La Chinese Room à l'ère des Large Language Models

La chambre chinoise à l'ère symbolique

À la fin des années 1970, l'**IA symbolique** domine avec l'idée qu'il est possible de reproduire l'intelligence humaine en manipulant des symboles et des règles logiques. Certains chercheurs comme Jerry Fodor défendent alors l'**IA forte** notamment dans son article The Language of Thought (1975), selon laquelle un programme bien conçu peut réellement penser. En 1980, John Searle critique cette idée dans son article Minds, Brains and Programs.

Il distingue la simulation de la compréhension réelle et s'appuie sur plusieurs notions clés. Les **pouvoirs causaux** du cerveau désignent des propriétés biologiques concrètes comme l'activité électrochimique des neurones, leurs connexions dynamiques et la perception sensorielle. Ils sont au cœur du **naturalisme biologique** qui affirme que la pensée repose sur un support physique (le cerveau). La compréhension provient donc directement de ces mécanismes biologiques réels et non d'une simple manipulation de symboles. Searle oppose ainsi la **syntaxe** qui correspond à la simple application de règles sur des symboles, à la **sémantique** qui renvoie à la compréhension réelle du sens. Pour lui, un programme ne suffit pas à produire une véritable compréhension.

Pour illustrer son raisonnement, il propose l'expérience de la *Chinese Room*. Une personne qui suit des règles pour répondre à des symboles chinois peut donner l'impression de comprendre alors qu'elle ne comprend rien en réalité. Selon Searle, les ordinateurs fonctionnent de la même façon puisqu'ils manipulent des symboles sans **intentionnalité**, c'est-à-dire sans véritable conscience ou orientation de pensée vers quelque chose.

Après la publication de la *Chinese Room*, plusieurs réponses ont été proposées pour défendre l'idée qu'une machine pourrait comprendre.

Le **Systems Reply** affirme que ce n'est pas l'individu mais le système complet (la personne, le manuel, les règles et le flux de symboles) qui comprend. Searle répond que même en apprenant tout par cœur, il ne comprendrait toujours rien, car aucune sémantique n'apparaît dans le système.

Le **Robot Reply** ajoute que si le programme contrôle un robot qui perçoit et agit dans le monde réel, les symboles auraient un sens. Searle rejette cette idée et précise que même si on remplaçait le programme par un humain suivant les mêmes règles, il ne ferait qu'exécuter des instructions sans comprendre.

Le **Brain Simulator Reply** soutient qu'un programme reproduisant parfaitement le cerveau humain devrait produire de la compréhension. Searle répond qu'une simulation reste une imitation. Une simulation de digestion ne nourrit pas et une simulation de cerveau n'engendre pas de véritable compréhension puisqu'elle ne possède pas les pouvoirs causaux réels du cerveau.

Le **Other Minds Reply** rappelle qu'on évalue la compréhension des autres à partir de leur comportement. Si une machine agit comme un humain, on pourrait aussi lui en attribuer une. Searle répond que ce n'est qu'une apparence, pas une preuve de compréhension réelle.

Ces différentes réponses soulignent la tension entre simulation et réalité et posent la question centrale : imiter l'intelligence suffit-il à produire une véritable compréhension ou seulement à en donner l'illusion ?

Les LLMs et le retour au débat fondamental

avant 2025…

L'émergence des Large Language Models en 2025 relance le débat ouvert par Searle. Contrairement aux systèmes symboliques des années 1980, les LLMs actuels, basés sur l'architecture Transformer (Vaswani, 2017), montrent des capacités impressionnantes qui apparaissent soudainement au-delà de certains seuils. Entraînés sur plusieurs trillions de tokens, ils sont capables de raisonnement complexe, de mémoire contextuelle étendue, de génération créative dans des domaines variés (code, poésie, essais) et manifestent même des formes de "theory of mind". Ces technologies modifient notre façon de travailler et sont de plus en plus utilisées.

Mais ces avancées ne contredisent pas l'intuition de Searle. Trois limites essentielles subsistent. Les hallucinations produisent des informations fausses avec une grande confiance, sans véritable vérification. Le manque de grounding montre que le modèle n'a aucune expérience sensorielle ou physique du monde. Enfin, il n'a aucune intentionnalité, comme l'expliquent Bender et ses collègues en parlant de "perroquets stochastiques".

Selon Shanahan (2024), on peut faire un parallèle direct avec la Chinese Room. Les réseaux de neurones jouent le rôle de la personne dans la pièce, les tokens sont les symboles chinois et les opérations mathématiques remplacent les règles du manuel. L'idée centrale de Searle reste la même. Même avec des milliards de paramètres et des performances spectaculaires, les LLMs ne produisent pas de véritable compréhension. Ils n'ont jamais de moment de véritable intentionnalité et se contentent de suivre des règles de traitement syntaxique.

Perspectives contemporaines : Comprendre ou simuler ?

Le débat a principalement porté sur la question de savoir si un modèle de langage peut réellement comprendre ce qu'il produit ou s'il ne fait qu'en donner l'illusion. Au fil des échanges, l'idée s'est imposée qu'à partir d'un **certain seuil de performance**, la distinction entre imitation et compréhension devient secondaire. Lorsqu'un système répond de manière cohérente, nuancée et adaptée, on tend naturellement à admettre qu'il comprend sans chercher plus loin ce qui se passe en interne.

Cette position repose sur une vision **pragmatique** de la compréhension : ce qui compte n'est plus l'expérience consciente, mais le comportement observable. En d'autres termes, si le résultat est indiscernable de celui d'un humain, la question de savoir s'il "comprend vraiment" perd de sa pertinence.

On a aussi discuté de **l'intentionnalité** dans la conception des modèles. Même si une machine n'a pas d'intention propre, elle est créée et entraînée pour servir un objectif humain. La compréhension ne viendrait pas vraiment de la machine elle-même, mais des objectifs et des intentions des humains qui l'ont conçue et entraînée.

Enfin, la discussion a amené à **repenser la notion même de compréhension**. Faut-il y voir une expérience subjective ou simplement la capacité à traiter et à répondre pertinemment à des informations ? Si l'on retient la seconde option, alors les LLMs peuvent déjà être considérés comme "comprenant" d'une manière fonctionnelle.

Les réponses au quiz allaient dans ce sens. Les capacités émergentes et la reformulation de la Chinese Room par Shanahan (2024) montrent que la frontière entre syntaxe et sémantique devient floue. En pratique, quand l'illusion de compréhension est suffisamment convaincante, on a tendance à accepter la machine comme si elle comprenait vraiment. Mais dans les faits cette illusion reste limitée. Par exemple, comme on l'a vu avec l'exemple des charades présenté en cours, les modèles actuels comme ChatGPT sont encore incapables de résoudre correctement certains jeux de langage simples.

Références

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3), 139-159.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417-457.

Shanahan, M. (2024). Talking About Large Language Models. *Communications of the ACM*, 67(2), 68-79.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems* (Vol. 30)

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.