ASRI Rida, BLUMET Thomas, EDHIB Awatef
Atelier IAC 2024
M2IA - Polytech Lyon

Synthèse : The Symbol Grounding Problem (SGP)

Notre capacité en tant qu'humain à pouvoir manipuler des **symboles**, c'est-à-dire quelque chose qui représente quelque chose, provient de notre capacité à construire et à mettre à jour notre propre **réseau sémiotique**. Ce réseau est notre moyen de pouvoir associer une "signification profonde" (**concept**) à un **symbole**, lui-même rattaché à un **objet** réel. La notion d'**ancrage** désigne la capacité à relier ce que l'on perçoit du monde avec un symbole. Ce réseau, qui se métamorphose en **paysage sémiotique** (Luc Steels) lorsqu'une signification commune d'un même symbole émerge entre des individus, est important afin que l'on puisse communiquer, comprendre ce que l'on entend et surtout se faire comprendre des autres. Tout ceci s'apparente à nos caractéristiques cérébrales qui semblent bien complexes à reproduire, mais un espoir est nourri de pouvoir résoudre un problème conséquent auquel font face les agents artificiels utilisant les systèmes de symboles.
C'est le fameux **problème de l'ancrage des symboles** (**Symbol Grounding Problem**).

Dans son article ***The Symbol Grounding Problem*** paru en **1990**, **Stevan Harnad** soulève la question de comment élaborer un moyen pour que les agents manipulant des systèmes de symboles puissent interpréter sémantiquement les symboles, autrement dit de pouvoir les ancrer avec la perception de leur environnement. A ce titre il propose **deux types de représentations**, **iconique et catégorielle**, dans l'objectif d'avoir les mêmes capacités humaines de **discrimination et d'identification** nécessaire à l'ancrage des symboles. D'ailleurs **divers types d'ancrages** existent (**sensorimoteur, communicatif, épistémique, relationnel et référentiel**), tous n'ayant pas la même façon de réaliser cette "opération".

Énoncé par **John Searle** en **1980** dans son article ***Minds, Brains and Programs*** , **l'argument de la Chinese Room** se présente sous la forme d'une **expérience de pensée** questionnant la capacité d'une réelle compréhension de la part d'un agent artificiel. L'expérience met en scène une personne ne comprenant pas le chinois enfermée dans une chambre. Recevant des messages écrits en chinois et disposant d'un manuel de règles de manipulation des symboles chinois uniquement en fonction de leur forme et non de leur sens, la personne peut produire des réponses cohérente, donnant l'impression à un observateur extérieur qu'elle comprend cette langue. Bien que les réponses soient correctes, la personne ne comprend pas réellement le chinois. De la même manière, un ordinateur qui manipule des symboles selon des règles syntaxiques ne possèderait pas une véritable compréhension sémantique selon Searle. Une IA basée uniquement sur la manipulation de symboles ne serait douée que d'une compréhension apparente, sans posséder de véritables états mentaux ou de conscience. Cela rejoint **la notion d'intentionnalité** de **Franz Brentano** qui introduit l'idée que tous les phénomènes mentaux sont caractérisés par l'intentionnalité, c'est-à-dire qu'ils sont toujours dirigés, "à propos de" quelque chose. L'IA ne possédant pas ces états mentaux, elle ne ferait donc pas preuve d'intentionnalité.
Harnad, de son côté, revisite cet argument en élaborant deux versions du problème (difficile et impossible) illustrant l'incapacité d'apprendre le chinois en se basant seulement sur l'utilisation d'un dictionnaire chinois/chinois. Avec la version impossible, dans laquelle le chinois serait appris comme première langue, l'humain fait face non seulement à un cycle infini de symboles sans jamais comprendre véritablement leur signification, mais surtout à l'incapacité de se rattacher à une autre langue ou connaissances/expériences dans le monde réel ! C'est justement **la difficulté à laquelle font face les modèles symboliques qui essayent de modéliser l'esprit sans avoir de base de connaissance pour commencer.**

Néanmoins même si la Chinese Room aborde comme le SGP la question des limites de la compréhension des machines, le but dans le SGP est de rechercher comment les symboles peuvent acquérir un sens alors que dans la Chinese Room il est simplement question de savoir si une machine peut réellement comprendre ou simuler simplement la compréhension.

Plusieurs approches ont été envisagées pour résoudre ce problème, en particulier **l'approche du modèle hybride** proposé par Harnad. Le principe est alors d'utiliser les points forts des approches symbolistes (source de sémanticité mais incapable de relier un objet à son symbole) et connexionnistes (capable d'extraire d'une perception visuelle des caractéristiques inhérentes à un objet grâce aux réseaux de neurones) afin de pouvoir ancrer un objet issu du monde réel avec le symbole qui lui est associé dans le réseau sémiotique de l'agent artificiel. D'autres solutions évoqués par Taddeo & Floridi dans ***Solving the Symbol Grounding Problem: a Critical Review of Fifteen Years of Research*** (2005) sont considérés comme étant incapable de résoudre le problème, ceci à cause du non-respect d'une condition nommé la **Z-condition** (ou condition de zéro engagement sémantique) émise par les auteurs. Elle stipule qu'**aucune ressource sémantique préexistante ou externe à l'agent artificiel ne doit être présente dans celui-ci** : il faut absolument qu'il développe de lui-même son propre réseau sémiotique.
D'autres recherches sont menées sur l'utilisation de **l'ancrage neuro-symbolique** (se basant sur la même idée que le modèle hybride d'Harnad), sur l'approche dite de **l'association symbole-symbole** soutenant que les symboles peuvent acquérir leur signification en se combinant entre eux plutôt que grâce aux représentations sensorielles, ou encore sur **la cognition incarnée** où l'ancrage des symboles se réalise via l'utilisation d'agents autonomes physiquement incarnés qui possèdent des capteurs, des actionneurs et une puissance de calcul suffisante pour interagir avec le monde réel sans intervention humaine.
Une perspective évoquée serait **l'utilisation potentielle de jumeaux numériques** (**Digital Twins**) afin d'améliorer l'ancrage des symboles, ceci grâce indirectement à la représentation numérique en temps réel d'objets physiques.

En conclusion le problème reste toujours ouvert, avec des recherches actuellement menées. Pour certains comme Steels le problème semble pourtant avoir été résolu (***The Symbol Grounding Problem has been Solved so what's next ?,*** 2008) dans le sens où ses techniques de jeux du langage (Guessing Game) ont apporté certains succès lorsqu'il travaillait au Sony Computer Science Laboratories à Paris sur l'émergence du langage avec les robots QRIO .

Il a été vivement question en débat de savoir à quel niveau nous, humains, souhaitons arrêter le niveau de compréhension d'une IA. Cela a ainsi amené la question de comment définir d'un commun accord la notion de compréhension, notamment en constatant, quand on demande à une IA générative (e.g. ChatGPT, Copilot) de générer du code informatique, que cette dernière n'était pas capable de se corriger lorsqu'on lui faisait remarquer une erreur dans le code produit. Le même code est en effet "recraché" pour résoudre le problème : on entre donc dans une boucle infinie d'incompréhension ! Cela met en évidence que l'IA ne comprend actuellement pas intrinsèquement ce qu'on lui demande. Mais selon l'avis qu'il est envisagé de défendre, on peut considérer que du moment que l'IA "comprend" ce qu'on lui demande (ce qui n'est qu'un faux-semblant) alors cela nous convient. Mais dans ce cas, pourquoi s'efforcer à vouloir utiliser un système de symboles dans l'optique que l'IA comprenne comme nous ? Souhaitons-nous vraiment avoir une IA qui puisse posséder un niveau de conscience proche de l'humain ? Dans cette perspective, cela pourrait arriver dans les 100 prochaines années, pourquoi pas grâce aux avancées dans le domaine de l'informatique quantique offrant une importante capacité de calcul.