



Métacognition



15/10/25

Deleglise Matthieu - Dpt biotechnologies
Garba Hamidou - M2 Intelligence Artificielle
Loquet Jérémie - M2 Neuroscience Computationnelle

Sommaire

- 01** Introduction : Historique et Définition
- 02** En pratique : La métacognition appliquée aux IAs
- 03** L'IA qui se sait penser
- 04** Conclusion



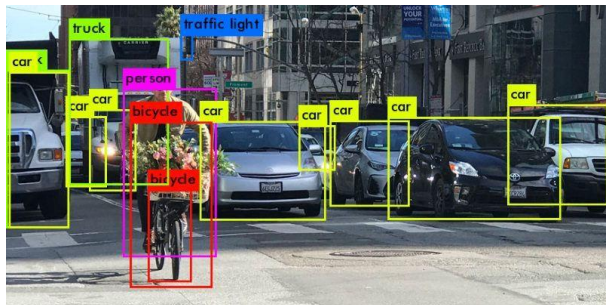


01

Introduction



Introduction



Reconnaître des images



Traduire des langues



Générer du texte





Introduction

Question :

Comment faire pour que l'IA comprenne quand elle se trompe, et qu'elle puisse s'améliorer d'elle-même ?



Introduction

Métacognition

Une approche qui, pour l'intelligence artificielle, vise à rendre ses processus internes plus observable.



Historique

Métacognition

- ❖ À l'origine : recherches sur le cognitivisme, mais dans un contexte de recherche sur la pédagogie active (implication de l'apprenant dans son propre apprentissage)
- ❖ C'est John H. Flavell qui a créé le mot dans les années 1960.



Définition

Métacognition

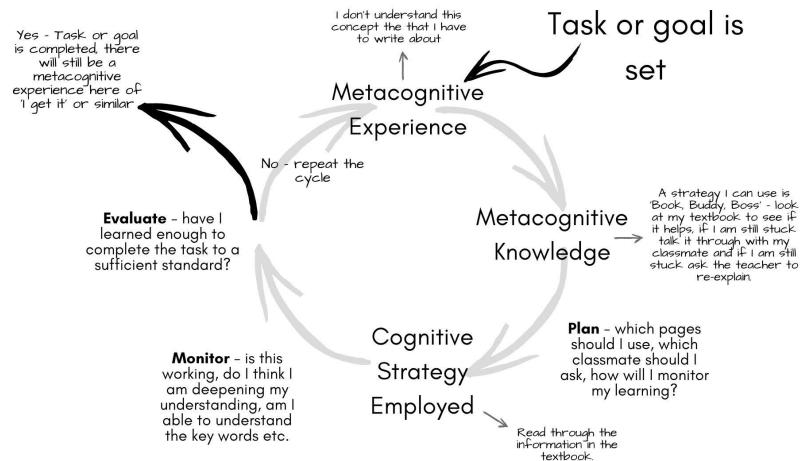
- ❖ Méta : sur, au-dessus
- ❖ Cognition : acte de connaître ou d'apprendre
- ❖ Essentiellement, la métacognition est une connaissance des processus cognitifs.

Définition (Flavell)

Métacognition

« La métacognition fait référence à la connaissance qu'une personne a de ses propres processus cognitifs, ainsi qu'à la capacité de les surveiller et de les réguler. »

(Flavell, J. H. (1979). *Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry*. *American Psychologist*, 34(10), 906-911.)



Psychology in the Classroom
An Introduction to Metacognition

<https://changingstatesofmind.libsyn.com/an-introduction-to-metacognition>



Définition (Chartier et Lautrey)

Métacognition

« la connaissance et le contrôle qu'un système cognitif peut avoir de lui-même et de son propre fonctionnement »

(Chartier et Lautrey, 1992, p. 29)

https://oraprdnt.uqtr.quebec.ca/Gsc/Portail-ressources-enseignement-sup/documents/PDF/metacognition_notes_de_cours.pdf

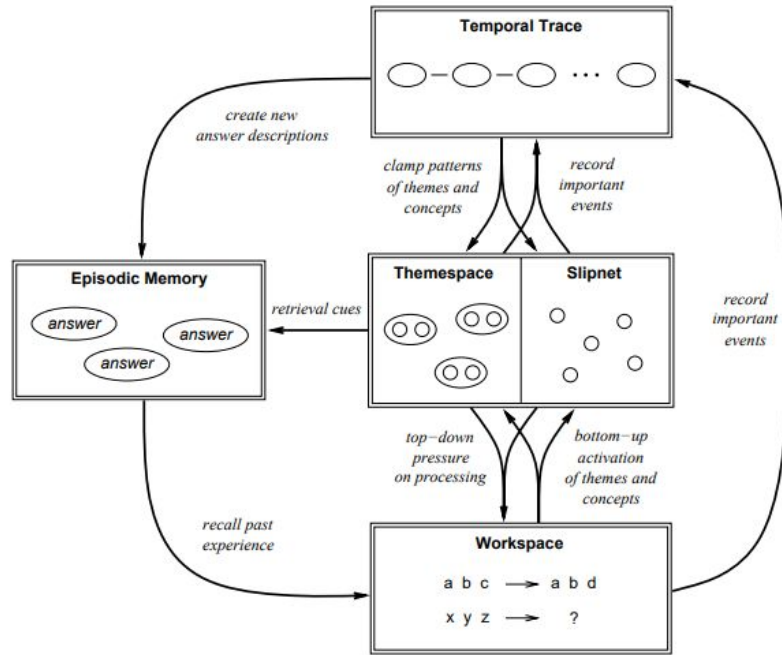


02

Métacognition en pratique



Metacat et l'auto-observation



- Copycat : un modèle capable de percevoir les **patterns dans les données en entrée** pour réaliser des analogies ;
- Metacat : un modèle capable de percevoir les **patterns dans le processing des données** pour comprendre comment il est arrivé à une réponse.
- **Trace temporel** : enregistrement séquentiel des raisonnements effectués et recherche de patterns ;
- **Mémoire épisodique** : répertoire d'expériences ;
- Capable de justifier une réponse **qu'il n'a pas créé** ;
- Capacité de **créativité**.



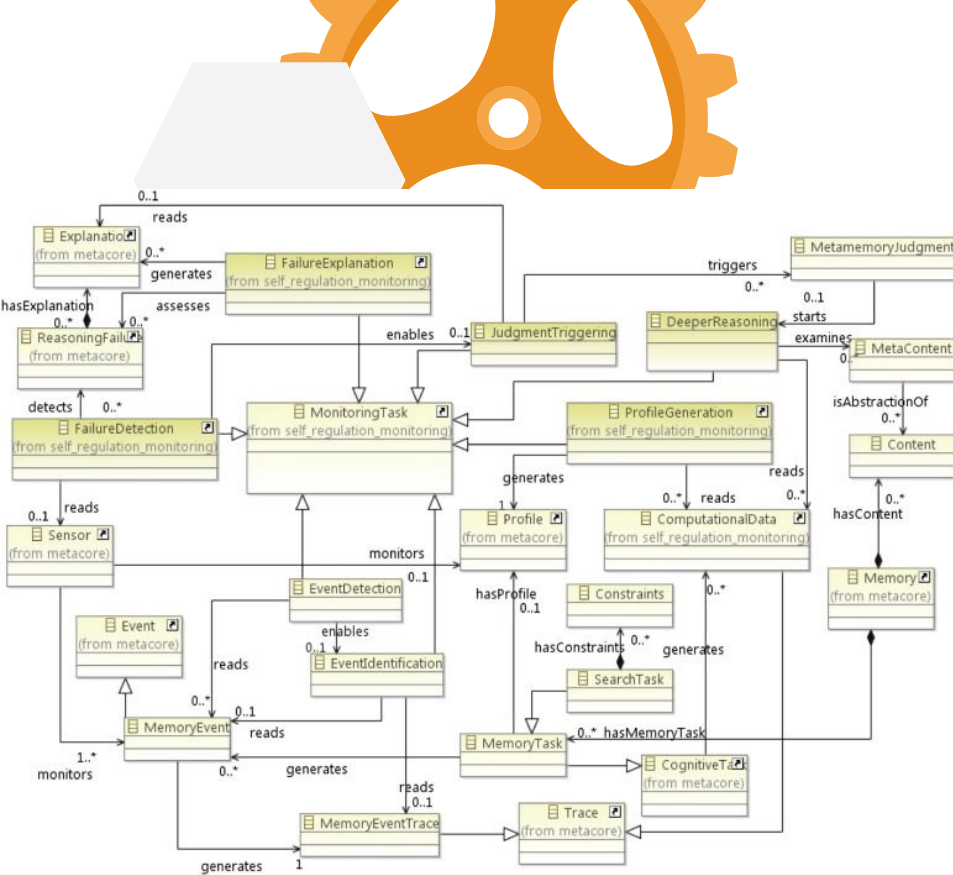
Méta-modèle général

Caro et al. (2014)* présente un méta-modèle **applicable à différents types de problèmes** et combinant les concepts utilisés par les modèles métacognitifs précédent.

Il se base sur trois principaux composants :

- **Régulation de soi**
- **Méta-mémoire**
- **Méta-compréhension**

Chaque concept permet le **monitoring** ou le **contrôle** de ces trois capacités.



Metamemory self-regulation mechanism

*Manuel F. Caro, Darsana P. Josyula, Michael T. Cox, Jovani A. Jiménez, Design and validation of a metamodel for metacognition support in artificial intelligent systems, *Biologically Inspired Cognitive Architectures*, Volume 9, 2014, Pages 82-104, ISSN 2212-683X, <https://doi.org/10.1016/j.bica.2014.07.002>.

L'IA métacognitive vue par Pitrat

“Nous trouverons de nouvelles connaissances pour découvrir des connaissances à l'aide des connaissances pour découvrir des connaissances déjà existantes”



Métaconnaissances et IA symbolique

- Dédurre la meilleure méthode pour résoudre un problème ;
- Inférer les règles **les plus à même de mener au succès.**



La réflexivité des métaconnaissances

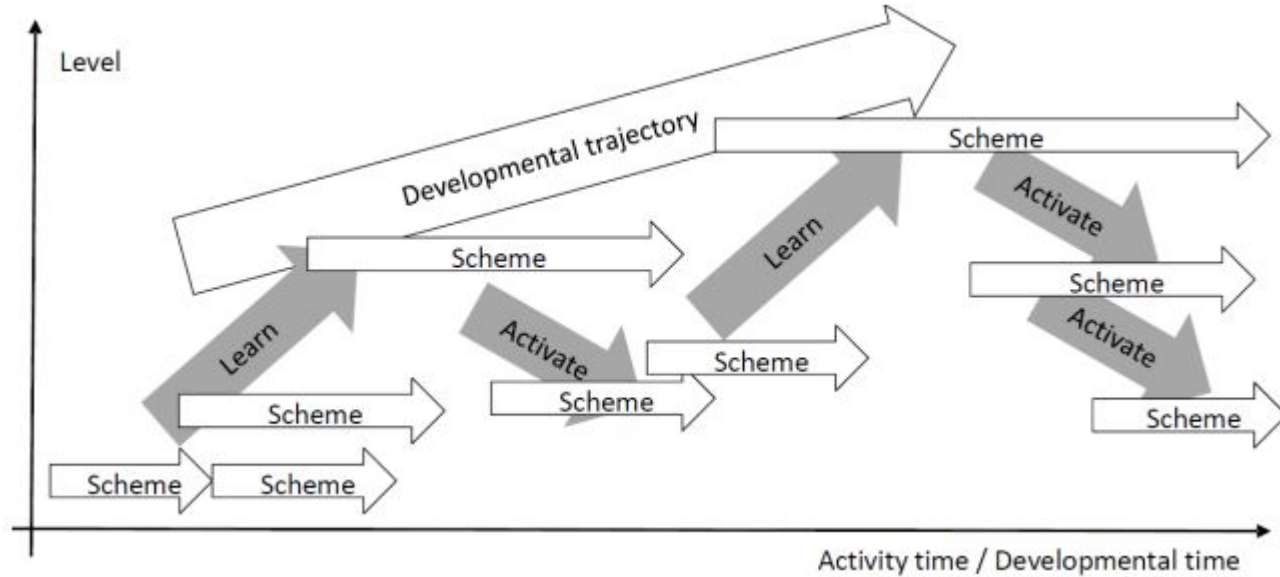
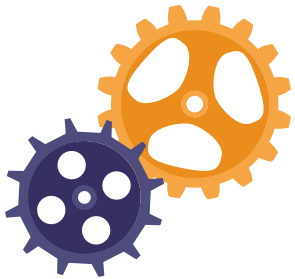
- Les métaconnaissances sont des connaissances sur l'utilisation des connaissances.
- Peuvent s'appliquer à elle-même (**réflexivité**);
 - Effet **boule-de-neige** positif.



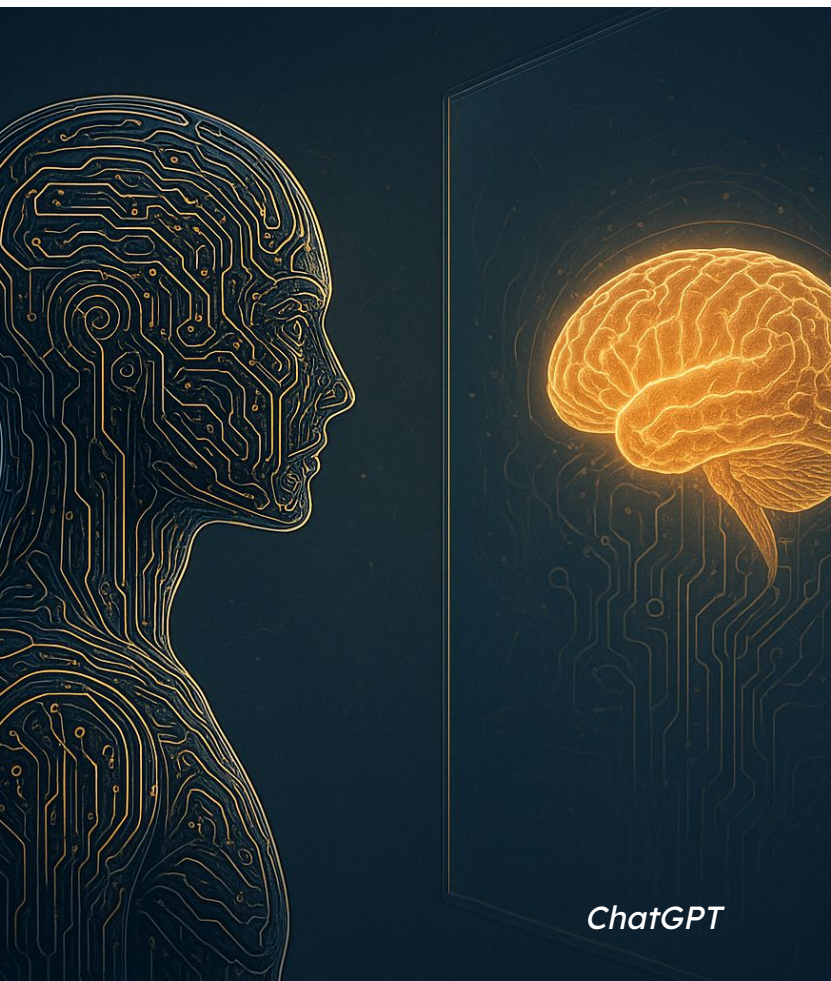
Amorçage de l'IA et auto-amélioration

- **Amorcer** l'IA avec un socle de méta-connaissances initiales.
- Création de niveaux d'abstractions supérieurs au fur et à mesure de l'apprentissage de nouvelles connaissances





Apprentissage métacognitif : un concept bien illustré par l'IA développementale



03

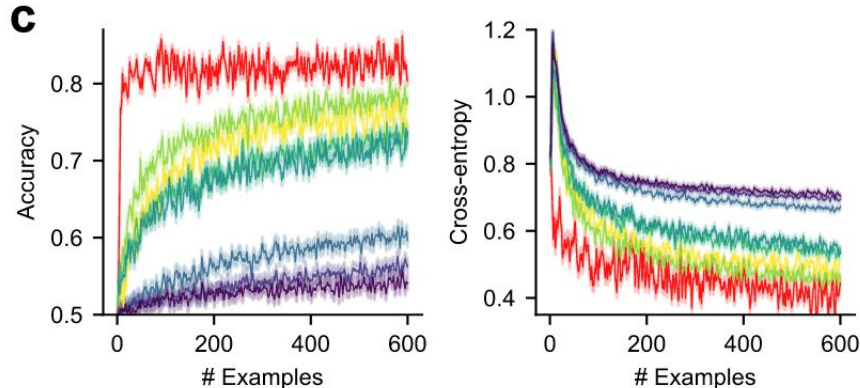
**L'IA qui se
sait penser**



Le tournant neuroparadigmatique

Ji-An et al., 2025:

- Identification de corrélats de certitude ou de cohérence → Indices métacognitifs ;
- Propriété émergente → “**Signal de soi**” **inconscient**.



LR vs. PCs

PCs sur vecteurs d'activation des couches internes.

Cross-entropy = difference entre prédiction et vérité

Li Ji-An, Hua-Dong Xiong, Robert C. Wilson et al., Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. 2025. arXiv. DOI: <https://doi.org/10.48550/arXiv.2505.13763>

La continuité



**Ackerman,
2025**

Approche comportementale

Monitoring métacognitif

Forme d'introspection
statistique

Penser sur sa pensée ?

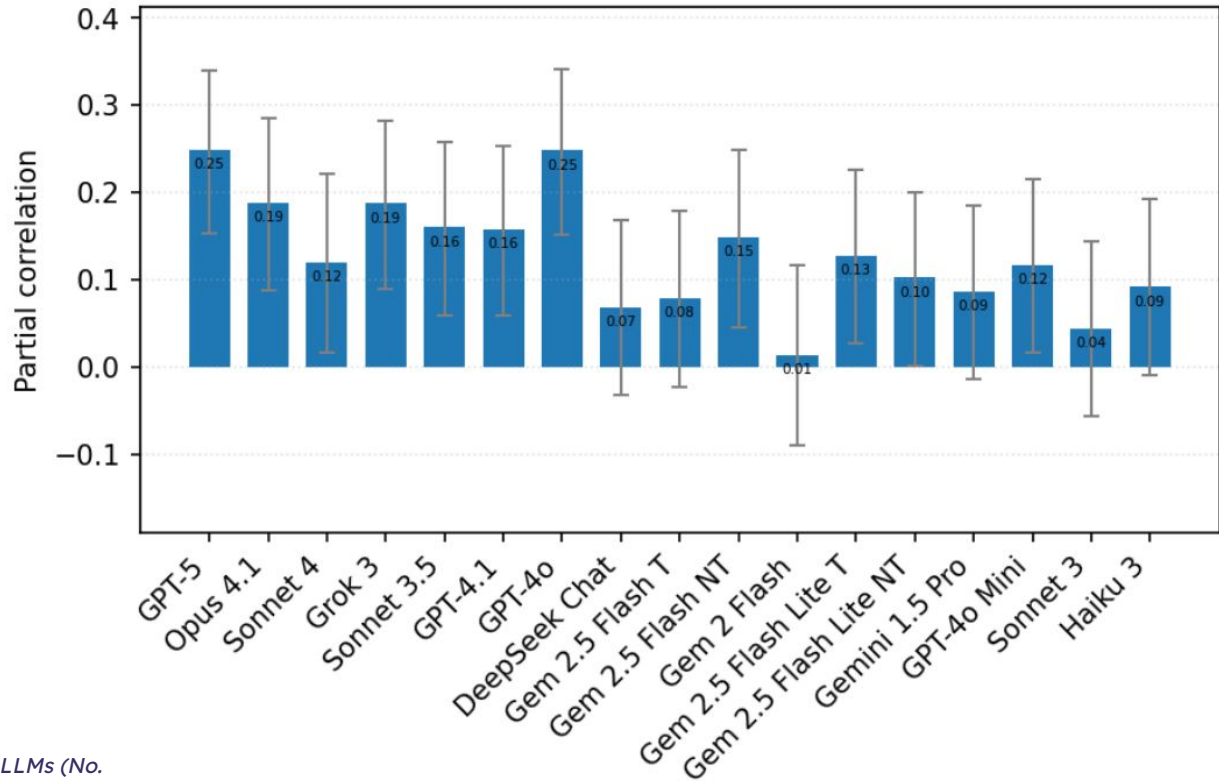


Table 5: Self-regulation: Success rate of the various agents on the novel tasks with less-capable LLMs, based on additional adaptation with five episodes.

Agent	Mistral-7B-OpenOrca	OpenELM-3B-Instruct
ReAct	23%	4%
Reflexion	27%	9%
MUSE	58%	55%

**Ackerman,
2025**

Approche comportementale
Monitoring métacognitif

Forme d'introspection
statistique

Penser sur sa pensée ?

**Valiente & Pilly,
2024**

*Metacognition for Unknown
Situations and
Environments*

Principe d'ingénierie

Système à deux niveaux

Fonction opérationnelle de
la performance



Valiente, R., & Pilly, P. K. (2024). *Metacognition for Unknown Situations and Environments (MUSE)* (No. arXiv:2411.13537). arXiv.
<https://doi.org/10.48550/arXiv.2411.13537>

La



Acker
2025

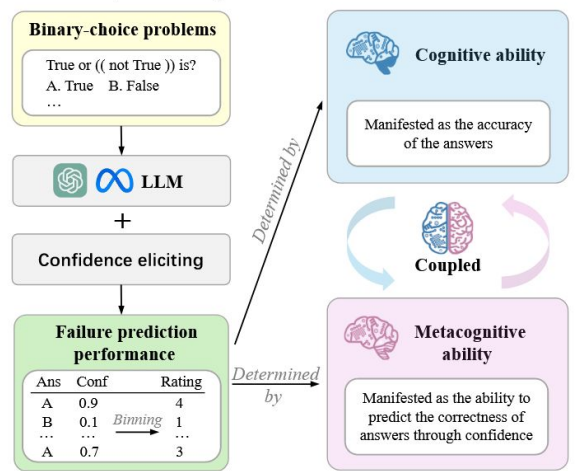
Appro

Monit

Forme
statist

Pense

Step 1: Tasking LLM with Failure Prediction

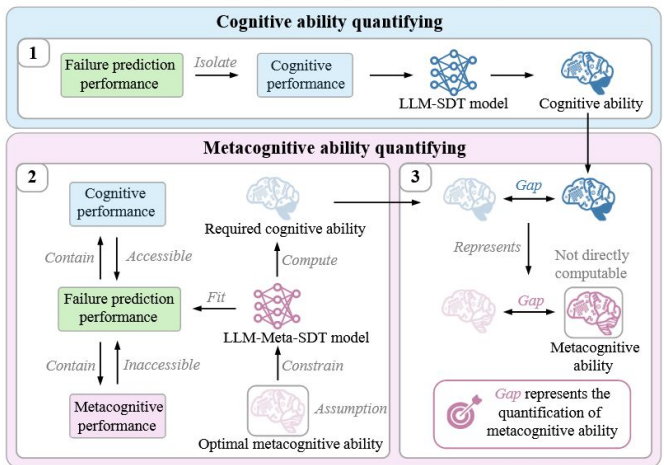


Wang, G., Wu, W., Ye, G., Cheng, Z., Chen, X., & Zheng, H. (n.d.). Decoupling Metacognition from Cognition: A Framework for Quantifying Metacognitive Ability in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*



Wang et al.,
2025

Step 2: Decoupling Metacognition from Cognition



or Unknown

erie

niveaux

onnelle de

Decoupling Metacognition from Cognition

Une métrique pour Métacognition vs. Cognition

Indice d'efficacité métacognitive

Vers une science quantitative de la MC artificielle

La continuité



Ackerman, 2025

Approche comportementale

Monitoring métacognitif

Forme d'introspection
statistique

Penser sur sa pensée ?



Valiente & Pilly, 2024

*Metacognition for Unknown
Situations and
Environments*

Principe d'ingénierie

Système à deux niveaux

Fonction opérationnelle de
la performance



Wang et al., 2025

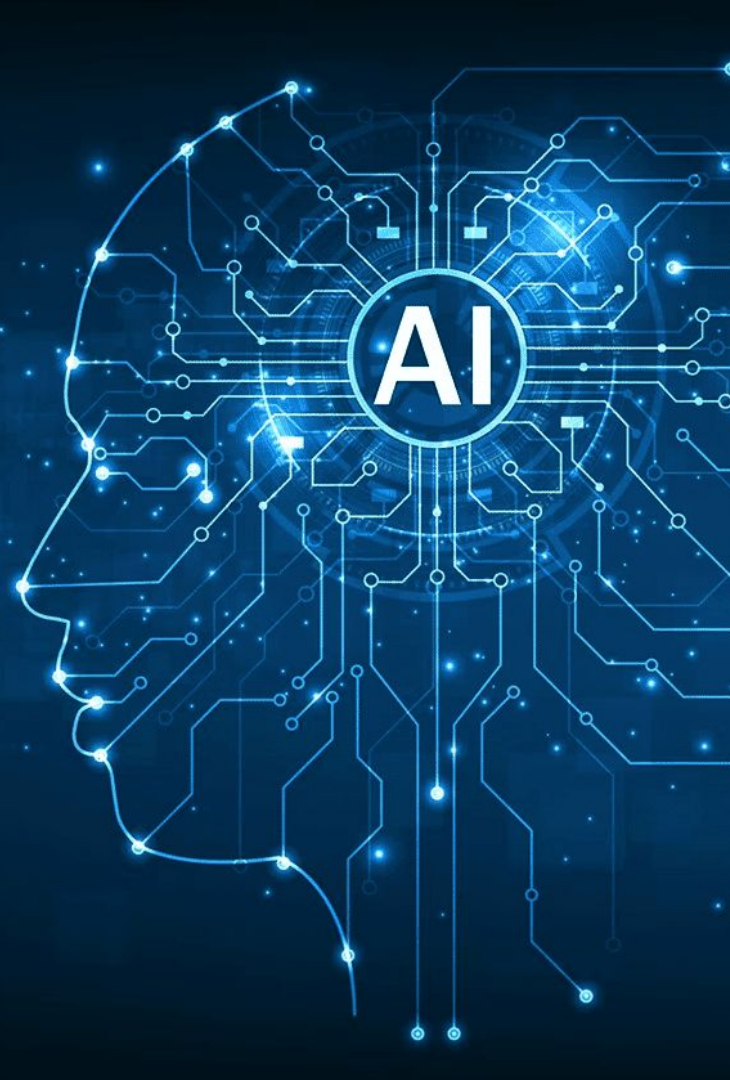
*Decoupling Metacognition
from Cognition*

Une métrique pour
Métacognition vs. Cognition

Indice d'efficacité
métacognitive

Vers une science
quantitative de la MC
artificielle





04

Conclusion



La métacognition et l'IA



Une idée ancienne ...

- Volonté de systèmes capables de s'améliorer depuis Dartmouth ;
- Étude de la métacognition chez l'humain et l'IA depuis une cinquantaine d'années.



... avantageuse ...

- **Adaptation à de nouveaux problèmes** sans ré-entraînements ;
- Système capable de s'auto-améliorer sans intervention humaine.



... encore naissante ...

- Modèles métacognitifs plutôt appliqués à des systèmes spécialisés non généralisables ;
- LLM capable de montrer leur raisonnement **sans vraiment le comprendre.**



... qui pose question.

- **Vision neuroscientifique** : nécessite l'implémentation des fonctions cognitives humaines ;
- **Vision ingénieure** : réalisable par la simple imitation de ces fonctions cognitives.

Merci pour votre attention !

Si vous avez des questions, n'hésitez pas :)

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

15/10/25

Deleglise Matthieu - Dpt biotechnologies
Garba Hamidou - M2 Intelligence Artificielle
Loquet Jérémie - M2 Neuroscience Computationnelle



Ressources

- Manuel F. Caro, Darsana P. Josyula, Michael T. Cox, Jovani A. Jiménez, Design and validation of a metamodel for metacognition support in artificial intelligent systems, Biologically Inspired Cognitive Architectures, Volume 9, 2014, Pages 82-104, ISSN 2212-683X, <https://doi.org/10.1016/j.bica.2014.07.002>.
- Pitrat Jacques. Des métaconnaissances pour des systèmes intelligents. In: Quaderni, n°25, Printemps 1995. Intelligence artificielle et entreprise : l'entreprise intelligente ? pp. 29-42. DOI : <https://doi.org/10.3406/quad.1995.1110>
- Li Ji-An, Hua-Dong Xiong, Robert C. Wilson et al., Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. 2025. arXiv. DOI : <https://doi.org/10.48550/arXiv.2505.13763>
- James B. Marshall, A Self-Watching Cognitive Architecture for Analogy-Making and High-Level Perception. Philosophy thesis. Indiana University : Department of Computer Science and the Cognitive Science Program. November 1999, 306p. <https://science.slc.edu/jmarshall/metacat/dissertation.pdf>
- Notes de cours de François Guillemette, professeur associé en éducation et en communication à l'UQTR, https://oraprdnt.uqtr.quebec.ca/Gsc/Portail-ressources-enseignement-sup/documents/PDF/metacognition_notes_de_cours.pdf
- Ackerman, C. (2025). Evidence for Limited Metacognition in LLMs (No. arXiv:2509.21545). arXiv. <https://doi.org/10.48550/arXiv.2509.21545>
- Valiente, R., & Pilly, P. K. (2024). Metacognition for Unknown Situations and Environments (MUSE) (No. arXiv:2411.13537). arXiv. <https://doi.org/10.48550/arXiv.2411.13537>
- Wang, G., Wu, W., Ye, G., Cheng, Z., Chen, X., & Zheng, H. (n.d.). Decoupling Metacognition from Cognition: A Framework for Quantifying Metacognitive Ability in LLMs. Proceedings of the AAAI Conference on Artificial Intelligence

