Synthèse Atelier Métacognition

Oct. 2025 DELEGLISE M., HAMIDOULLAH GARBA H., LOQUET J.

1 Présentation

Concept inventé dans les années 1970 par John Flavell, la métacognition désigne la connaissance qu'une personne a de ses propres processus cognitifs ainsi que la capacité de les surveiller et de les réguler (Flavell, 1979). C'est un niveau *au-dessus* de la cognition : elle permet à un système d'observer et de contrôler son propre fonctionnement (Guillemette, 2012). Dans le domaine de l'intelligence artificielle (IA), la métacognition vise à rendre les processus internes plus observables pour lui permettre l'auto-amélioration.

En pratique, elle a d'abord servi à perfectionner les systèmes d'IA symboliques. Ainsi, *Metacat* ajoute des capacités métacognitives à un modèle d'analogies pour lui permettre d'observer des motifs dans son raisonnement, de réutiliser ces processus dans d'autres problèmes similaires, de se corriger, d'expliquer son raisonnement, voire de créer de nouveaux problèmes (Marshall, 1999). Ce genre de modèle restait toutefois très spécialisé. Plus tard, *MISM* a regroupé divers concepts de monitoring, de contrôle pour la régulation de soi, de métamémoire et de métacompréhension dans un métamodèle applicable à différents contextes. Il fonctionne notamment grâce à la détection d'erreurs cognitives et la proposition de nouvelles stratégies pour guider la réflexion vers une meilleure solution (Caro et al., 2014). Plus globalement, Pitrat voyait dans la métacognition le moyen d'amorcer des IA capables de se développer par elles-mêmes : un niveau de base exécuterait les tâches, tandis qu'un niveau métacognitif observerait et contrôlerait ce niveau ainsi que lui-même pour accroître leur efficacité, créant des couches de complexité croissante. Cette architecture repose sur la *réflexivité* des métaconnaissances : une métaconnaissance peut s'appliquer à elle-même, si bien que mieux utiliser les connaissances permet aussi d'améliorer les métaconnaissances, produisant un effet *boule-de-neige* (Pitrat, 1995).

Plus récemment, il a été montré que les LLMs pouvaient observer leurs activations internes et identifier des corrélats de certitude ou de cohérence (Ji-An et al., 2025). La métacognition ne serait alors plus une couche ajoutée, mais une propriété émergente du réseau lui-même. Ackerman (2025) a montré qu'ils pouvaient aussi juger la fiabilité de leurs réponses, sans toutefois exploiter cette information pour s'améliorer (Ackerman, 2025). Un indice, le DMC (*Decoupling Metacognition from Cognition*), permet désormais de distinguer la capacité d'un modèle à donner une bonne réponse de celle de savoir qu'il la connaît (Wang et al., 2025), rendant les capacités métacognitives mesurables.

Ces capacités ont également été intégrées à *MUSE* pour rendre des agents adaptatifs : ils apprennent à prédire leurs propres compétences face à une tâche nouvelle et à choisir la stratégie la plus prometteuse (Valiente et al., 2024). Avec ces avancées, la métacognition en IA n'imite plus la pensée humaine : elle devient un principe d'autorégulation computationnelle. De la surveillance des activations internes aux architectures réflexives mesurables, elle devient un outil d'adaptation et d'évaluation – et peut-être, un jour, le fondement d'une forme de conscience artificielle.

2 Débat

2.1 Philosophie

Dans un premier temps, les participants se sont interrogés sur la nécessité de doter les IA de capacités métacognitives. Certains y ont vu un moyen efficace pour leur permettre de corriger leurs erreurs en modifiant leur mode de raisonnement, d'autres ont craint qu'une mauvaise adaptation ne produise de nouvelles erreurs, annulant l'effet boule de neige de la métacognition. D'où l'importance d'une réelle capacité à juger la fiabilité des réponses, aspect fondamental du monitoring métacognitif. La discussion s'est ensuite tournée vers les LLMs : leurs réponses statistiques, associées à des probabilités, constituent déjà une forme d'évaluation. Pourquoi, dès lors, aller plus loin ? L'exemple de *ChatGPT* créant de nouvelles charades défectueuses a fourni la réponse : il doit être capable non seulement de remarquer l'erreur, mais aussi de la corriger. Cela suppose qu'il conserve la trace de son raisonnement (*chain-of-thought*), et sache l'exploiter pour guider sa recherche. De plus, pour avoir une vraie métacognition, un aspect manquant grandement à des modèles est le principe de réflexivité : en plus de pouvoir observer ce qu'il se passe, ils devraient aussi pouvoir observer ce processus d'observation lui-même, et ainsi de suite.

2.2 Pragmatisme

Bien que l'idée d'une IA capable d'apprendre en autonomie et de développer des capacités de plus en plus complexes et efficaces puisse être scientifiquement attirante, elle a aussi été considérée comme effrayante pour plusieurs membres de l'assistance. Le fait qu'il puisse arriver un moment où l'humanité ne soit plus capable de suivre les IA "fortes", invite à se demander si l'on devrait vraiment continuer dans cette direction. Les IA actuelles sont développées pour nous aider. Il est donc légitime de se demander pourquoi aller plus loin si l'on ne peut plus les comprendre et que nous ne servons plus à rien. Pourtant, depuis des millénaires, l'Homme éduque les générations suivantes et les rend meilleures, voire hors de portée lors de grands changements sociétaux/technologiques. Toute l'assistance n'a cependant pas été d'accord avec ce point, pensant plutôt que les parents éduquent leurs enfants du mieux possible pour leur permettre de survivre, voire de bien vivre, plutôt que pour les supplanter. Cette partie du débat s'est donc terminée sur une question importante à laquelle notre société devra répondre pour guider le développement des IA : cherchons-nous à créer des IA pour nous aider à comprendre le monde, ou pour leur permettre de mieux comprendre le monde que nous ? La nuance est faible, mais changera pourtant du tout au tout les limites que nous poserons à nos créations.

2.3 Neurologie

La métacognition, par définition, nécessite de bien comprendre et représenter les mécanismes cognitifs de bas niveaux pour fonctionner. Si l'on prend l'exemple de la mémoire, les architectures IA actuelles (souvent sous forme de *buffer*) ne ressemblent en rien aux engrammes hyper-dynamiques humains. Bien que certaines techniques permettent aujourd'hui de s'approcher de ces représentations en utilisant des graphes ou en imitant certaines parties de la mémoire humaine (Cf projet *SOAR*), il semble que les solutions informatiques actuelles soient encore trop triviales pour permettre une bonne imitation du fonctionnement du cerveau humain. La métacognition, niveau encore supérieur, est-elle donc réellement accessible aux IA d'aujourd'hui ?

Références

Ackerman, C. (2025). Evidence for Limited Metacognition in LLMs. (arXiv :2509.21545). https://doi.org/10.48550/arXiv.2509.21545

Caro, M. F., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically Inspired Cognitive Architectures*, *9*, 82-104. https://doi.org/10.1016/j.bica.2014.07.002

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring A New Area of Cognitive — Developmental Inquiry. *AMERICAN PSYCHOLOGIST ASSOCIATION*. https://doi.org/10.1037/0003-066X.34.10.906

Ji-An, L., Xiong, H.-D., Wilson, R. C., Mattar, M. G., & Benna, M. K. (2025). Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. (arXiv :2505.13763). https://doi.org/10.48550/arXiv.2505.13763

Marshall, J. B. (1999). A Self-Watching Cognitive Architecture for Analogy-Making and High-Level Perception. https://doi.org/10.1080/09528130600758626

Pitrat, J. (1995). Des métaconnaissances pour des systèmes intelligents. *Quaderni*, *25*(1), 29-42. https://doi.org/10.3406/quad.1995.1110

Valiente, R., & Pilly, P. K. (2024). Metacognition for Unknown Situations and Environments (MUSE). (arXiv :2411.13537). https://doi.org/10.48550/arXiv.2411.13537

Wang, G., Wu, W., Ye, G., Cheng, Z., Chen, X., & Zheng, H. (2025). Decoupling Metacognition from Cognition : A Framework for Quantifying Metacognitive Ability in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v39i24.34723

Notes de cours de François Guillemette, professeur associé en éducation et en communication à l'UQTR. https://oraprdnt.uqtr.uquebec.ca/Gsc/Portail-ressources-enseignement-sup/documents/PDF/metacognition_notes_de_cours.pdf