Bonhoure Timothe
Lefebvre Julien
Ravella Nino
Somny Maeva

AIC
M2 IA
2023/2024

METACOGNITION
Synthèse

La métacognition en quelques mots

La métacognition peut être définie comme le fait d'avoir une activité mentale sur ses propres activités mentales. Ce concept est fortement lié à celui des méta-connaissances, qui correspondent aux connaissances que l'on peut avoir sur nos propres connaissances (leurs utilités, leurs véracités ,..).

Un débat philosophique lie métacognition et évolution humaine, abordant le fait que notre espèce serait la seule, ou non, à avoir une capacité de métacognition. Si tel était le cas, il pourrait être nécessaire d'importer cette métacognition dans les systèmes d'apprentissage, si l'on souhaite les rendre réellement "intelligents".

Les apports de la métacognition dans l'IA

Depuis l'émergence de la recherche en intelligence artificielle dans les années 1950, la possibilité d'avoir un système capable de décrire le monde dans lequel il évolue, et donc de se décrire lui-même puisqu'il est dans ce monde, s'est fait ressentir. Si à cette date la notion de "métacognition" n'était pas clairement énoncée, il s'agit cependant bien de ce concept.

L'intérêt d'une capacité de métacognition dans des systèmes d'IA, peut avoir plusieurs avantages, à commencer par l'amélioration des systèmes d'apprentissages actuels qui, bien qu'efficaces, comportent tout de même certains problèmes, souvent liés à un jeu de données mal préparé. Ainsi, à l'heure actuelle, nous sommes confrontés à des problèmes d'interprétation, voire à des problèmes éthiques, qui pourraient être résolus par la capacité d'un système à "réfléchir" sur sa propre cognition et sur ses propres connaissances.

De plus, en considérant qu'un système d'intelligence artificielle est capable de décrire le cheminement de sa "pensée" l'ayant mené à une conclusion, il serait alors possible d'expliquer les raisons justifiant cette conclusion. La métacognition serait alors un atout vis-à-vis de l'explicabilité des décisions des modèles d'IA.

La métacognition en IA dans la pratique

Depuis 1950, plusieurs essais pratiques ont été menés, même si le concept de métacognition en intelligence artificielle reste surtout un concept philosophique. Cependant, on peut noter que l'essentiel des recherches à été réalisé sur des systèmes experts (car étant le domaine de prédilection de l'époque). Par ailleurs, ces essais pratiques n'ont pas permis la réflexivité des méta-connaissance (une méta-connaissance qui s'applique à elle-même), se contentant alors de multiples couches de connaissances, chacune supervisant une couche inférieure. On peut également noter que la plupart des tests usant de couches de métaconnaissances ont obtenus de moins bons résultats que des systèmes experts "classiques".

Certaines recherches actuelles se penchent sur l'implémentation de méta-connaissances dans des systèmes à base de réseaux de neurones, qui semblent mieux performer que des systèmes "classiques". Cependant, un point essentiel à la mise en pratique de la métacognition, est la nécessité d'avoir une base de connaissances, métaconnaissances, cognition et métacognition. Or, ce sont des choses que les être humains font de manière tout à fait inconsciente. Ainsi, il est très compliqué de déterminer des règles de métacognition et métaconnaissance, pour ensuite les implémenter dans les systèmes.

Une solution serait alors d'amorcer l'IA, c'est à dire créer une première IA à laquelle on fournirait des connaissances, qui créerait ses métaconnaissances, pour les fournir à une autre IA, qui améliorerait cette base de métaconnaissances, pour la fournir à une autre IA, et ainsi de suite, pour obtenir une IA métacognitive.

La métacognition, un pas vers l'IA forte ?

Avec les prouesses théoriques d'une IA métacognitive qui peuvent être avancées, il est naturel de se demander si la métacognition, une fois maîtrisée, serait une avancée vers l'obtention d'IA dites "fortes". Mais en réalité, la question est bien plus large que cela. En effet, il serait plutôt sensé de se demander s'il est possible d'obtenir des IA fortes sans la métacognition. On pourrait même se demander si cette capacité de métacognition n'émergerait pas d'elle-même dans le cas de l'obtention d'une IA forte.

Ces questionnements peuvent être mêlés au débat philosophique sur le lien entre métacognition et évolution humaine. Si en effet, l'être humain dispose d'une intelligence différente des autres espèces en raison de sa capacité de métacognition, cela voudrait donc dire qu'une IA devrait être capable de métacognition (ou de simuler cette capacité), afin d'être catégorisée comme IA forte.

La métacognition, cachée dans toutes les IA ?

Avant de parler d'IA forte, il peut être utile de s'intéresser à l'état actuel des modèles d'apprentissage. Pourrait-on dire que, d'une certaine manière, et à leurs échelles, ces modèles font de la métacognition ? En effet, la métacognition, n'est autre que le fait d'éviter de rentrer à la main toutes les connaissances d'un monde, ainsi que les règles de celui-ci, et pourquoi et comment les appliquer. En bref, la métacognition dans l'IA peut être vue comme le fait de ne pas faire de l'IA symbolique.

Ainsi, on peut considérer que les modèles d'apprentissages, qui apprennent par principe de récompense (IA par renforcement), ajustent leurs paramètres (réseaux de neurones), modifient leurs connaissances sur l'environnement (planification), font évoluer leur décisions en fonction des résultats précédents (IA développementale), ou s'appliquent à eux même (mécanismes d'attention), font de la métacognition. En bref, il n'est pas possible de faire de l'IA sans métacognition, puisque celle-ci se cache derrière tous raisonnements, même basiques.

La métacognition, une caractéristique permettant l'explicabilité ?

Il n'est pas erroné de se dire qu'actuellement, c'est déjà la métacognition, des modèles d'apprentissage et des humains, qui permet de faire de l'explicabilité. En effet, dans le cas d'une explicabilité par pixel (dans le cas d'une classification d'image par ex.), ce sont les mécanismes de métacognition (basiques) des modèles qui permettent de faire le lien entre les différents poids des différents paramètres et la mise en avant de certains pixels. Mais c'est aussi la capacité de métacognition des êtres humains qui permet de détecter les incohérences entre la signification de la classification et ces pixels censés l'expliquer.

Une question serait alors de se demander si une métacognition telle que celle des humains, c'est-à-dire réflexive, permettrait d'améliorer cette explicabilité. On peut également se demander si le problème actuel de l'explicabilité est lié a la capacité métacognitive des modèles, ou au problème de l'ancrage des symboles. En effet, il est possible que les modèles actuels soient assez performants mais ne sont pas capables de nous fournir l'explicabilité de telle manière que l'on puisse la comprendre. Il est également possible qu'un modèle capable de métacognition telle que celle des humains ne permettent pas l'explicabilité, pour la même raison.

La métacognition, une solution aux biais de l'IA ?

Enfin, on peut se demander si le fait qu'une IA dispose de métaconnaissances et soit capable d'appliquer des réflexions sur ses propres processus cognitifs ne permettrait pas d'éviter les biais actuels de l'IA.

Si cette piste est plutôt prometteuse, elle n'est cependant pas à toute épreuve. En effet, même en imaginant qu'une IA construise ses propres bases de métaconnaissances, et métacognitions (sans qu'elles ne soient implémentées), celles-ci seraient tout de même alimentées par les êtres humains et leurs biais (à l'image d'un enfant qui n'apprend pas tout tout seul). De plus, la capacité de métacognition pourrait être la porte à des biais plus "humains" et difficiles à détecter, comme le biais de confirmation. En effet, puisque l'apprentissage de nouvelles connaissances est influencé par la métacognition, il est possible d'obtenir une IA négligeant ce qui est éloigné ou qui ne confirme pas ses métaconnaissances, plutôt que de les réexaminer.