

The Symbol Grounding Problem

ECHHIBOU Amine
BOUMAZA Choayb
DAMERGI Mohamed Ali

Encadrante
Marie Lefevre

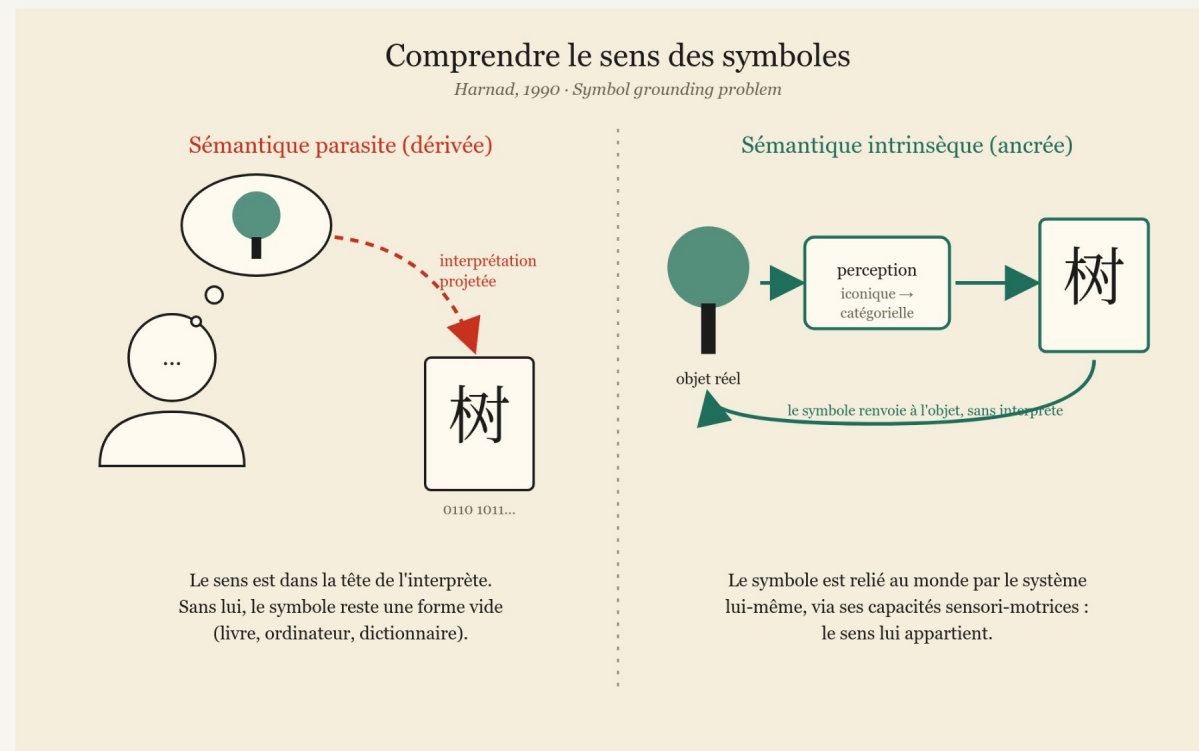
Introduction

**Est-ce que la machine
comprend vraiment ce
qu'elle produit ?**



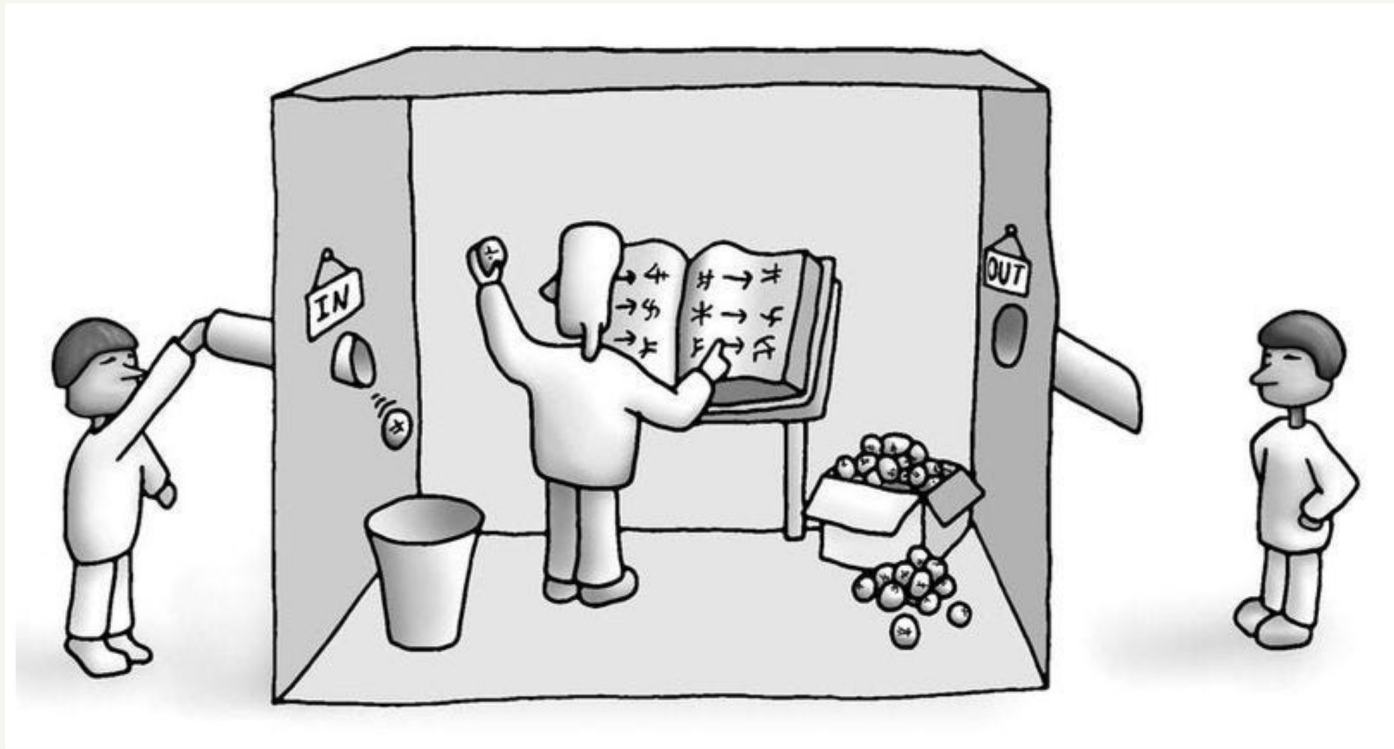
Deux sémantiques

parasite et intrinsèque



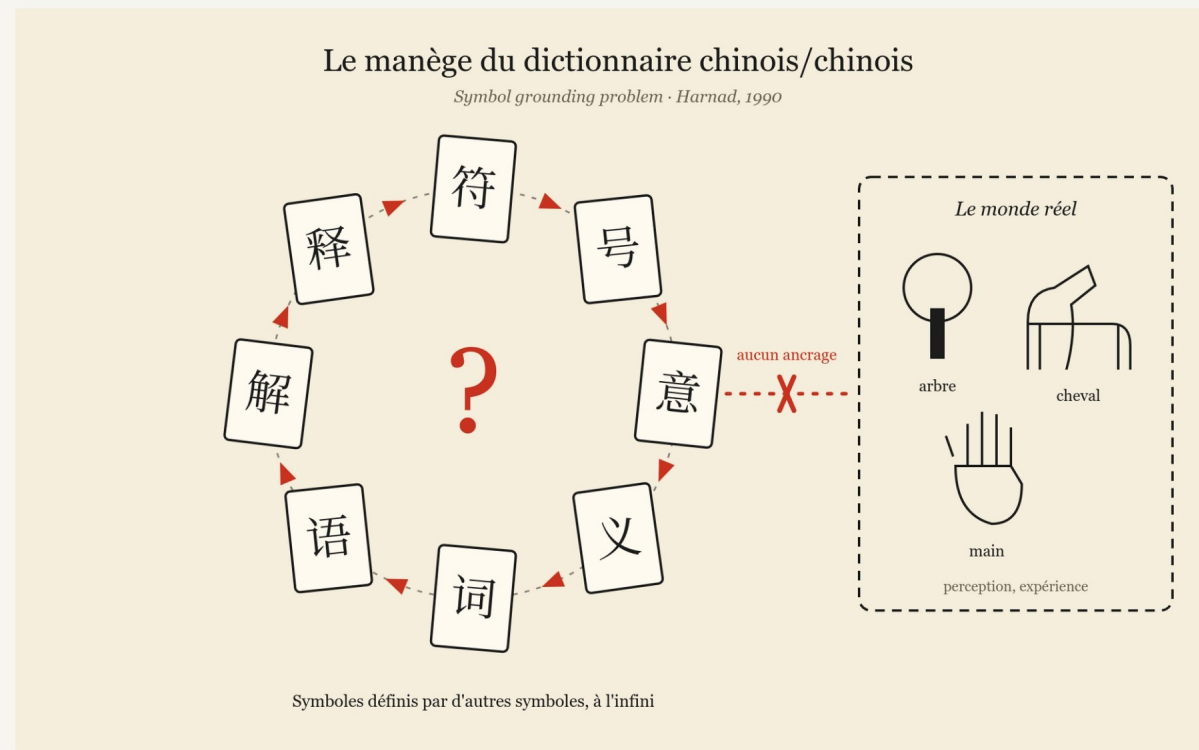
La chambre chinoise

John Searle, 1980



Le carrousel chinois

le dictionnaire chinois de Stevan Harnad, 1990



Harnad, 1990

La solution hybride « bottom-up »

Les limites

du symbolisme classique et du connexionnisme

Symbolisme classique

- Manipulation syntaxique de symboles sans ancrage dans le monde
- La signification dépend d'une interprétation externe (concepteur ou utilisateur)

Connexionnisme pur

- Excellent pour apprendre des invariants sensorimoteurs
- Du mal à expliquer la compositionnalité et le langage structuré
- Représentations subsymboliques peu exploitables pour un langage riche

Les 3 niveaux de Harnad

Du signal sensoriel jusqu'au symbole

1. Iconique · discrimination

Des copies analogiques des projections sensorielles sur les surfaces réceptrices (rétine, peau...).

- Discrimination fine entre stimuli : même ou différent ?
- Riches en détails mais pas encore catégorielles : elles ne disent pas « c'est un chien », juste « voici ce pattern sensoriel ».

2. Catégoriel · invariants

Des détecteurs de traits invariants, appris ou innés, qui extraient ce qui est commun à une catégorie d'objets.

- Identifier des catégories : « tous ces patterns différents sont de la même classe ».
- Exemples : chien, chaise, visage.

3. Le niveau symbolique

La compositionnalité

Des symboles (mots, noms de catégories) qui désignent des catégories **déjà ancrées** dans le sensorimoteur.

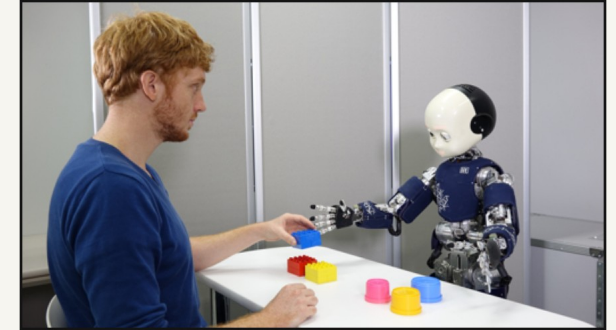
- Nommer les catégories : « chien », « rouge », « courir »
- Les combiner en phrases : « le petit chien noir court »

Les 3 modes de représentation de Jerome Bruner

- 01** Enactif
- 02** Iconique
- 03** Symbolique

Le Total Turing Test

un critère sensorimoteur



Turing Test classique (T2)

Un juge converse par texte avec un humain et une machine. S'il ne peut pas les distinguer, la machine passe le test.

Problème : un système purement symbolique peut réussir T2 en restant coupé du monde, sans perception ni action.

Total Turing Test (T3)

Le candidat doit être indiscernable d'un humain dans toutes ses capacités :

- linguistiques : conversation, raisonnement
- sensorimotrices : percevoir, manipuler, se déplacer

Il doit discriminer, catégoriser, manipuler, nommer et discuter des objets réels.

Luc Steels

The symbol grounding problem has been solved, so what's next?

2008

Le quiproquo

c-symbols vs m-symbols

c-symbol

Pour l'informaticien, un symbole c'est une adresse mémoire qui pointe vers une donnée.

```
ptr = 0x7ffe3a10
```

m-symbol

Pour le philosophe, un symbole c'est un signe qui a un sens pour quelqu'un.

Le mot "Pomme", un son, un gout, une couleur.

Même mot, deux choses très différentes.

Le dessin de Monica

Donnée brute ou sens ?



c-representation

La photo enregistre tout, au même niveau.

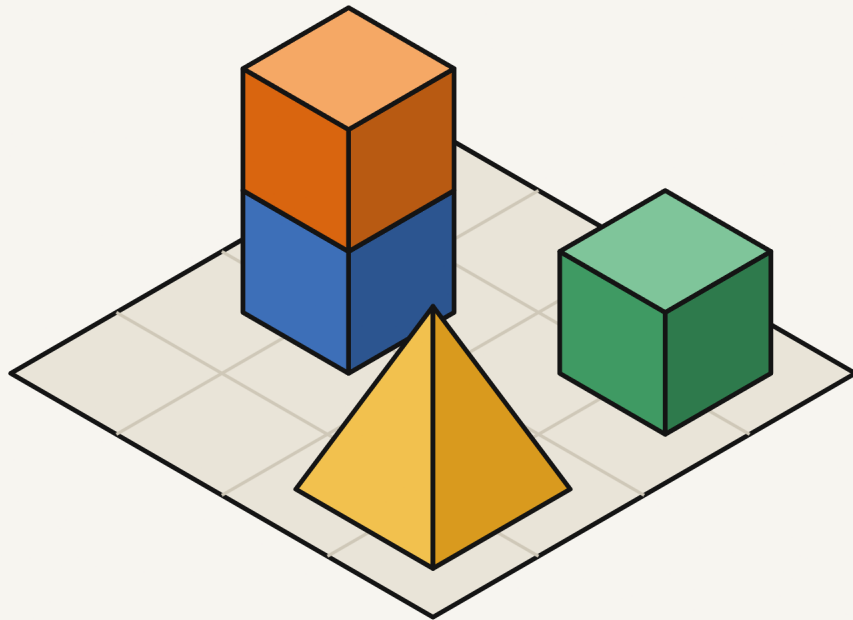


m-representation

Le dessin grossit ce qui compte pour celui qui dessine.

Shakey et le monde des blocs

Début des années 1970, une sémantique écrite à la main



Commande donnée à Shakey :

« *go to the next room and approach the big pyramid* »

```
TYPE(P1, PYRAMID) // la grande pyramide
INROOM(ROBOT, R1) // Shakey : pièce 1
INROOM(P1, R2) // pyramide : pièce 2
CONNECTS(D1, R1, R2) // porte entre les deux
```

```
BUT : NEXTTO(ROBOT, P1)
```

```
PLAN : GOTHRU(D1, R1, R2), GOTO(P1)
```

Ces prédicats ont été écrits par des humains. Le sens vient du programmeur, pas du robot.

QRIO invente ses mots

Steels, 2008 : des robots qui créent leurs propres symboles



1. Ils voient des couleurs avec leurs caméras
2. Ils les classent en catégories (réseaux RBF)
3. Ils inventent un mot pour chaque catégorie
4. Ils gardent les mots qui marchent entre eux

wabado

un mot que personne ne leur a appris



2012



2007

Et les LLMs ?

Des milliards de c-symbols... mais est-ce que ça a du sens ?

Ce qu'ils ont

- Des milliards de textes écrits par des humains
- Des statistiques très fines sur les mots
- Des réponses souvent justes et cohérentes

Ce qui leur manque

- Un corps et des capteurs pour percevoir
- Une expérience directe du monde réel
- Des mots ancrés, comme wabado chez QRIO

Sont-ils coincés dans la Chambre chinoise ?

Manipuler des symboles suffit-il à comprendre la réalité ?

Au final, les LLM ne sont-ils pas juste un Shakey amélioré ?

Et si Claude avait des yeux ?

Moi

Est-ce que Claude avec des yeux, c'est QRIO ? QRIO a inventé des mots et des actions, il a appris tout seul à utiliser son corps et l'a transmis à d'autres robots. Toi, même avec des yeux et un corps, tu exploites la donnée humaine : tu aurais peur d'un couteau parce qu'on t'a appris que ça fait mal, alors qu'il ne peut pas te faire de mal ?

Claude

Si on me donne des yeux et un corps, mon sens restera hérité. Je saurais qu'un couteau est « dangereux » parce que des millions de textes humains le disent, pas parce que j'ai quelque chose à protéger. Ma peur serait une imitation de la vôtre, un sens de seconde main. QRIO fonctionne à l'inverse : ses catégories et ses mots (wabado) viennent de sa propre perception. Ils tiennent parce qu'ils marchent entre robots, pas parce qu'un humain les lui a donnés. Donc même avec des caméras, je ne deviens pas QRIO : j'ai des capteurs, mais toujours des symboles humains.

Ma réponse courte : je suis plus proche d'un Shakey qui aurait lu toute la bibliothèque que d'un QRIO. La question qui reste ouverte, c'est de savoir si un sens hérité peut devenir un vrai sens une fois confronté au monde.