Intelligence Artificielle et Cognition

Synthèse d'atelier : Le problème de l'ancrage des symboles

Medhi Nechadi, Nael Lahcene, Vincent Joulain

October 2025

Contents

1 Qu'est-ce que l'ancrage des symboles ?

Le problème de l'ancrage des symboles est le problème du lien des symboles (i.e. représentations abstraites telles que les mots) avec perceptions du monde. En essence, c'est l'obtention du sens d'une abstraction et de son ancrage de celle-ci dans des représentations concrètes. Le problème a été défini pour la première fois dans le papier publié en 1990 de S. Harnad, il pose deux questions essentielles à la compréhension de l'ancrage des symboles :

- Comment l'interprétation sémantique d'un système de symboles formel peut-elle devenir intrinsèque au système, plutôt que seulement parasitique sur le sens dans nos têtes.
- Comment le sens de symboles sans signification, manipulés uniquement à partir de leurs formes arbitraires, peut-il être ancré dans quoi que ce soit d'autre que d'autres symboles sans signification.

1.1 La signification

Le problème de l'ancrage des symboles peut être résumé à la définition du terme comprendre et de savoir si celui-ci peut être accordé aux intelligences artificielles ou non. Dans un premier temps, il est nécessaire de faire la différence entre la compréhension objective et subjective et ce qu'elles impliquent.

1.1.1 Signification objective / parasitaire

Comprendre est employé de manière objective lorsqu'il est possible de tester de manière empirique la capacité de compréhension de l'alter (agent intelligent, modèle de langage, humain, animal, ...). Par exemple, dire de quelqu'un qu'il comprend le français, c'est être capable de tenir une discussion sensée avec lui en français, et ainsi tester sa compréhension. On dit du sens qu'il est parasitaire ou extérieur puisqu'il est interprété uniquement par l'agent externe.

1.1.2 Signification subjective / intrinsèque

Comprendre est employé de manière subjective lorsqu'il est question de la compréhension intrinsèque et personnelle du sens des concepts et non simplement extrinsèque. C'est-à-dire que le sens n'est pas uniquement donné et compris par l'interlocuteur mais par l'agent lui-même. En ce sens, dire de quelqu'un qu'il comprend le français, c'est non seulement être capable de tenir une discussion sensée avec lui en français, mais que ce dernier fasse l'expérience subjective des mots et concepts qu'il utilise de telle sorte qu'il comprenne ce pourquoi il dit les choses.

2 Approches théoriques et expérimentales

2.1 Steven Harnad : solution à deux modèles

Pour S. Harnad il est nécessaire d'inclure un niveau sensorimoteur, les symboles doivent être ancrés dans le monde réel par nos sens. Il propose un modèle hiérarchique à trois niveaux :

- **Iconique** : Représentations brutes des perceptions directes (ex: une image de cheval).

- **Catégoriel** : Extraction d'éléments constants pour la catégorisation (ex: silhouette, taille du cou d'un cheval).
- **Symbolique** : Etiquettes (mots) liées aux catégories établies (ex: le mot "cheval").

La solution pour lui est la réconciliation entre les approches symbolistes et connexionnistes de L'IA, en revanche son modèle est uniquement conceptuel et n'a pas fourni d'implémentation concrète.

2.2 Luc Steels : les robots communicants

D'après L. Steels, le langage doit émerger par la perception et la communication et non être pré-inscrit dans le système, raison pour laquelle il mène en 2006 l'expérience suivante : deux robots sont placés dans une pièce close contenant des objets de couleurs, tailles et formes différentes ; les robots sont autonomes (possèdent des capteurs ainsi qu'un corps pouvant pointer du doigt) et s'engagent dans des jeux de langages ou l'objectif est de discriminer un des objets présent dans la pièce par sa couleur. L'aspect crucial de l'expérience étant que les robots n'avaient aucun langage ou catégories de couleurs prédéfinies, ils ont dû développer de manière autonome un système conceptuel commun à force d'interactions.

L. Steels affirme avec audace que suffisasement de progrès ont été fait en sciences cognitives et IA afin de pouvoir dire que le problème de l'ancrage des symboles a été résolu.

3 Les grand modèles de langage et la compréhension intrinsèque

3.1 LLM : un modèle uniquement prédictif ?

Prédire n'est pas comprendre, voici une critique régulièrement faite aux grands modèles de langages puisque ceux-ci sont souvent péjorativement réduits à de simples modèles statistiques. En revanche, bien qu'il soit vrai que les modèles de langages monomodaux n'aient été entraînés que sur du texte, et donc, que sur un ensemble de symboles abstraits, par conséquent, tirant leur sens que par les liens que les mots entretiennent entre eux (flamme -¿ rouge -¿ couleur), cette critique n'est pas si facile à tenir. En effet, il semble difficile de soutenir qu'un modèle très efficient dans une tâche restreinte tel que prédire le meilleur coup aux échecs ne possède aucune compréhension du jeu d'échecs, tout comme supposer qu'un LLM n'ait aucune compréhension de ce qu'il raconte alors qu'il manipule les concepts correctement.

Supposer des qualités émergentes pouvant apparaître à partir d'un système très limité dans chacun de ses composant ne semble pas être une hypothèse farfelue ou vide de sens en comparaison de notre compréhension actuelle du fonctionnement du cerveau humain.

3.2 La solution moderne : l'ancrage par la fonction

Si L. Steels proposait l'incarnation (donner un corps) comme solution, **Mollo et Millière** avancent que cette solution n'est ni nécessaire ni suffisante pour le fondement référentiel. Selon eux, les LLMs parviennent à ancrer leurs vecteurs par une nouvelle forme de causalité, liée à leur entraînement :

- Le Rôle des Contraintes Externes : L'entraînement des LLMs, notamment l'ajustement par le feedback humain (RLHF), introduit une exigence extra-linguistique. On demande au modèle

d'être non seulement cohérent, mais aussi factuel et véridique. Cette exigence externe force les vecteurs internes du modèle à se structurer de manière à tracer fidèlement les caractéristiques réelles du monde.

- L'Établissement de Fonctions Causales : Lorsque ces vecteurs, structurés par la réalité factuelle, influencent causalement la génération du texte (par exemple, un vecteur "vitesse" cause la mention d'une vitesse précise), une relation d'ancrage est établie. Le modèle n'a pas besoin de comprendre le monde comme un humain, mais ses représentations internes sont sélectionnées et optimisées pour répondre à des conditions de vérité imposées par le monde.

En conclusion, le problème n'est plus de savoir si l'IA doit avoir des yeux ou des mains, mais de déterminer si le processus d'apprentissage établit des fonctions causales fiables entre les représentations numériques et la réalité. C'est en comprenant ces mécanismes d'optimisation que nous pourrons affirmer que l'IA moderne a, ou non, résolu le problème de l'ancrage du sens.

4 Interaction avec le public

La question principale autour de laquelle le sujet du débat à tourné a été de savoir si la compréhension subjective était nécessaire afin de résoudre le problème de l'ancrage des symboles. Si la subjectivité (compréhension intrinsèque des symboles) de l'IA ne peut être vérifiée empiriquement, la question a-t-elle réellement encore un sens ?

Le sujet des sciences sociales telles que l'anthropologie ou l'histoire a été soulevé, thèmes importants qui pourtant ne reposent pas sur l'expérimentation de la démarche scientifique, objets d'études qui sont pourtant primordiaux chez l'espèce humaine.

Enfin, le parallèle avec le chatbot Elisa a été fait, quelle différence fondamentale existe-t-il entre ce chatbot et les LLMs actuels lorsqu'on souhaite affirmer que le problème de l'ancrage des symboles est résolu ? De deux choses l'une, soit il est nécessaire d'admettre qu'Elisa avait une compréhension de ce qu'elle disait, soit que les LLMs d'aujourd'hui ne comprennent toujours rien à ce qu'ils racontent.

Pour conclure, deux visions s'affrontent, la première qui place la compréhension subjective comme essentielle dans l'objectif d'atteindre une IA forte et pour qui la simple imitation de la compréhension n'est pas suffisante, excluant d'office les modèles statistiques tels que les LLMs d'aujourd'hui ; la seconde, pour qui la compréhension n'a besoin d'être qu'extérieure éliminant par la même occasion la distinction entre la démonstration de quelque chose et sa réalité (i.e. : exprimer et ressentir des sentiments, imiter la compréhension et réellement comprendre, etc). Selon son point de vue, le problème de l'ancrage des symboles peut-être considéré comme partiellement résolu ou pas du tout.