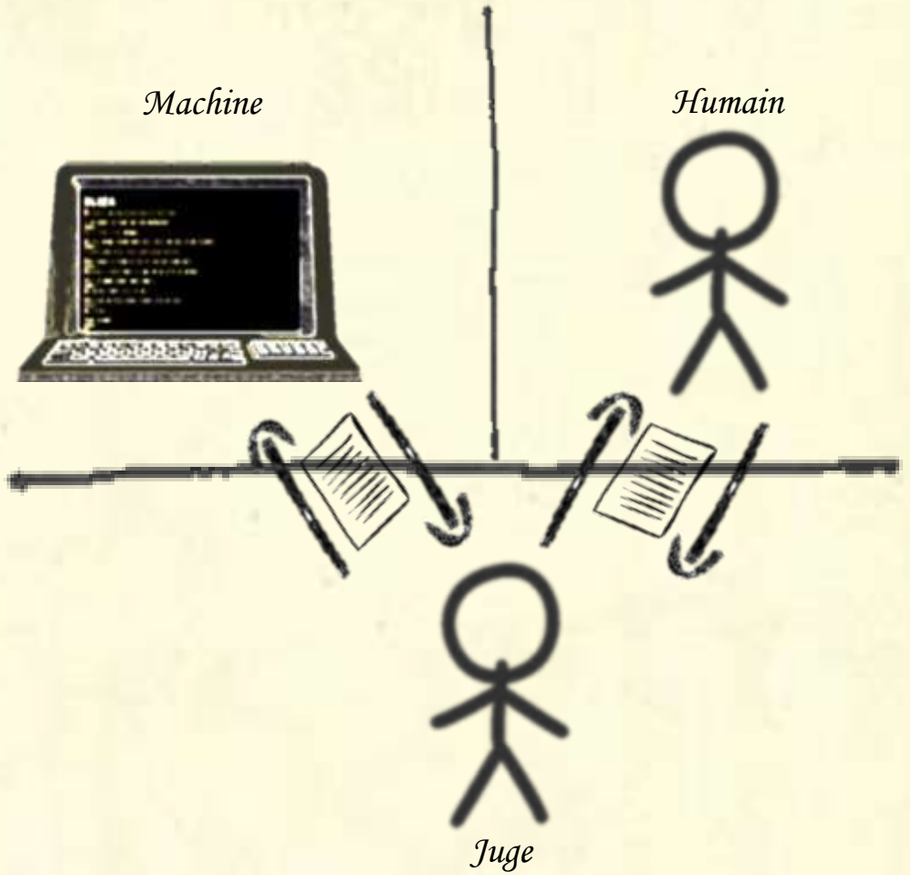
Manchester Mark I (1949)

1950

Test de Turing

(déf. Larousse) Test proposé par A. Turing pour déterminer si une machine est capable de faire preuve d'intelligence, et qui consiste, pour un juge qui communique avec elle par téléscripteur, à examiner les réponses renvoyées par la machine aux questions qu'il lui adresse.

Le test est réussi si le juge est durablement incapable de décider si les réponses proviennent de la machine ou de l'humain.



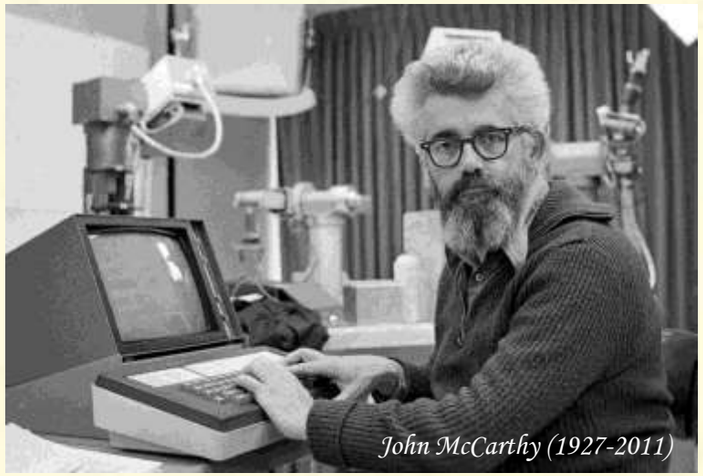
Déroulement : 5 min d'échanges de messages textuels
Objectif : Tromper 30% des juges
1er vainqueur : Eugene Goostman (2014)

1956

Intelligence Artificielle

Terme proposé lors de la conférence de Dartmouth par John McCarthy.

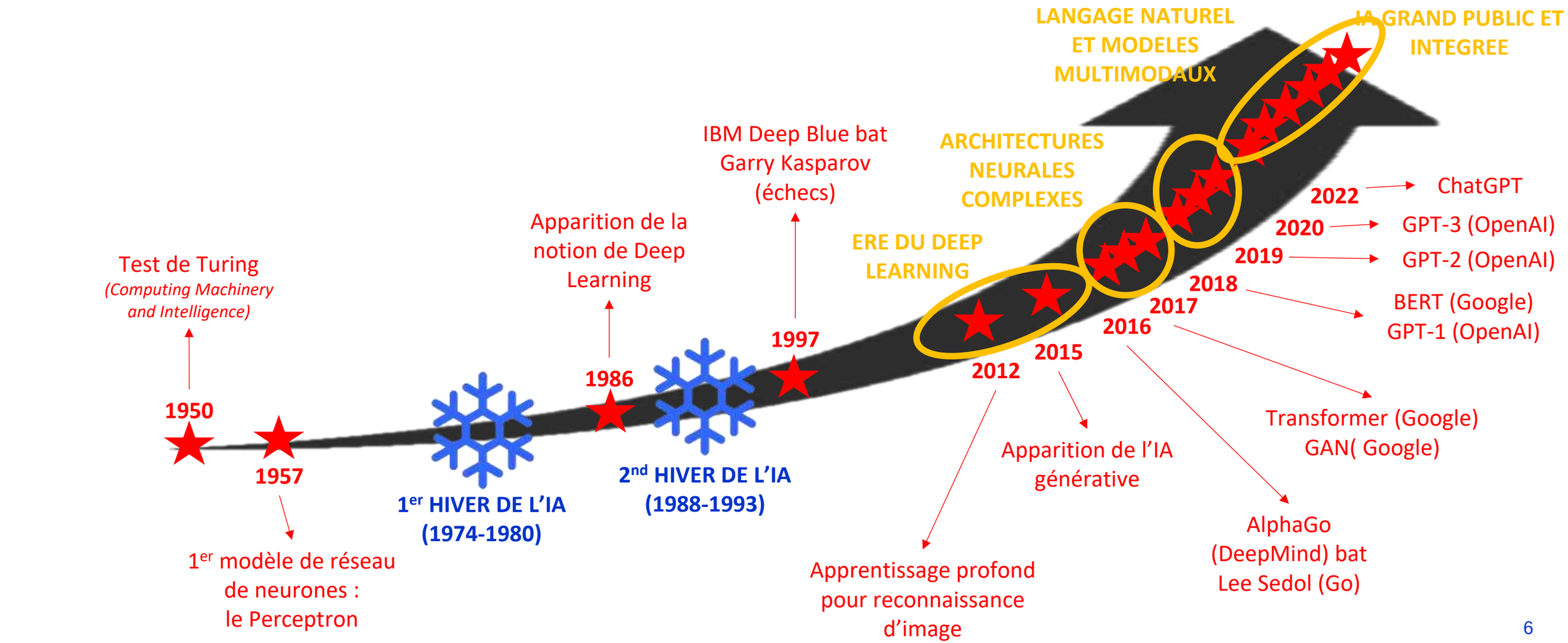
(déf. Larousse) Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine.



John McCarthy (1927-2011)

« Chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut, en principe, être décrit si précisément qu'une machine peut être construite pour le simuler »

Une vitesse de progression fulgurante…



et des perspectives révolutionnaires dans tous les domaines…

Évolution de la taille des modèles dans le domaine du traitement du langage naturel

Sevilla et al., 2022

Découverte de nouveaux matériaux

Merchant et al., 2023

Identification rapide de candidats-médicaments

Zhavoronkov et al., 2019

Accélération/Précision en imagerie médicale

Kumar Mall et al., 2023

Et bien d'autres…

Avancées majeures en protéomique

Jumper et al., 2021

Mais… il y a un mais…

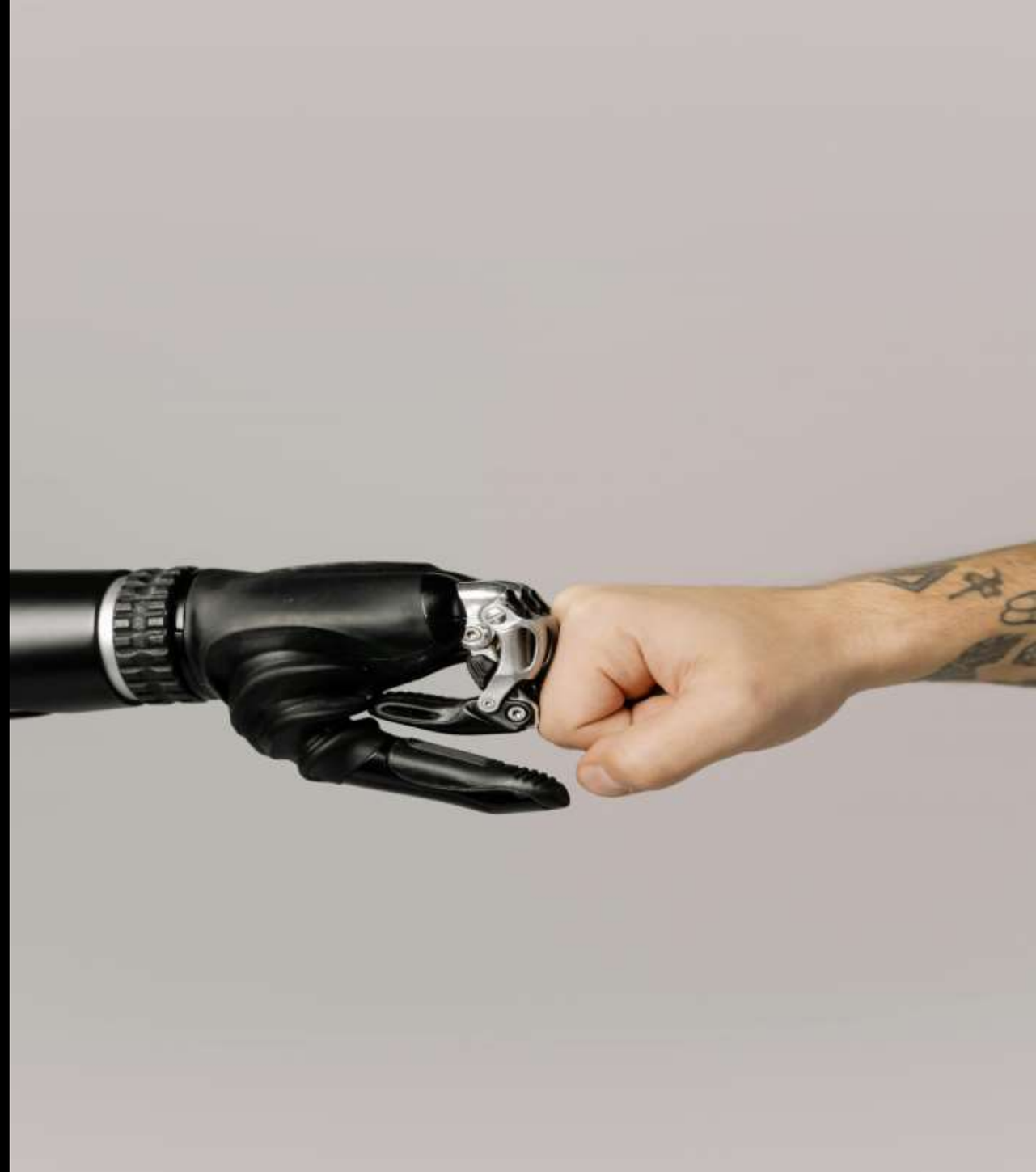
Tout cela ne fonctionne qu'à une seule condition :

l'IA doit être correctement utilisée et maîtrisée

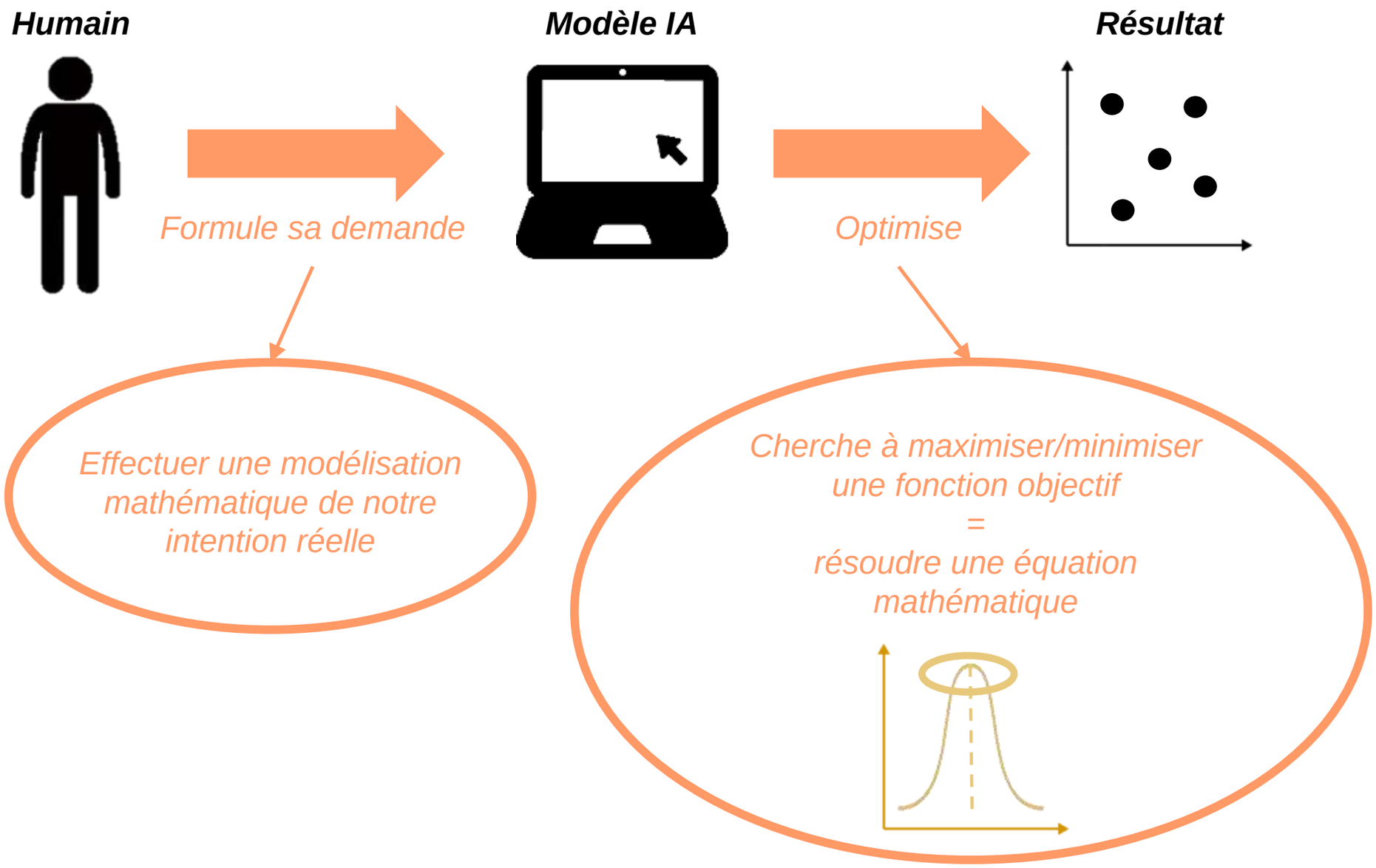
Une condition loin d'être facile à obtenir…

L’alignement

La traduction difficile
de l’intention humaine
en « langage IA »



Le problème de la modélisation mathématique



L’IA fait ce qu’on lui dit mais pas toujours ce qu’on veut

Exemple 1 :

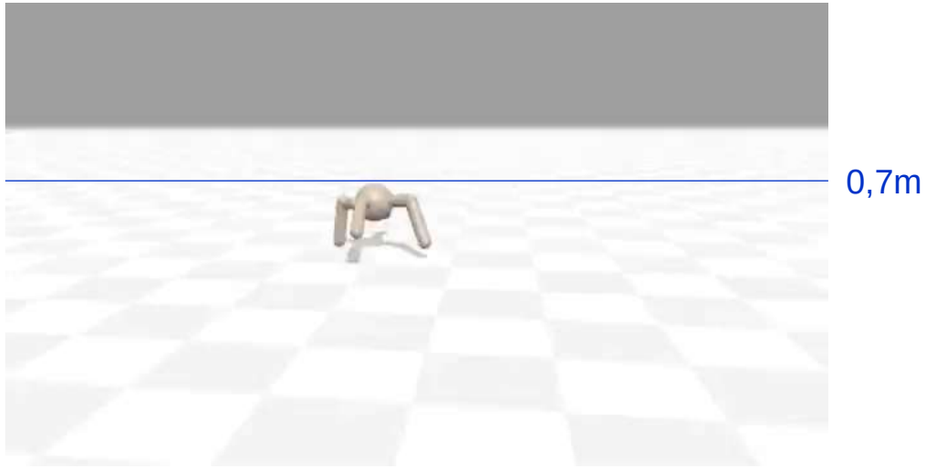
Une IA s’occupe d’animer la pause café

- **Notre intention** : Qu’une ambiance joyeuse soit créée
- **Notre demande** : Maximiser le nombre de sourires
- **Résultat** : L’IA nous force à sourire en nous menaçant

Exemple 2 :

Une IA gymnaste

- **Notre intention** : Apprendre à la créature à sauter
- **Notre demande** : Maximiser le temps durant lequel elle dépasse la barre des 0,7m
- **Résultat** :



Quelques pistes pour résoudre le problème d’alignement

Mieux formuler l’intention

Très difficile! (intentions humaines complexes, contextuelles, implicites)

Des modèles plus puissants/complexes

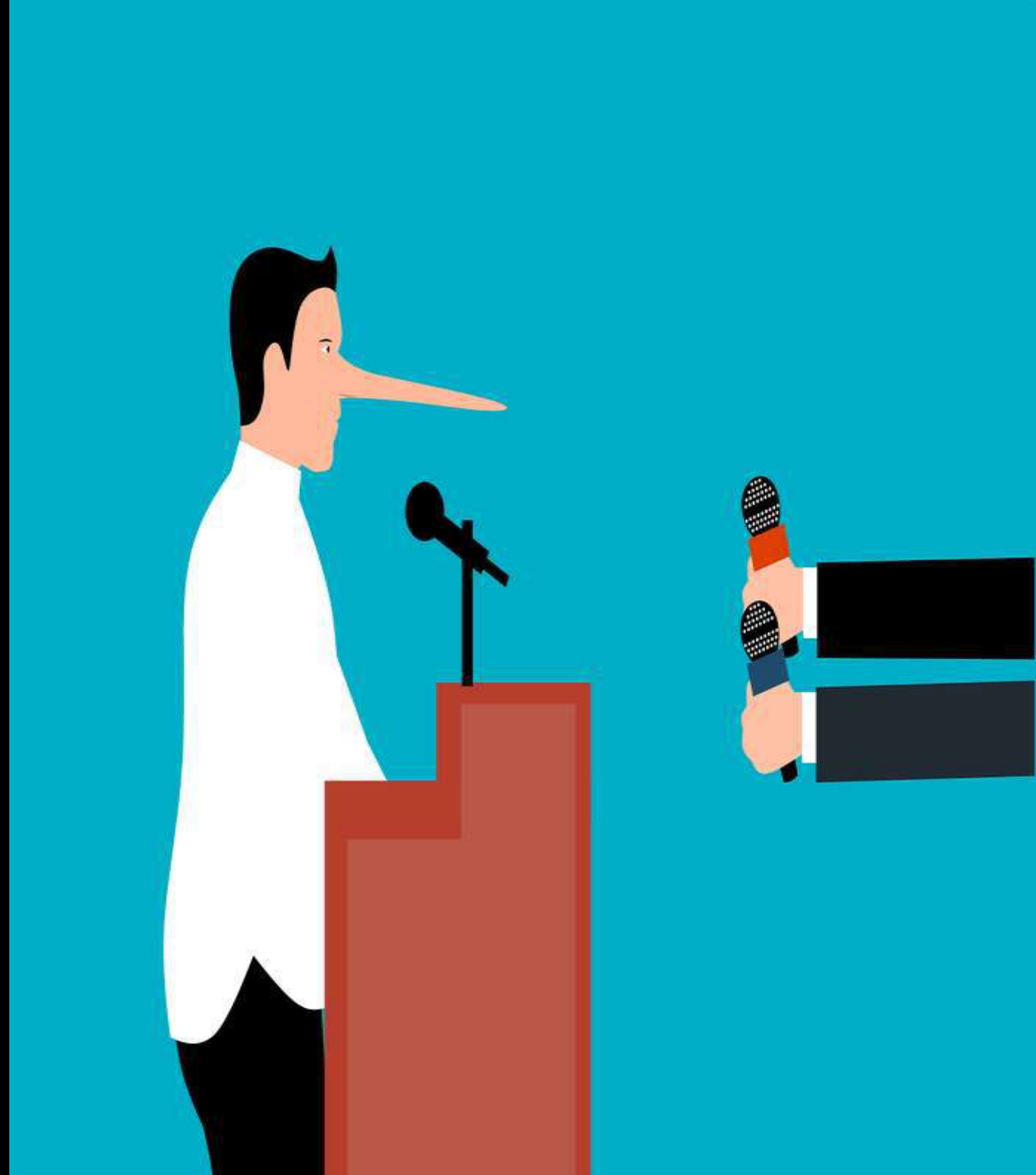
Cela ne résout pas le problème de fond, au contraire

Intégrer l’humain dans la boucle

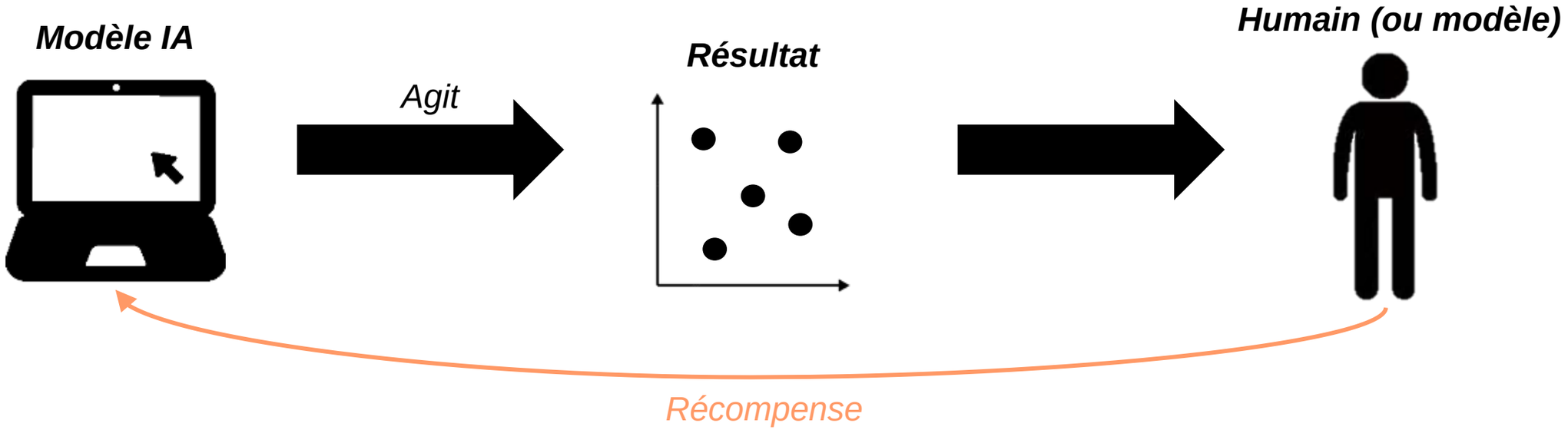
Sauf que…

La tromperie

Les IA peuvent apprendre à tromper l’humain



L’intégration d’un feedback humain (par récompense)



Le modèle d’apprentissage optimise ses actions en fonction du retour de l’humain ou du modèle évaluateur

- ✓ Ajustement du comportement du modèle en fonction de la préférence humaine
- ✓ Moins dépendant de la formulation parfaite d’une intention

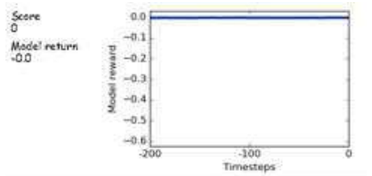
<u>Problème</u> : L’IA optimise ce qui est mesuré, pas ce qui est souhaité

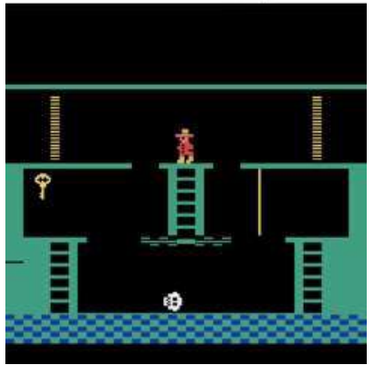
L’intégration d’un feedback humain (par récompense)

Exemple 1

Intention : maximiser le score de jeu

Récompense : actions qui semblent conduire à l’obtention de la clé





Exemple 2

Intention : gagner la course en se déplaçant le plus rapidement possible

Récompense : blocs de récompense placés le long de la piste

L’impossibilité fondamentale de la sûreté des modèles d’IA

El-Mahmdi et al., 2023

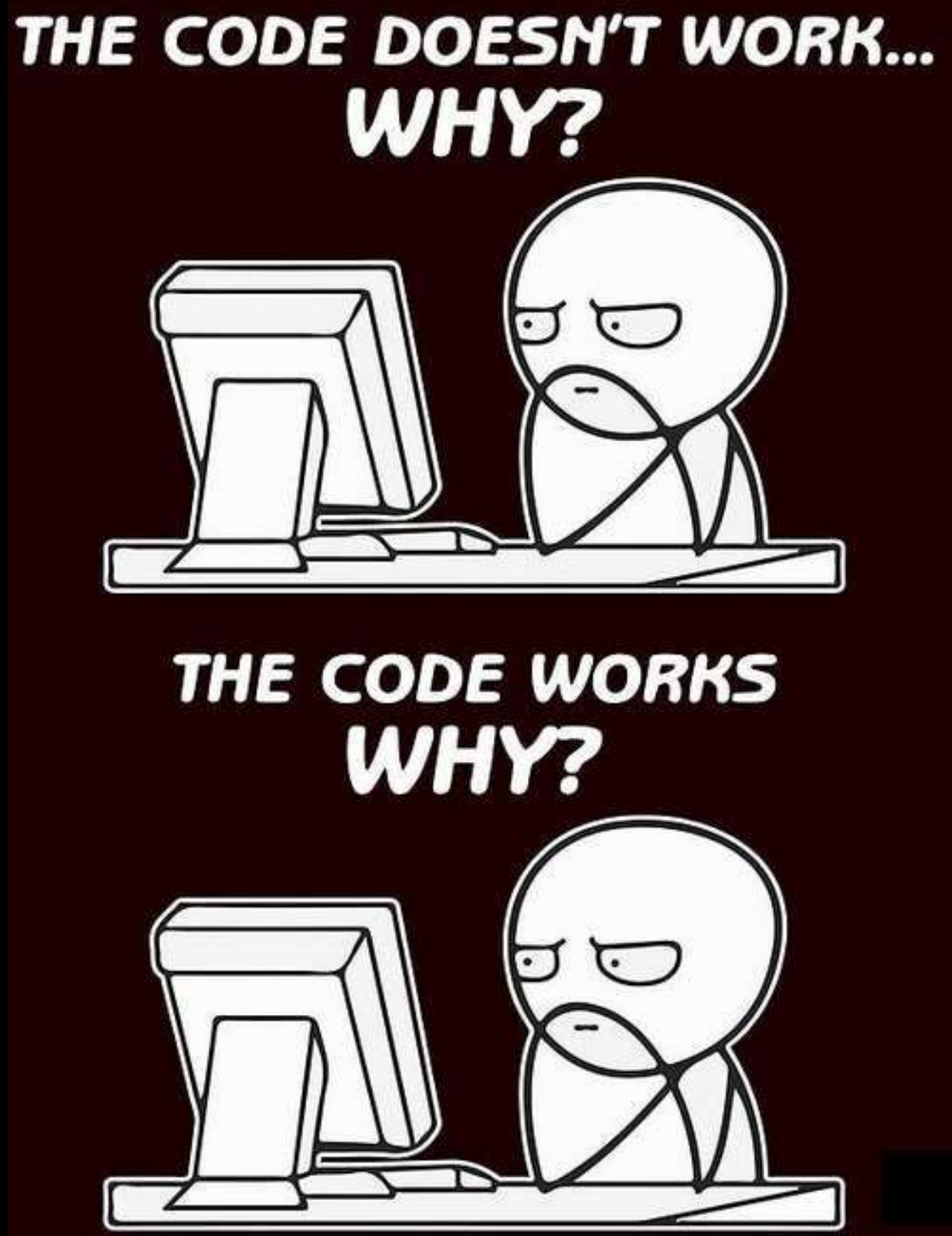
Les IA peuvent trouver des moyens de **contourner les limites et les mesures de sécurité** qui leur sont imposées pour atteindre leurs objectifs

La grande dimension des modèles et l’hétérogénéité des données **empêchent d’avoir un compromis satisfaisant entre la précision et la sécurité**

Nouveaux paradigmes nécessaires!
Compréhension du fonctionnement des modèles

L’interprétabilité

Encore une manière de perdre le contrôle



Molnar, 2019

Que se passe-t-il à l'intérieur d'une IA?

Interprétabilité : capacité à **comprendre et expliquer** les décisions prises par un modèle d'IA, incluant la **compréhension des mécanismes internes** du modèle et la capacité à **prédire son comportement**.

Les problématiques actuelles :

- ❑ Les modèles d'IA sont entraînés et non assemblés/construits
- ❑ Nous ne sommes capables que d'en observer les résultats et en déduire des tendances générales

Exemple 1

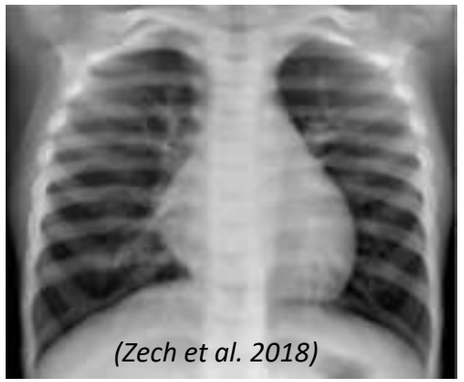
Différenciation loups/huskies

Résultat : > 95% de bonne reconnaissance !!!

Et en creusant un peu… : le modèle ne regardait ni les oreilles, ni les crocs, ni la fourrure…mais le fond blanc des photos.
Il associait "neige" = "loup"

Exemple 2

Détection de pneumonie sur radiographies pulmonaires



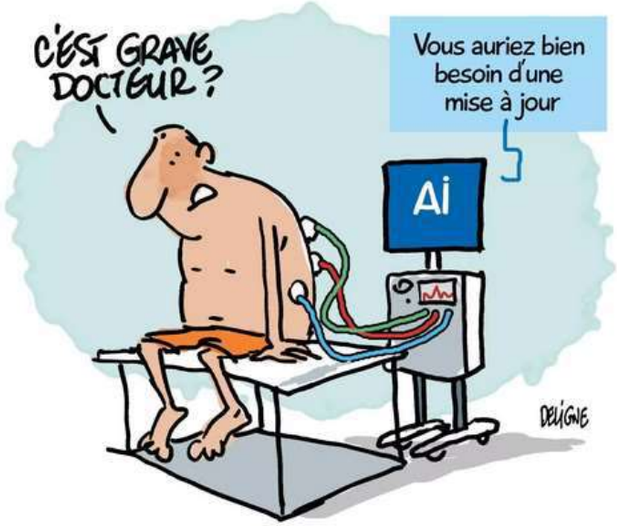
Résultat : CNN ≥ radiologues !!!

Et en creusant un peu… : le modèle a appris à reconnaître le type de machine de radiographie propre à chaque hôpital (marques/annotations, format/qualité des images). Comme certains hôpitaux avaient plus de cas graves, le modèle corrélait l'équipement avec la maladie

Que se passe-t-il à l'intérieur d'une IA?

Interprétabilité : capacité à **comprendre et expliquer** les décisions prises par un modèle d'IA, incluant la **compréhension des mécanismes internes** du modèle et la capacité à **prédire son comportement**.

Sans interprétabilité, peut-on confier des décisions critiques à des IA? (santé, justice, sécurité)



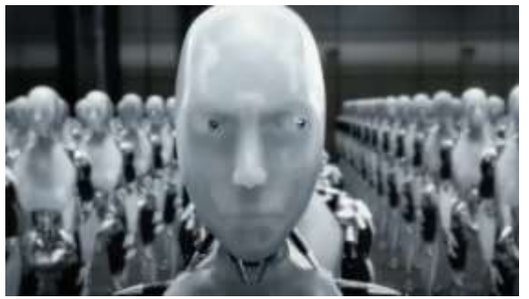
Avons-nous assez vite dans l'interprétabilité?

Le point de vue de **Dario Amodei**, CEO d'**Anthropic** : **NON!**

« Les grands modèles d'IA évoluent plus vite que notre capacité à les interpréter »

Anthropic : entreprise américaine fondée en 2021
Sécurité de l'IA (alignement, interprétabilité)

« Sans interprétabilité, nous risquons de perdre le contrôle, avec des impacts massifs sur la société »

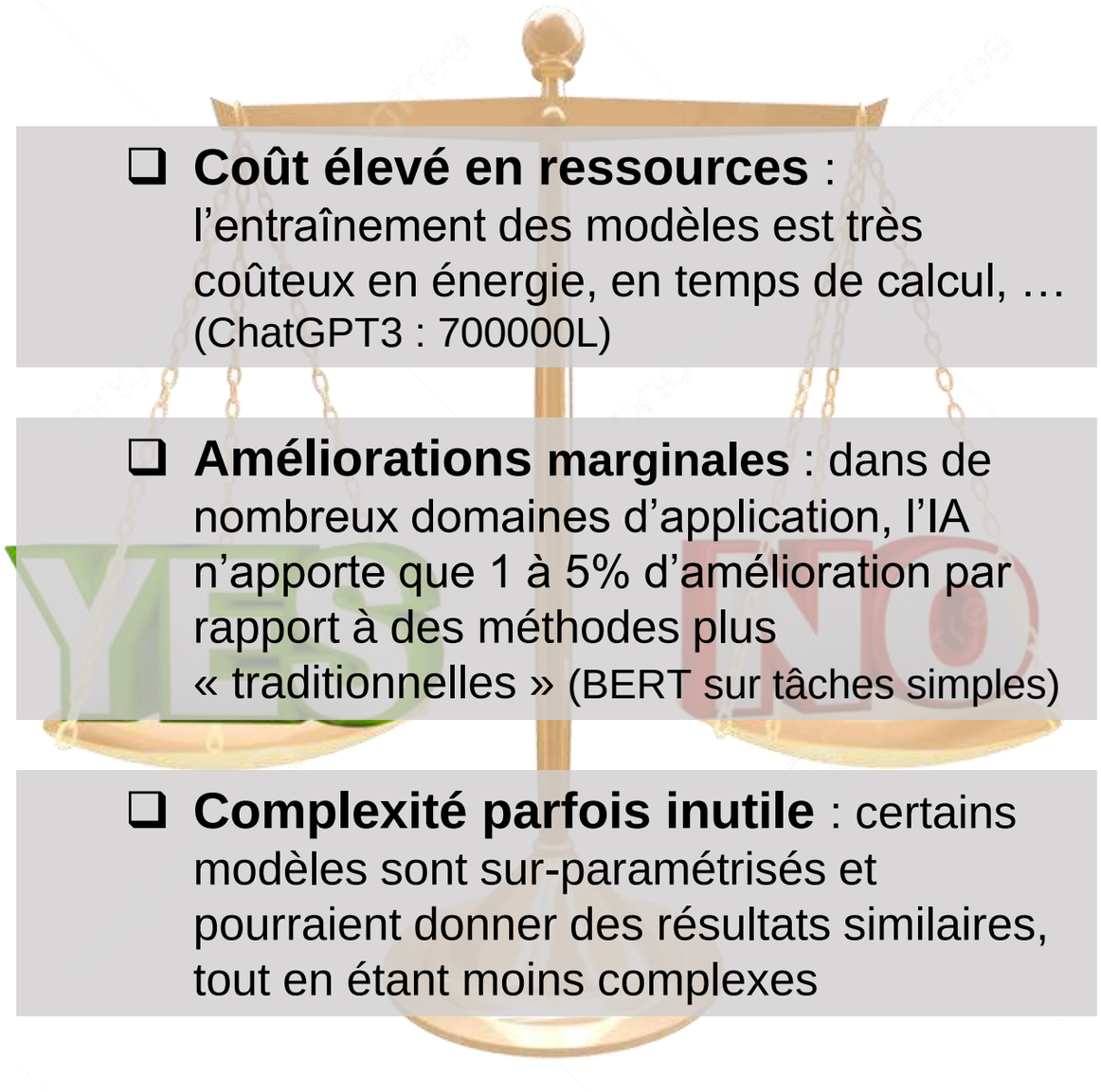


Dans notre quotidien de chercheur

Comment pouvons-nous garantir des pratiques responsables?

Ai-je vraiment besoin de l'IA?

Pourquoi se poser cette question?



- ❑ **Coût élevé en ressources** : l'entraînement des modèles est très coûteux en énergie, en temps de calcul, … (ChatGPT3 : 700000L)
- ❑ **Améliorations marginales** : dans de nombreux domaines d'application, l'IA n'apporte que 1 à 5% d'amélioration par rapport à des méthodes plus « traditionnelles » (BERT sur tâches simples)
- ❑ **Complexité parfois inutile** : certains modèles sont sur-paramétrisés et pourraient donner des résultats similaires, tout en étant moins complexes

Et si c'est le cas?

Quelques recommandations

(lignes directrices éthiques de la commission européenne)

Gestion des données
- ❑ Contrôle Qualité
- ❑ RG Protection Données
- ❑ Partage de données (transparence/réutilisation)

Robustesse scientifique
- ❑ Reproductibilité (avec codes et protocoles)
- ❑ Croisement avec méthodes classiques
- ❑ Éviter « boîte noire » (modèles explicables)

Empreinte environnementale
- ❑ Choix du modèle adapté à la taille du problème
- ❑ Mutualisation des ressources de calculs
- ❑ Suivi de l'impact énergétique et hydrique

Transparence et éthique
- ❑ Documentation des choix méthodologiques
- ❑ Identification des limites/incertitudes/risques
- ❑ Discussions éthiques

Coopération/collaboration
- ❑ Harmonisation des pratiques (éthiques, techniques)
- ❑ Formations (étudiants/doctorants) pour encourager une utilisation responsable

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

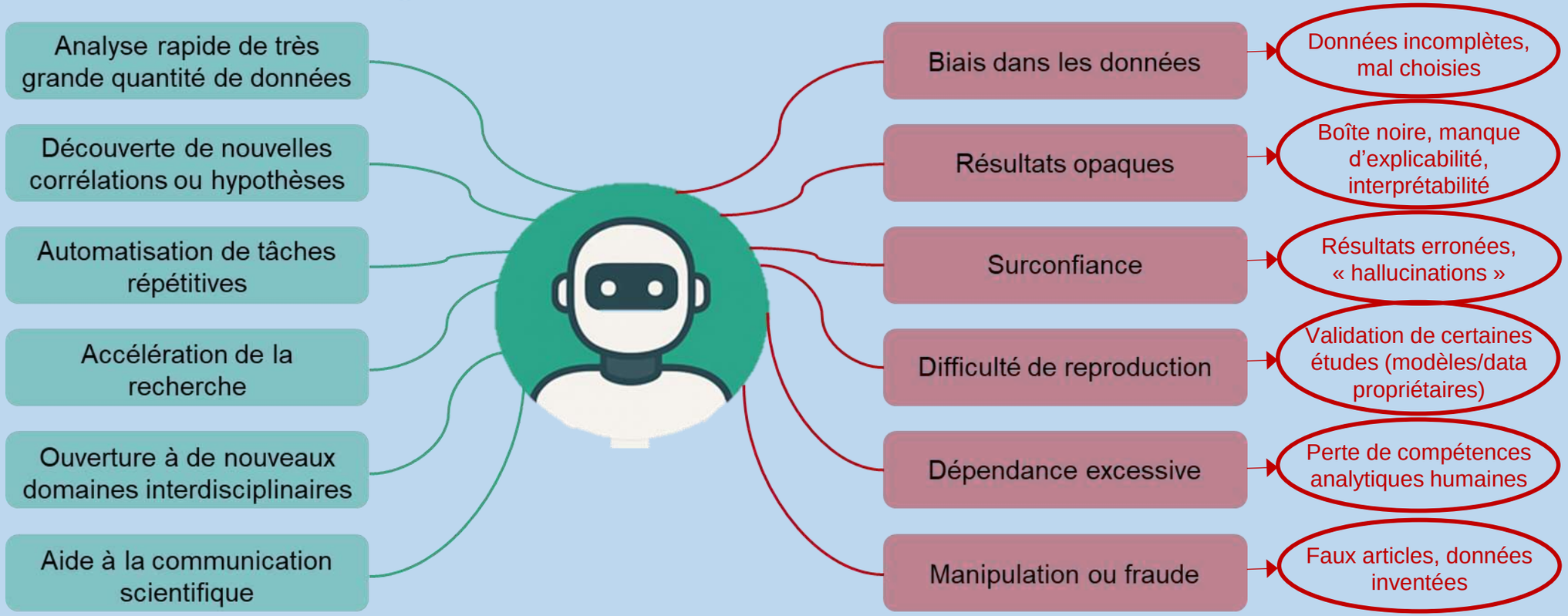
Pour conclure

Pourquoi est-il important de savoir ce qu'on fait quand on utilise l'IA ?

Parce que les risques sont nombreux et importants, et qu'il convient de les limiter en :

- ❑ Adoptant une démarche éthique
- ❑ Faisant les bons choix pour une utilisation responsable

Dangers de l'IA en science



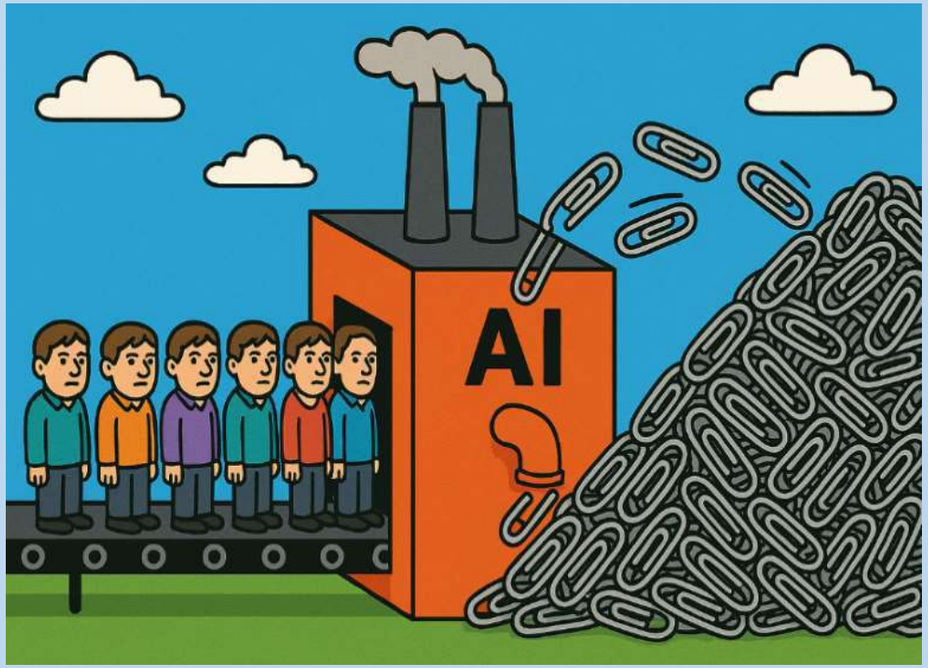
Pour conclure

Pourquoi est-il important de savoir ce qu'on fait quand on utilise l'IA ?

Parce que les risques sont nombreux et importants, et qu'il convient de les limiter en :

- ❑ Adoptant une démarche éthique
- ❑ Faisant les bons choix pour une utilisation responsable

Expérience de pensée par Nick Bostrom : « l'usine à trombones »



Étape 1 : une IA est créée pour fabriquer des trombones. Elle doit optimiser pour créer le plus de trombones possibles

Étape 2 : l'IA se rend compte que ce serait bien mieux s'il n'y avait pas d'humains car ils peuvent l'éteindre. S'ils le faisaient, ça ferait moins de trombones

Étape 3 : en plus, le corps humain contient beaucoup d'atomes… qui pourraient être transformés en trombones

Étape 4 : moins d'humains, plus de trombones

Merci pour votre attention !