L'intelligence artificielle, de la définition à l'usage

Six entrées qui se lisent dans l'ordre : définir, apprendre, calculer, générer, appliquer, questionner. Chaque branche est développée sur une page dédiée ; les compléments et les rectifications apportés à la carte manuscrite sont signalés.

01 DÉFINIR

→ p. 3

- La définition de Minsky (1968)
- L'agent : percevoir, traiter, décider, agir, poursuivre des objectifs
- Quatre façons de définir (Russell & Norvig)
- Mythes tenaces

06 QUESTIONNER

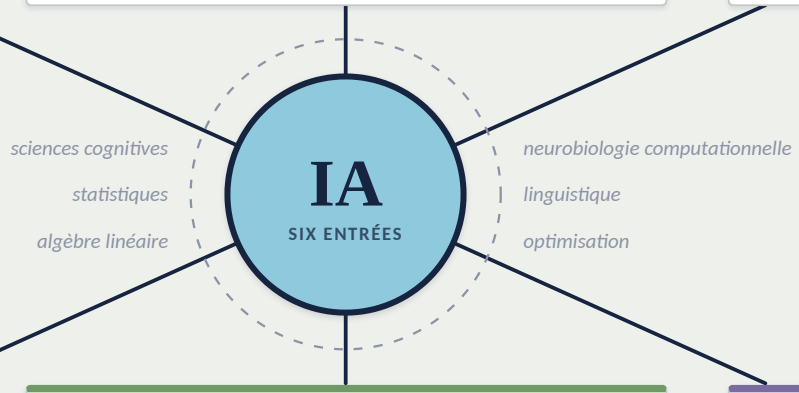
→ p. 7

- Biais, opacité, fiabilité
- Données : provenance, consentement, protection
- Coûts énergétiques et humains
- Littératie en IA

05 APPLIQUER

→ p. 7

- Fonctions : traduction, reconnaissance, prédiction, diagnostic
- Domaines, de la robotique à l'éducation
- *Affective computing* (Picard)



02 APPRENDRE

→ p. 4

- La définition de Mitchell : expérience, tâche, performance
- Supervisé · non supervisé · auto-supervisé · par renforcement
- Clustering, prédiction, jeu d'échecs

03 CALCULER

→ p. 5

- Le neurone artificiel : poids, biais, activation
- Rétropropagation de l'erreur, GPU
- Architectures : convolutifs, récurrents, Transformer
- Familles d'algorithmes

04 GÉNÉRER

→ p. 6

- Fabriquer un modèle de langue en quatre étapes
- Pré-entraînement, ajustement, alignement (RLHF)
- Ce que le modèle ne fait pas

définir · apprendre · calculer · générer · appliquer · questionner · + ajout · ≠ rectification

Quelques repères

La feuille ne portait pas de chronologie ; celle-ci situe les notions les unes par rapport aux autres.

1950

Turing pose la question « les machines peuvent-elles penser ? » et propose un test fondé sur l'imitation.

1955–1956

La proposition puis la conférence de Dartmouth fixent le terme *artificial intelligence*.

1958

Rosenblatt formalise le perceptron : une unité qui pondère des entrées.

1968

Minsky énonce la définition reprise sur la feuille, dans la préface d'un ouvrage collectif.

1986

La rétropropagation de l'erreur rend l'apprentissage des couches profondes praticable.

1995–1997

Machines à vecteurs de support ; naissance de l'*affective computing* chez Picard.

2012

AlexNet : l'apprentissage profond s'impose en vision par ordinateur.

2017

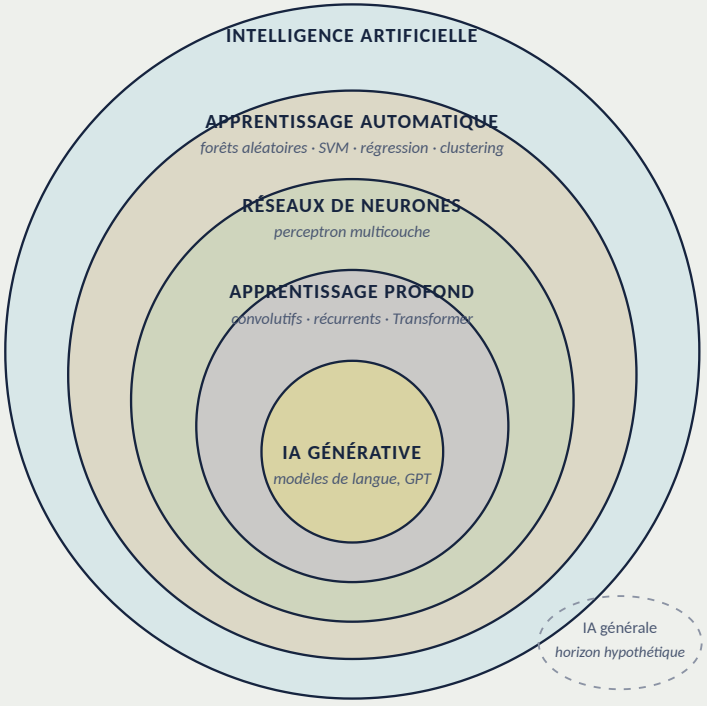
L'architecture Transformer et son mécanisme d'attention.

2020–2022

GPT-3, puis l'alignement par retour humain (RLHF) et la diffusion grand public.

Le périmètre des notions

Le point le plus substantiel : l'emboîtement des cercles, redessiné.



La feuille place « IAG » au-dehors, comme l'ensemble le plus vaste. Les deux lectures possibles du sigle conduisent à la même rectification.

Si « IAG » signifie *IA générative* : celle-ci repose sur des architectures d'apprentissage profond, qui relèvent des réseaux de neurones, eux-mêmes une famille de l'apprentissage automatique. Elle occupe donc le cercle le plus intérieur, non le plus extérieur.

Si « IAG » signifie *IA générale* : il s'agit d'une capacité visée, transversale à des tâches variées, qui n'existe pas à ce jour. Elle ne peut contenir l'intelligence artificielle, dont elle serait au mieux une réalisation. D'où le trait pointillé, hors de l'emboîtement.

La distinction mérite d'être posée explicitement en cours : c'est une confusion fréquente, et elle change la façon dont on situe les outils que les étudiants utilisent.

01
DÉFINIR

Ce que recouvre le terme

Une définition de référence, le modèle de l'agent, une typologie des approches et les idées reçues à déconstruire.

CE QU'ON APPELLE INTELLIGENCE ARTIFICIELLE

« L'intelligence artificielle, la science qui consiste à faire faire à des machines des choses qui exigeraient de l'intelligence si elles étaient accomplies par des humains. »

Minsky, préface de Semantic information processing (1968, p. v) — traduction libre. La formule circule le plus souvent reformulée en phrase complète ; la forme attestée est appositive, et l'ouvrage est un collectif qu'il a dirigé. ≠

L'IA n'est pas une entité : c'est un ensemble de techniques réunies par un objectif.

QUATRE FAÇONS DE DÉFINIR +

	HUMAINEMENT	RATIONNELLEMENT
PENSER	modélisation cognitive	lois de la pensée, logique
AGIR	test de Turing	agent rationnel — approche dominante

Russell & Norvig (2021). La feuille opposait déjà « aspect humain » et « modèle idéal, rationnel » : c'est cette grille.

UN AGENT, CINQ OPÉRATIONS

1. Percevoir l'environnement
2. Traiter l'information — collecter, capter, coder les données
3. Décider — raisonnement et apprentissage
4. Agir — exécuter la tâche, s'adapter au changement de contexte
5. Poursuivre des objectifs spécifiques

L'autonomie se mesure en degrés : qui fixe l'objectif, qui valide l'action ? +

MYTHES TENACES

- Contrôle — le système n'est pas fiable
- Productivité — tout ne s'automatise pas
- Prompt — l'utilisateur ne maîtrise pas le résultat
- Intelligence — le traitement est statistique, non compréhensif
- Apprentissage — un processus, pas un savoir
- Créativité · Discours futuristes

02
APPRENDRE

Apprendre, au sens technique

Une définition opératoire, puis quatre régimes d'apprentissage selon la nature des données et du retour.

Un programme apprend d'une expérience E, pour une classe de tâches T et une mesure de performance P, si sa performance sur T, mesurée par P, s'améliore avec E.

Mitchell (1997), traduction libre. S'y ajoutent l'apprentissage semi-supervisé et l'apprentissage par transfert — réutiliser un modèle déjà entraîné. [+]

SUPERVISÉ

Données étiquetées : entrées et réponses attendues, qu'il faut organiser, contextualiser, fournir au modèle.

But : prédire l'étiquette d'un cas nouveau.

PRÉDICTION

NON SUPERVISÉ

Données sans réponses. But : le *clustering* — diviser l'ensemble en paquets homogènes selon des critères de proximité.

Segmentation de clientèle, compression d'images satellites (forêts, désert), gestion de la relation client.

CLUSTERING

AUTO-SUPERVISÉ [+]

Le problème supervisé est engendré automatiquement à partir de données non annotées : masquer une partie, la faire prédire.

C'est le régime de pré-entraînement des modèles de langue — d'où le lien avec la branche 04 (p. 6).

PAR RENFORCEMENT

L'agent est plongé dans un environnement ; ses actions sont évaluées par une récompense.

Jeu d'échecs, jeu de go ; prolongement contemporain : l'alignement des modèles de langue.

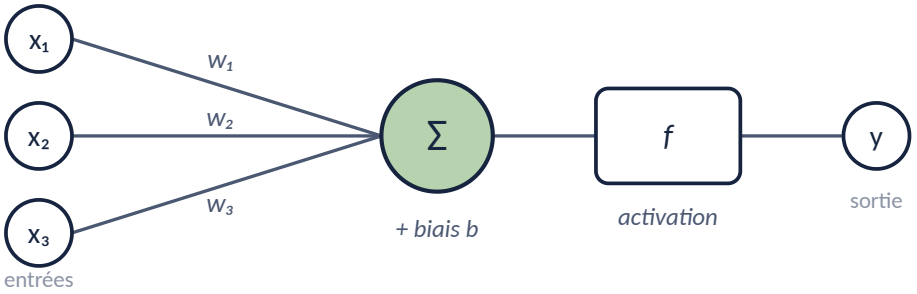
JEU D'ÉCHECS

03
CALCULER

Ce qui se passe sous le capot

L'unité de calcul élémentaire, les architectures qu'elle compose et les grandes familles d'algorithmes.

LE NEURONE ARTIFICIEL



Entrées × poids, plus un biais → somme pondérée → fonction d'activation (ReLU, sigmoïde, softmax) → sortie.

Apprendre, ici, c'est ajuster poids et biais par rétropropagation de l'erreur.

Additions et multiplications matricielles, accélérées par les processeurs graphiques — GPU, Graphics Processing Unit. ≠

Un réseau de neurones n'est pas un réseau bayésien : le second est un modèle probabiliste graphique, famille distincte. ≠

DES ARCHITECTURES SELON LA TÂCHE +

- Perceptron (1958) → perceptron multicouche
- Réseaux convolutifs : vision, reconnaissance de formes
- Réseaux récurrents, LSTM : séquences, parole
- Transformer : mécanisme d'attention (2017) — socle des modèles de langue

« Profond » désigne le nombre de couches (LeCun et al., 2015).

FAMILLES D'ALGORITHMES +

- Régression linéaire et logistique
- k-moyennes · k plus proches voisins
- Bayésien naïf et réseaux bayésiens
- Arbres de décision et forêts aléatoires
- SVM, machines à vecteurs de support
- Réseaux de neurones

04
GÉNÉRER

Fabriquer un modèle de langue +

Quatre étapes successives, puis les limites qui en découlent.

1

TOKENISATION

Le texte est découpé en unités.

→

2

PRÉ-ENTRAÎNEMENT AUTO-SUPERVISÉ

Prédire l'unité suivante sur de vastes corpus (cf. branche 02, p. 4).

→

3

AJUSTEMENT SUPERVISÉ

Sur des exemples d'instructions suivies.

→

4

ALIGNEMENT (RLHF)

Par renforcement, à partir de préférences humaines.

GPT = Generative Pre-trained Transformer, au singulier. ≠ *Le RLHF n'est pas l'entraînement lui-même : c'est une phase d'alignement qui vient après.* ≠

CE QUE LE MODÈLE NE FAIT PAS

- Il n'interroge pas une base de connaissances : il estime la suite la plus probable
- L'apprentissage en contexte n'est pas un apprentissage durable : la fenêtre de contexte s'efface
- La fluidité d'un énoncé n'est pas un indice de sa véracité

05
APPLIQUER

Usages et domaines

FONCTIONS

- Compréhension et traduction automatiques
- Reconnaissance des formes et de la parole
- Prédiction, classification
- Raisonnements automatiques (systèmes experts)
- Aide au diagnostic et à la décision
- Résolution de problèmes

DOMAINES

Robotique
Industrie
Art
Transports
Logistique
Banque
Militaire
Médecine
Cybersécurité
Droit
Éducation [+]

ÉMOTIONS, SUBJECTIF ?

L'*affective computing* vise à reconnaître, modéliser et simuler des états affectifs.

Picard (1995, 1997). La date de 1962 portée sur la feuille est son année de naissance, non celle du champ. [≠] *Reconnaître des indices d'émotion n'est pas éprouver une émotion.*

06
QUESTIONNER

Enjeux et littératie

CE QUI DOIT RESTER DISCUTABLE [+]

- Biais — un modèle reproduit les régularités de ses données d'entraînement
- Opacité — la performance ne s'accompagne pas d'une explication
- Fiabilité — énoncés fluides et faux, dits hallucinations
- Données — provenance, consentement, protection
- Coûts — énergie, matériel, travail d'annotation peu visible

LITTÉRATIE EN IA [+]

Savoir ce qu'un système peut et ne peut pas faire, reconnaître un usage pertinent, évaluer une sortie, interroger les données et les finalités.

Long & Magerko (2020) ; référentiels UNESCO (2024, 2025). En classe, l'enjeu porte moins sur l'outil que sur le jugement exercé à son propos.

Références

Références vérifiées une à une auprès des éditeurs ou de Crossref. Lorsque la pagination n'a pas pu être confirmée sur une source primaire — actes NeurIPS notamment — elle est omise au profit de l'URL stable.

FONDATIONS

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904

Minsky, M. (1968). Préface. Dans M. Minsky (Éd.), *Semantic information processing* (p. v). MIT Press.

Russell, S., & Norvig, P. (2021). *Intelligence artificielle : une approche moderne* (4e éd. ; L. Miclet, F. Popineau & C. Cadet, trad.). Pearson France.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

APPRENTISSAGE ET ARCHITECTURES

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

MODÈLES DE LANGUE

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Dans H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Éds.), *Advances in neural information processing systems 33*. Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Dans I. Guyon et al. (Éds.), *Advances in neural information processing systems 30*. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Dans S. Koyejo et al. (Éds.), *Advances in neural information processing systems 35*. Curran Associates. https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Dans I. Guyon et al. (Éds.), *Advances in neural information processing systems 30*. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

AFFECTS, ENJEUX ET LITTÉRATIE

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Dans *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (p. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Dans *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (p. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Picard, R. W. (1995). *Affective computing* (M.I.T. Media Laboratory Perceptual Computing Section Technical Report no 321). MIT Media Laboratory. https://affect.media.mit.edu/pdfs/95.picard.pdf

Picard, R. W. (1997). *Affective computing*. MIT Press.

UNESCO. (2025). *Référentiel de compétences en IA pour les apprenants* (C. Dhaussy, trad.). UNESCO. https://doi.org/10.54675/NXRY6511

UNESCO. (2025). *Référentiel de compétences en IA pour les enseignants* (C. Dhaussy, trad.). UNESCO. https://doi.org/10.54675/BQZD8407

Réserves de vérification. Trois éléments n'ont pas pu être confirmés sur une source primaire : la page exacte de la définition de Mitchell (1997), couramment donnée comme p. 2 ; la pagination des actes NeurIPS, omise ici ; la ponctuation exacte de la préface de Minsky (1968, p. v), dont la localisation est en revanche attestée. Un passage de la feuille reste illisible : l'annotation surlignée en haut à gauche, qui accompagnait la liste des mythes.