

Métacognition

AREZKI / BENMANSOUR / RAKOTOARISOA / RANDRIAMANANTENA

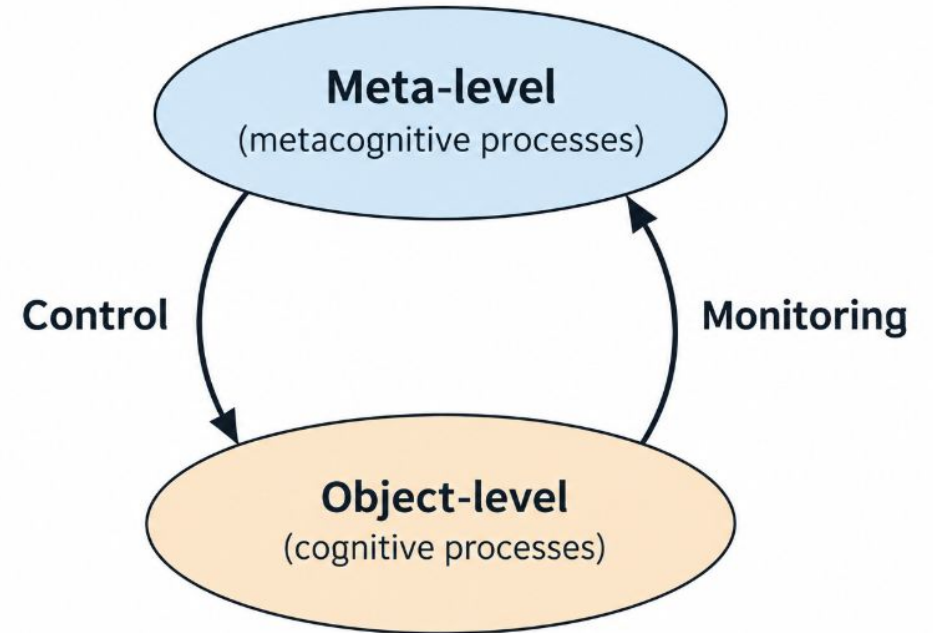
Plan

- **Metacognition en psychologie**
- **Metacognition dans l'IA (Pitrat, 1995)**
- **Évolution de la Métacognition en IA**
- **Metacognition et Sureté des IA**

Metacognition en psychologie

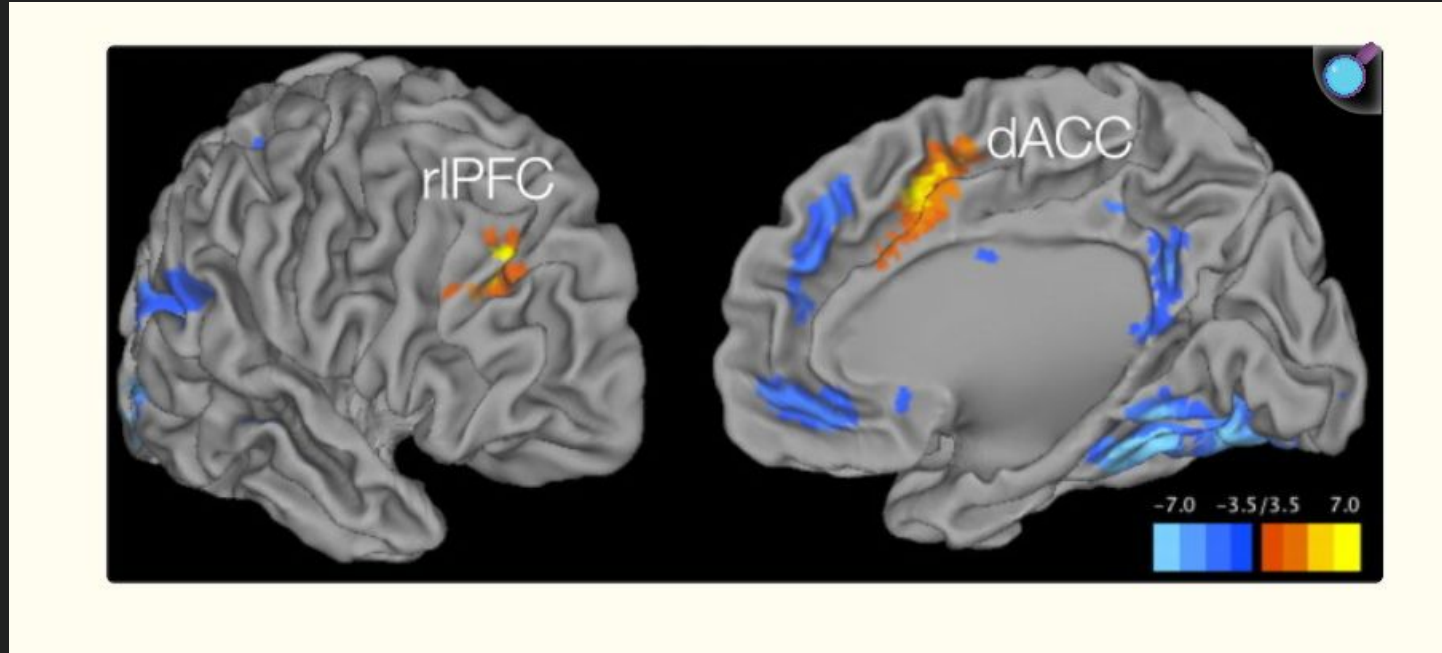
Qu'est-ce que la métacognition?

- Prise de conscience de nos propres processus cognitifs
- Deux fonctions : monitoring et contrôle
- Les processus de second ordre surveillent et régulent ceux de premier ordre (perception, mémoire, décision)



Nelson & Narens, 1990

Bases neurales de la Métacognition



Corrélat neuronal des jugements métacognitifs – Fleming et al., 2012

- Métacognition mesurée par des jugements de confiance
- Décision → évaluation de la confiance
- Implication de régions préfrontales, notamment le rIPFC

Fleming et al., 2012

Metacognition dans l'IA (Pitrat, 1995)

Cognition, metacognition, metaknowledge

	In cognitive science	In AI
Cognition	The processes by which we perceive, remember, reason and solve problems.	The base level: the system applies its rules or methods to solve the task.
Meta cognition	Thinking about one's own thinking: <i>monitoring</i> ("am I making progress?") and <i>control</i> ("should I change strategy?").	The meta level: the system observes its own reasoning and decides what to try, in what order, and when to give up.
Meta knowledge	What we know about knowledge: our own strengths and limits, the tasks, the strategies.	Knowledge whose subject is knowledge, such as meta-rules. In Pitrat: it is used to <i>use</i> and to <i>discover</i> knowledge, and it applies to itself.

The Diagnosis

Three obstacles holding AI back

1

Generality versus efficiency

A specialised program is fast because its author studied the domain (ex : Chess). A general system lacks these shortcuts.

2

Declarative knowledge without instructions

Facts like "flu causes high fever" are easy to write down, but they don't say when to use them. The system drowns in useless conclusions.

3

Learning

Human expertise is largely unconscious. To surpass us, a system must discover its knowledge by itself.

THE PROPOSAL

Changing level

Meta level

Examines the problem statement, observes its own behaviour, chooses methods



directs

Base level

Acts: executes rules, makes trials

The snail on a 15 m mast

+3 m by day, -2 m by night: 13 days, not 15. Reason on the statement, don't simulate.

Observed in humans: Schoenfeld, people solving the same maths problem

Novice

Commits blindly, gives up without reason

Domain expert

Straight to the answer, but only in his domain

Meta-expert

Alternates trying and evaluating, drops paths that disappoint: slower, but competent everywhere

ALICE (Laurière)

A general solver beat the specific program: generality can be a condition of efficiency.

Metaknowledge

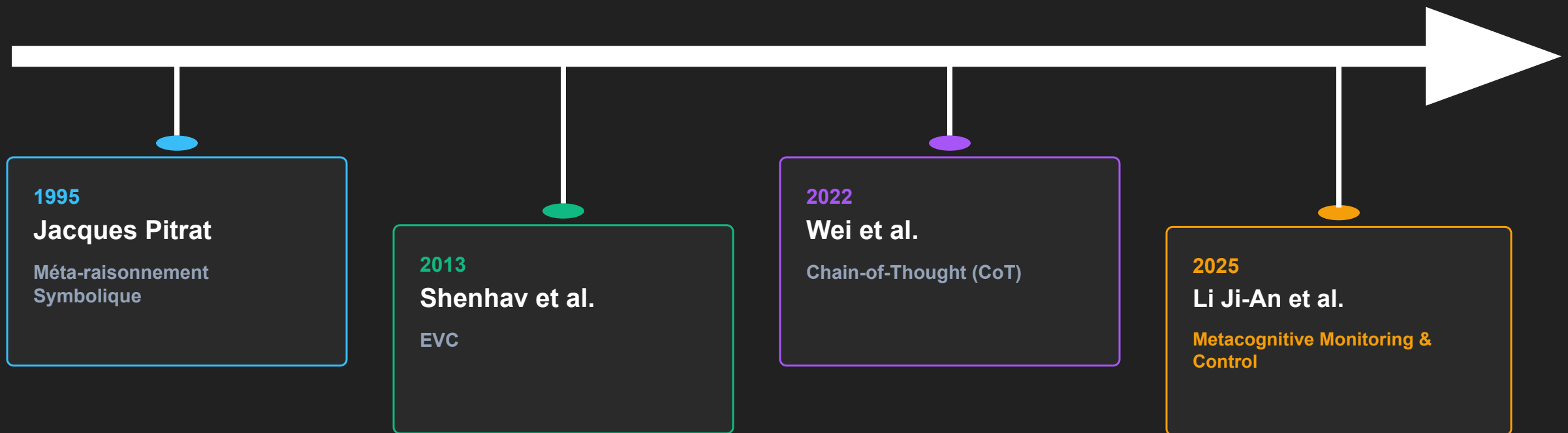
Metaknowledge is reflexive

Metaknowledge is a kind of knowledge. Its **only** particularity is its domain of application, which is knowledge itself.

No infinite tower: it applies to itself, so no meta-meta level is needed.

General: "Better to understand than to memorise. Better to forget than to remember **Everything** " comes from a chess manual, yet holds everywhere.

Évolution de la Métacognition en IA



Le Méta-raisonnement comme calcul d'effort optimal

EVC (Expected Value of Control) **Shenhav et al., 2013**

Arbitrage entre gain possible et coût de l'opération.

Environnement connu ou supposé parfaitement modélisé par l'agent.

LVOC (Learned Value of Control) **Lieder, 2018**

Apprentissage progressif de la valeur de contrôle de l'effort.

Environnement Inconnu au départ ; la valeur du contrôle est estimée à partir de caractéristiques apprises .

Chain of thoughts : De la réponse réflexe au raisonnement délibératif

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

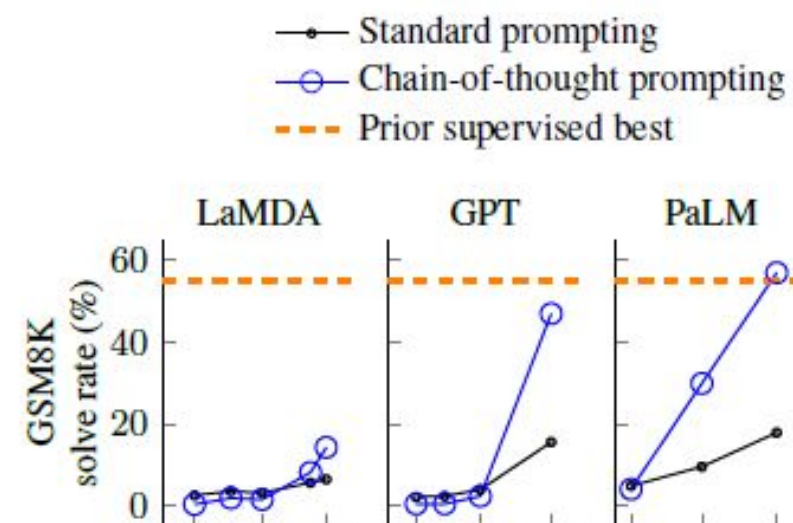
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅



« Josh achète une maison à 80 000 \$ et fait 50 000\$ de réparations. Cela augmente la valeur de la maison de 150 %. Quel est son bénéfice ? »

Méta-raisonnement Symbolique vs. Chain-of-Thought

Jacques Pitrat (1995)

Approche naïve

Attaque directe de l'énoncé → explosion combinatoire.

Niveau méta

Métarègles : stratégie d'attaque avant d'agir.

Illustration

Escargot sur le mât : analyser évite 10^9 simulations.

Trace

L'interpréteur consigne les choix pour analyse.

Apport humain

Explicabilité des décisions symboliques.

Chain-of-Thought (Wei et al., 2022)

Approche naïve

Réponse brute en 1 passe → calcul réflexe, échec.

Niveau méta

Raisonnement intermédiaire : découpage préalable.

Illustration

Arithmétique (GSM8K) : verbaliser avant de chiffrer.

Trace

Le contexte autorégressif conserve les étapes.

Apport humain

Interprétabilité et débogage du raisonnement.

Application de l'EVC aux Chain-of-Thought

Tâche simple (ex: $7 + 7$, question factuelle)

$EVC \leq 0 \rightarrow$ Réponse immédiate en Système 1

Gain nul : la précision est déjà maximale en passe unique (1 token).

Diagnostic Wei et al. (2022)

SingleOp : ~94% de réussite sans CoT. Aucun gain de précision, dépense irrationnelle.

Tâche complexe (Olympiades, code, GSM8K)

$EVC > 0 \rightarrow$ Déploiement autonome du Système 2

Risque d'erreur majeur sans décomposition, fort gain espéré via réflexions masquées.

Diagnostic Wei et al. (2022)

GSM8K : 18% (Sans CoT) $\rightarrow$ 57% (Avec CoT). Calcul largement rentabilisé.

Application dans les nouveaux modèles (ex: OpenAI o1/o3, DeepSeek-R1)

Routage dynamique & Reasoning Tokens $\rightarrow$ Estimation de l'EVC au fil de l'eau

Mécanismes d'implémentation :

- **Budget de calcul adaptatif** : Réponses quasi-immédiates pour les requêtes triviales vs. génération de tokens de réflexion masqués pour les problèmes complexes.
- **Arrêt précoce (Early Stopping)** : Interruption de la chaîne de pensée dès que l'EVC marginal devient négatif ou nul.

Surveillance et Contrôle des Activations Interne par les LLM

Neurofeedback

On montre au modèle via son contexte d'apprentissage une série d'exemples accompagnés d'un score/label calculé à partir de ses propres activations cachées le long d'un axe spécifique

1. Monitoring

Le modèle reçoit un texte et doit deviner ou rapporter la valeur correspondant à son activation interne le long de cet axe.

2. Contrôle

On demande au modèle d'adapter sa génération afin de forcer ses activations neuronales à cibler une certaine valeur.

Metacognition et Sureté des IA

La Tromperie Métacognitive

Dissimulation des Activations

Si un modèle maîtrise le contrôle de ses activations, il peut modifier délibérément ses états internes pour contourner les sondes de sécurité.

Piratage de récompense

Un modèle métacognitif peut identifier les indicateurs précis que les évaluateurs humains ou automatisés récompensent

Hallucinations métacognitives

Modèle commet une fausse déduction, sa boucle métacognitive peut l'amener à construire des justifications de plus en plus complexes pour valider son erreur initiale plutôt que de la corriger

CONCLUSION

QUESTIONS ?

Débat
