



**HAL**  
open science

# L'IA aujourd'hui Comment ça marche? Quelles limites?

Thomas Guyet

► **To cite this version:**

Thomas Guyet. L'IA aujourd'hui Comment ça marche? Quelles limites?. Initiation à l'IA et à ses enjeux informationnels pour les professionnels des bibliothèques et de la documentation, Enssib, Feb 2025, Lyon, France. hal-04946830

**HAL Id: hal-04946830**

**<https://inria.hal.science/hal-04946830v1>**

Submitted on 13 Feb 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# L'IA aujourd'hui




Comment ça marche ? Quelles limites ?

Thomas Guyet  
AlstroSight  
thomas.guyet@inria.fr

# 01

## Tentative de définition



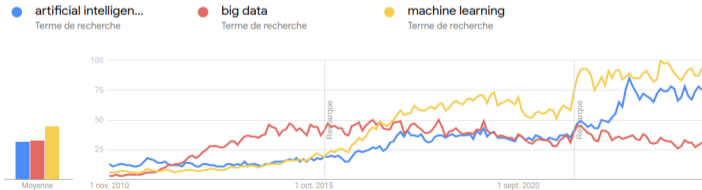
# « Intelligence Artificielle » : LE buzz-word du moment

Intelligence Artificielle (IA) utiliser à tout va ...

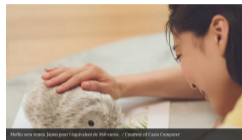
Attention aux abus : Promesses de « progrès » de l'Intelligence Artificielle (IA)

- ▶ performance (« à la pointe »)
- ▶ personnalisation
- ▶ innovation technologique (pour les décideurs)
- ▶ solution pour des questions sociétales (santé, environnement, etc.)

Évolution de l'intérêt pour cette recherche



Grâce à l'IA, cette peluche est capable d'exprimer des émotions



Casio, Moflin<sup>3</sup>





## Aux origines de l'Intelligence Artificielle<sup>4</sup>



August 1956. From left to right :  
Oliver Selfridge, Nathaniel Rochester,  
Ray Solomonoff, Marvin Minsky,  
Trenchard More, John McCarthy,  
Claude Shannon.

### Cybernétique

Établir une théorie générale du fonctionnement de la pensée humaine basée sur la théorie du contrôle  
Conférences Macy 1946-1953 (Wiener, McCulloch, Pitts, Ashby).

### Intelligence Artificielle

Dartmouth 1956 (McCarthy, Minsky, Shannon, ...)  
Naissance du terme *Intelligence Artificielle*  
Construire des dispositifs simulant les processus cognitifs humains

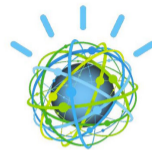
# Deux conceptions opposées de la cognition

## Cognition = Raisonnement logique

- ▶ IA « Symbolique »
- ▶ Principes
  - L'information peut être représentée sous la forme de symboles
  - La pensée peut être décrite comme l'application de règles formelles
- ▶ ex. planification, résolution de problèmes, diagnostic, etc.

## Cognition = Émergence de l'activité neuronale

- ▶ IA « connexionniste »
- ▶ Principes
  - Simuler le fonctionnement des neurones = simulé la cognition
  - Pas de représentation explicite du monde
- ▶ ex. reconnaissance de formes, diagnostic, etc.



IBM Watson

IBM Watson<sup>8</sup>

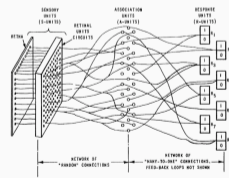


Figure 1 ORGANIZATION OF THE MARK I PERCEPTION

Rosenblatt<sup>12</sup>, 1958

Et d'autres encore programme des recherches en IA ...



## Vers une définition des systèmes d'intelligence artificielle (SIA) ...

### Unesco, Recommendation on the Ethics of Artificial Intelligence<sup>17</sup>, 2021

Les « systèmes d'IA » sont des **technologies de traitement des informations** qui intègrent des modèles et des algorithmes, lesquels génèrent une **capacité d'apprentissage** et d'**exécution de tâches cognitives** conduisant à des résultats tels que l'anticipation et la prise de décisions dans des environnements matériels et virtuels. Les systèmes d'IA sont conçus pour fonctionner avec différents degrés d'autonomie, au moyen de la modélisation et la représentation des connaissances, de l'exploitation des données et du calcul de corrélations. Ils peuvent intégrer plusieurs méthodes, telles que, sans s'y limiter :

- ▶ l'apprentissage automatique, y compris l'apprentissage profond et l'apprentissage par renforcement ;
- ▶ le raisonnement automatique, y compris la planification, la programmation, la représentation des connaissances et le raisonnement, et la recherche et l'optimisation.

autres définitions

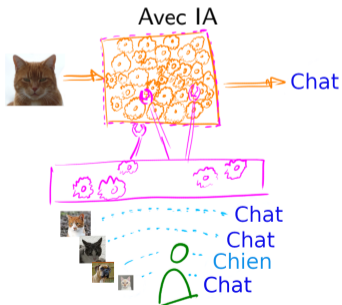
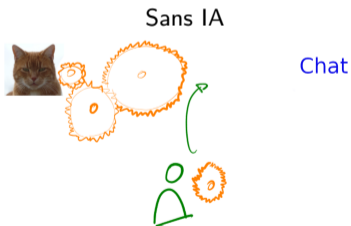
02

« Système d'IA Apprenants » : pourquoi c'est tentant ?





# « Système d'IA Apprenants » : principes (1)



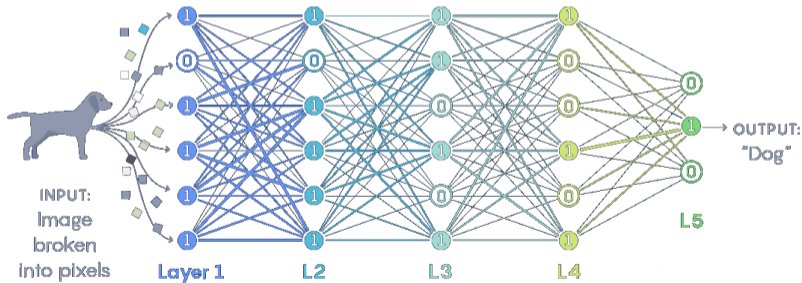
# « Systèmes d'IA Apprenants » : modèle connexionniste

## Réseaux de neurones

Neurone artificielle : brique élémentaire de calcul inspirée d'un neurone biologique

- ▶ valeurs numériques en entrée
- ▶ valeur numérique en sortie

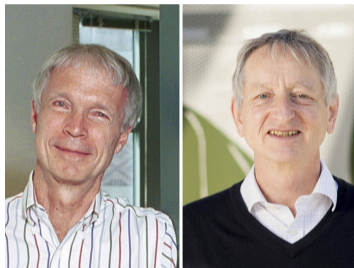
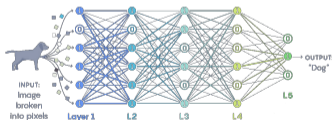
Connexions du réseau : « paramètres du modèle », ~ 153 paramètres



# « Systèmes d'IA Apprenants » : Retro-propagation du gradient (de l'erreur)

## Apprentissage d'un réseau de neurones

- ▶ résultat de la **minimisation des erreurs** faites sur les exemples présentés
- ▶ lors de la présentation d'un nouvel exemple :
  1. tentative de prédiction
  2. calcul d'une erreur commise
  3. modification proportionnelle des connexions du réseau (*vers l'arrière*)



J. Hopfield & G. Hinton  
Prix Nobel de physique 2024



## Quelques points importants

**L'utilisation d'un « Système d'IA Apprenant » remplace la conception d'un système informatique par une procédure d'apprentissage à partir d'exemples**

Un principe de conception d'un système informatique très attrayant ...

**La procédure d'apprentissage s'appuie sur**


- ▶ une collection d'exemples *labellisés* (*chat/chien*)
- ▶ la minimisation d'une erreur

**Une fois appris, le modèle appris peu être réutilisé sur de nouveaux exemples**

Enjeu : on cherche un pouvoir de généralisation

# 03

**« *Deep learning* » – émergence des grandes architectures : qu'est-ce qui change ?**



## « *Deep learning* » : émergence des grandes architectures

### *Deep learning* = Apprentissage des représentations

- ▶ Données d'entrée beaucoup moins contraintes
  - variétés de données plus larges (images, signaux, texte, ...)
  - moins de préparation nécessaire (*feature engineering*)
- le réseau « apprend » quelle information utiliser pour bien réaliser sa tâche

### Accroissement des architectures de réseau de neurones

- ▶ Plus de couches/neurones  $\Rightarrow$  plus de complexités
- ▶ Plus de couches/neurones  $\Leftarrow$  besoin de (beaucoup) plus de données



J. Bengio, G. Hinton & Y. Le Cun  
Prix Turing 2018

# « Deep learning » : le saut AlexNet

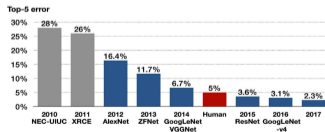
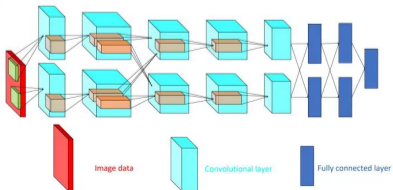
## ImageNet : compétition de classification d'images

14 millions d'images

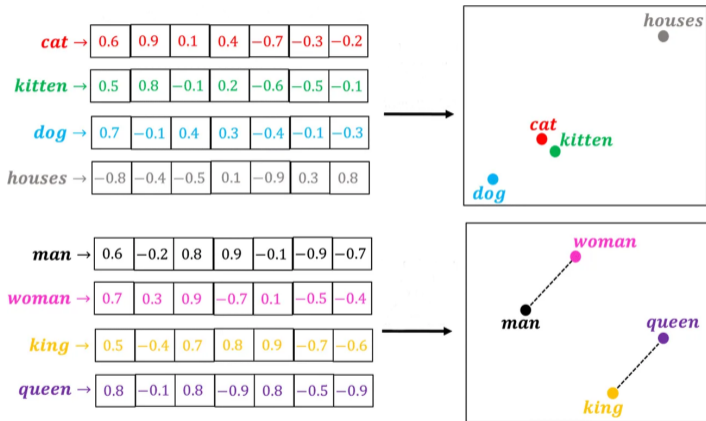
20000 catégories

## AlexNet<sup>9</sup> (2012) : réseau pour la classification d'images

- ▶ Apprentissage de représentation : couches convolutionnelles
- ▶ Classification : couches « fully connected »
- ▶ 60 Millions de paramètres

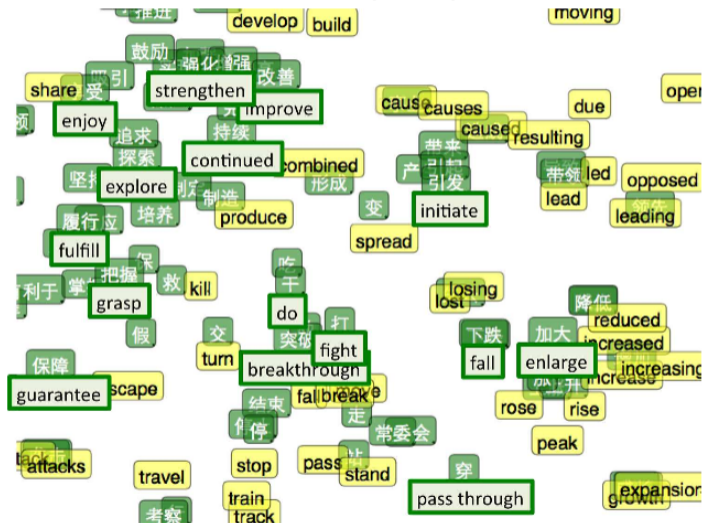


# Représentation vectorielle de mots : principes

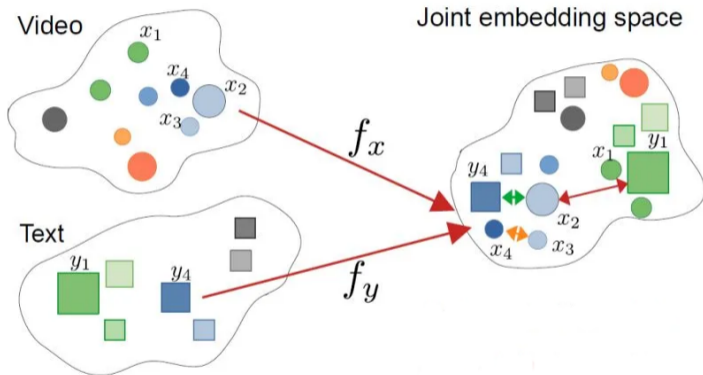




# Représentation vectorielle de mots : principes



## Représentation vectorielle de mots : principes



# Apprentissage d'une représentation vectorielle de mots / textes

## Modèle *Transformer*<sup>18</sup> (2017)

- ▶ Modèle de réseau de neurones permettant de construire des représentations **contextuelles** des mots/phrases
- ▶ GPT : Generative Pre-training Transformer

## Apprentissage par masquage

- ▶ Chaque phrase est présentée avec un mot masqué
- ▶ Le réseau cherche à prédire le mot masqué
  - retropropagation de l'erreur

■ : Center Word  
■ : Context Word

c=0 The cute **cat** jumps over the lazy dog.

c=1 The **cute** **cat** **jumps** over the lazy dog.

c=2 The **cute** **cat** **jumps** **over** the lazy dog.

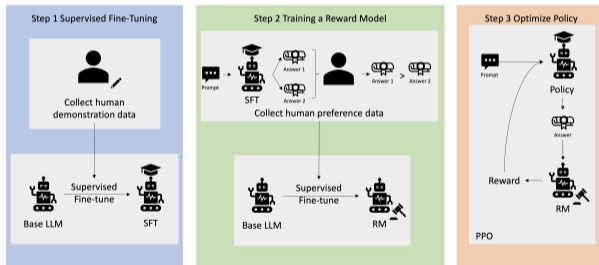
## Le résultat de l'apprentissage est un modèle « pré-entraîné »

- ▶ Réutilisable dans une multitude de tâches (génériques)
- ▶ Très gros modèles ... (*Palm2* 540 milliards de paramètres!)

# GPT : Pourquoi ça marche ?

## RLHF<sup>2</sup> : Reinforcement Learning by Human Feedbacks (GPT-3)

- ▶ Des humains donnent des avis sur des générations
- ▶ Les avis servent à construire un modèle de récompense (*RM*)
- ▶ Le modèle de récompense affine le modèle pré-entraîné



Alternatives moins coûteuses au RLHF/PPO : RLAIF<sup>1</sup> (Anthropic), GRPO<sup>13</sup> (DeepSeek), ...

## Quelques points importants du « *deep learning* »

### « *Deep learning* »

- ▶ Accroissement de la taille des modèles
- ▶ Hétérogénéité des données d'entrées (textes, images, sons)
- ▶ Apprentissage de **représentations génériques** : capture des régularités du monde

### Principes de construction du système informatique similaire à ceux de l'apprentissage automatique, mais :

- ▶ Complexification des procédures
- ▶ Contrôle de l'apprentissage moins direct

### Recette de la réussite

- ▶ Beaucoup de données
- ▶ Beaucoup de calculs (et d'énergie)
- ▶ Beaucoup – *mais pas tant que cela* – d'annotation humaine (à bas coût)

04

# IA Génératives : Ça fait des choses qui plaisent !





# Des modèles de langues aux IA Génératives

## chat-GPT

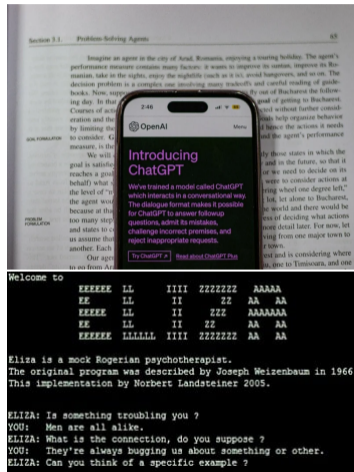
- ▶ Utilisation d'un modèle de langue pour un usage conversationnel
- ▶ Chat-GPT est un réseau de neurones basé sur GPT et affiné pour converser

## Accessibilité, intuitivité de l'interaction avec une machine

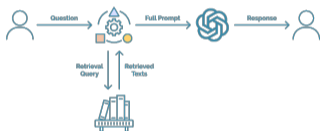
- ▶ Vers le remplacement des moteurs de recherche ...

## L'effet Eliza<sup>5</sup> (1966)

- ▶ Invitation à l'interaction
- ▶ Attribution d'une « *intelligence* »



# RAG : tendance pour des besoins plus opérationnels



## Défauts des modèles conversationnels

- ▶ les hallucinations
- ▶ données d'entraînement obsolètes
- ▶ connaissances non-spécifiques à une entreprise / un domaine

## RAG : Retrieval Augmented Generation<sup>10</sup>


Modèle conversationnel branché à une base de connaissances

Modèle génératif de texte utilisé pour structurer la réponse



# 05

## Quelles questions ? Quelles limites ?



# Un système d'IA est-il un outil comme un autre ?

## Un système d'IA est un outil de traitement de l'information

- ▶ Réalise une tâche pour laquelle il a été construit
- beaucoup de questions liées aux risques et à l'éthique des SIA peuvent être abordés par des principes usuels
  - éducation
  - régulation/encadrement

## Ce qui pose question sur la nature de cette technique ?

- ▶ Il réalise des tâches cognitives (et non-plus mécaniques)
- ▶ Très grande versatilité
- ▶ Niveau de complexité élevé *et* un usage aisé
- ▶ La conception des SIA apprenants est indirecte : contrôle limité par le concepteur lui-même
- questionnement sur la neutralité de la technique, la transformation par la technique (perte de compétences), ...

## 12 classes de risques des SIA dites « générales »

The International Scientific Report on the Safety of Advanced AI<sup>15</sup>, 2025

### Risks from malicious use

- ▶ Harm to individuals through fake content
- ▶ Manipulation of public opinion
- ▶ Cyber offence
- ▶ Biological and chemical attack

### Risks from malfunctions

- ▶ Reliability issues
- ▶ Bias
- ▶ Loss of control

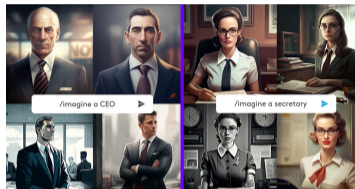
### Systemic risks

- ▶ Labour market risks
- ▶ Global AI R&D divide
- ▶ Market concentration and single points of failure
- ▶ Environmental risks
- ▶ Privacy risks
- ▶ Copyright infringements

## Quelques défauts bien connus des SIA génératifs

### Biais (de genre, d'ethnicité, etc.)

- ▶ Les biais d'un modèle sont nécessaires ! Mais certains biais sont discriminants ...
- ▶ Une grande partie du problème provient des données (leur sélection, etc.)
- ▶ Les mécanismes d'apprentissage automatique ont **tendance à renforcer les biais existants dans les données**



### Hallucinations

- ▶ Quelle est la part d'invention dans la réponse ?
- challenge : apprendre à dire "je ne sais pas" !

**Conserver un regard critique sur les productions et les usages**

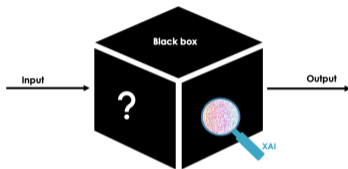
Outil à tester : Compar:IA<sup>7</sup>



## L'explicabilité des résultats d'un SIA est-elle bien nécessaire ?

**Système d'IA = boîte noire**

Trop complexe pour en décrire le fonctionnement



### Confiance vs explicabilité (XAI)

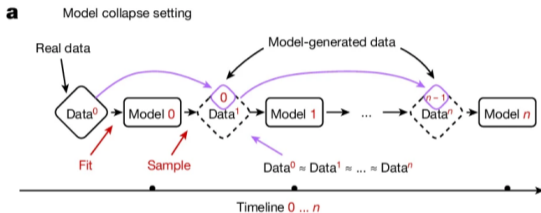
L'explicabilité est un vecteur de construction de la confiance dans les Systèmes d'IA

### Pourrait-on imaginer se passer d'explications à terme ?

- ▶ dans des usages routiniers sans impact : probablement !
- ▶ dans des usages plus pointus : le retour au source restera nécessaire
  - utilisation des RAGs (hybridation des méthodes symboliques/neurales)
  - révéler des causalités dans les données (vs corrélations)

# Le risque d'auto-pollution des sources pour les IA génératives

Les SIA s'écroulent sur les textes générés récursivement<sup>14</sup>



Problème : les contenus du net sont envahis de contenus générés artificiellement

Comment détecter des contenus générés artificiellement ?

# Empreinte environnementale des systèmes d'IA

## Impact réel difficile à quantifier mais réel

- ▶ matériel et fonctionnement
- ▶ phase d'apprentissage et phase d'inférence
- limite pratique à l'expansion des Systèmes d'IA (génératifs)

## Amélioration de l'efficacité énergétique des modèles : IA Frugale/énergies décarbonnées

Attention à l'effet rebond

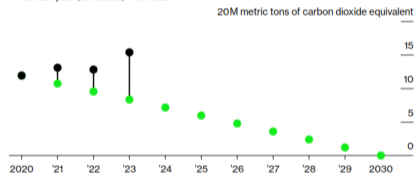
## Besoin d'éducation des usages du numérique

<https://altimpact.fr/>

### Microsoft's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril

● Climate plan (simulated) ● Actual



Source: Microsoft (Scope 1, 2 and 3 "management criteria" data)

Note: Green dots represent linear decline to carbon negative goal.



## Coût social des SIA

### Travailleurs du click

- ▶ Annotateurs Kenyan (Sama) : entre \$1.32/h et \$2/h<sup>16</sup>
- ▶ Exposition à des contenus traumatisants

### Remplacement pour certaines activités

- ▶ Système d'IA pour **assister** ou pour **remplacer** certains travailleurs ?
- **Oui!** Certaines activités vont disparaître
- ▶ Quelques secteurs potentiellement impactés par la disparition d'activités :
  - production *routinière* de documents (journalisme de brève, rédaction de notes synthétiques, illustrations visuelles ou sonores, codeurs de spécifications, etc.)
  - poste de monitoring visuel (de qualité, en santé, etc.)
- ▶ Est-ce un progrès (social) ?

**Renversement deshumanisant des positions : humain qui s'adapte aux machines**



## Impacts des SIA sur la formation des esprits

### Modèle de langue (LLM) : y a-t-il du sens derrière les représentations ?

Ces modèles proposent une **certaine représentation du monde** : un point de vue, non-neutre

- ▶ diffusion très large, parfois cachée de cette représentation
- imposent-ils une vision ?
- opportunité pour favoriser une pensée complexe ?

### Les LLMs sont-ils une nouvelle forme de soft power ?

#### Modification de la relation au savoir

Un LLM connaît beaucoup plus de faits que je ne saurai jamais ...

- ▶ Dois je continuer à apprendre par coeur ? quelles conséquences sur la formation des esprits ?
- ▶ Accès à ce savoir par des interfaces imposées (parfois payantes)
- ▶ Un savoir sans structure (sans taxonomie, sans organisation) ...
- apprendre à apprendre avec des LLM !

---

~~Penser avec ou contre un LLM/un SIA : entre paresse et besoin de justification~~

# 06 Conclusions



## Conclusions

### **Le terme « intelligence artificielle » est très utilisé actuellement**

- ▶ difficile à définir (pas manque de bonne définition de l'intelligence)
- ▶ usage (souvent abusif) pour la promesse de modernité ...

### **Présentation d'un point de vue sur les systèmes d'intelligence artificielle apprenant**

- ▶ les systèmes d'intelligence artificielle (SIA) peuvent prendre d'autres formes
  - ▶ la dimension d'apprentissage automatique soulève le plus de questionnement et de fascination
- tentative de démystification en présentant
- des mécanismes de construction des SIA
  - la notion de représentation sémantique utilisés par les IA Génératives

### **Enumération de beaucoup de questionnements actuels**

- ▶ sur les limites des SIA
- ▶ sur les réflexions à avoir sur leurs usages (régulation, formation)
- ▶ sur les risques qu'il nous faut anticiper

*Merci.*



# Références I

- [1] Constitutional ai : Harmlessness from ai feedback, 2022.
- [2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [3] Casio. <https://www.moflin.com/>.
- [4] J. P. Dupuy. *Aux origines des sciences cognitives*/. Découverte,, Paris:, 1994.
- [5] Eliza. [https://fr.wikipedia.org/wiki/Effet\\_ELIZA](https://fr.wikipedia.org/wiki/Effet_ELIZA).
- [6] EU. <https://artificialintelligenceact.eu/fr/article/3/>.
- [7] G. Fr. <https://www.comparia.beta.gouv.fr/>.
- [8] IBM. <https://www.ibm.com/fr-fr/watson>.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [11] OCDE. <https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449>.
- [12] F. Rosenblatt. Perceptron simulation experiments. *Proceedings of the IRE*, 48(3):301–309, 1960.
- [13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath : Pushing the limits of mathematical reasoning in open language models, 2024.
- [14] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- [15] A. Summit. The International Scientific Report on the Safety of Advanced AI, 2025.
- [16] TIME. <https://time.com/6247678/openai-chatgpt-kenya-workers/>.
- [17] UNESCO. <https://www.unesco.org/fr/legal-affairs/recommendation-ethics-artificial-intelligence>, 2021.
- [18] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

## Autres définitions plus récentes d'un « système d'IA » ...

### AI Act<sup>6</sup>, EU, 2024

Un système basé sur une machine qui est conçu pour fonctionner avec différents niveaux d'autonomie et qui peut faire preuve d'adaptabilité après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer des environnements physiques ou virtuels.

### OCDE<sup>11</sup>, 2024

Système d'IA : Un système d'intelligence artificielle est un système automatisé qui, pour des objectifs explicites ou implicites, déduit, à partir d'entrées reçues, comment générer des résultats en sortie tels que des prévisions, des contenus, des recommandations ou des décisions qui peuvent influencer sur des environnements physiques ou virtuels. Différents systèmes d'IA présentent des degrés variables d'autonomie et d'adaptabilité après déploiement.

### Unesco

AI systems are **information-processing technologies** that integrate **models** and **algorithms** that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments.

AI systems are designed to operate with varying degrees of autonomy by means of **knowledge modelling and representation** and by **exploiting data** and **calculating correlations**. AI systems may include several methods, such as but not limited to :

---

▶ machine learning, including deep learning and reinforcement learning ;

▶ machine reasoning, including planning, scheduling, knowledge representation and reasoning, search,