Artificial intelligence is stupid and causal reasoning will not fix it

Résumé de Kilian Debraux

## Thème

Les similitudes et différences entre un humain et une intelligence artificielle

## Problématique

Est ce qu’une intelligence artificielle pourra un jour fonctionner comme un humain (effectuer les mêmes tâches, comprendre des concepts...) ?

## Thèse

L’intelligence artificielle ne pourra jamais réfléchir et comprendre comme un être humain.

## Concepts

* Fonctionnalisme : le fonctionnalisme est le fait de penser que l’esprit humain n’est que le résultat de calculs dans notre cerveau et que donc, si nous arrivions à reproduire ces calculs sur un autre support (par exemple en informatique), nous pourrions créer un esprit semblable à l’humain *« if computers can adequately model the brain, then, theory goes, it ought to be possible to program them to act like minds »*. L’auteur est par ailleurs contre cette thèse.
* IA forte/faible : Une IA forte est un modèle d’IA étant capable d’explorer pour comprendre un environnement inconnu et donc de s’adapter comme un humain le ferait. Ce modèle est pour l’instant seulement hypothétique car il n’existe aucune IA capable aujourd’hui de faire cela. Nous ne connaissons donc que des IA faibles : des programmes certes très efficaces mais hyper spécialisés dans une seule tâche *« Judea Pearl believes that we will not succeed in realizing strong AI until we can create an intelligence like that deployed by a 3-year-old child »*.
* Test de turing : Le test de turing, mis au point par Alan Turing, est un test pour savoir si une IA est devenu forte. Pour cela, il suffit de faire discuter un humain avec des programmes informatiques et avec d’autres humains, sans savoir qui est humain et qui ne l’est pas. Si à la fin du test, le cobaye ne peut dire qui est une machine et qui est un humain, le programme a passé le test de turing. Ici il est fait mention du mini test de turing (MTT), un test de turing où le nombre de questions possibles est limité, et du test de turing total, où l’on peut poser les questions que l’on veut. Bien qu’il soit un test pilier dans l’informatique, il est aujourd’hui remis en cause. En effet, certaines IA ont réussi plusieurs fois à duper des êtres humains sans pour autant être une IA forte. On peut penser à ChatGPT, mais en réalité, depuis les années 60 les IA sont capable d’avoir une conversation plausible avec un humain sans forcément comprendre de concepts (<https://fr.wikipedia.org/wiki/ELIZA>)
* Chambre chinoise : La chambre chinoise est une expérience de pensée visant à mettre en lumière les astuces utilisées par l’IA pour faire croire qu’elle fait preuve de compréhension et de réflexion. L’expérience est la suivante : on enferme un humain dans une pièce seul et vide. Avec lui, il a un manuel avec des questions en chinois et les réponses qu’il faut donner. Puis on lui passe une question en chinois et il doit donner une réponse. Pour la donner il peut regarder dans son manuel quelle réponse convient à la question, et ainsi il pourra retourner la réponse juste. Si l’on répète plusieurs fois l’expérience, et que l’humain retourne toujours la bonne réponse, on pourrait croire qu’il arrive à comprendre le chinois et les concepts exprimés dans cette langue. Or ce n’est pas le cas, il ne comprends rien de ce qu’il lit, il a simplement des règles prédéfinies. Or un IA qui arrive à répondre à des questions de façon pertinente pourrait, de même, ne pas comprendre la question, mais simplement avoir un bon dictionnaire de règles lui permettant de répondre juste à toutes les questions.
* Théorème de Gödel : Le 1er théorème de Gödel est un théorème mathématique qui affirme l’énoncé suivant : « Un système mathématique ne peut être à la fois complet et cohérent». Un système mathématique est comme une sorte d’univers dans lequel on construit des théorèmes à partir d’axiomes. Un système est donc cohérent, c’est à dire qu’il ne contient aucune contradiction, soit complet, c’est à dire que tous les théorèmes de ce système sont démontrables. Ainsi, dans un système cohérents, il existe des théorèmes qui sont vrais, mais que l’on ne peut pas démontrer.

## Arguments

* Une intelligence artificielle ne comprend pas de concept au contraire des humains *« Siri’s response clearly demonstrates that it does not understand language, and specifically the ontology of books and liquids, in anything like the same way that my 5-year-old daughter does. .* Elle ne fait que donner une réponse à une situation donnée, et en évaluant sa réponse, un humain va la corriger où non, ce qui donnera lieu à une amélioration de son modèle. Elle ne peut donc pas évoluer dans un environnement inconnu.
* « Shadow of mind » est un argument avancé par le mathématicien et physicien Roger Penrose. Il avance que la compréhension humaine des mathématique ne peut être écrit sous la forme d’un ensemble de règles (I.e. un algorithme). Pour cela, il imagine un ensemble de solutions pour résoudre un problème donné, et essaie de formaliser cet ensemble (le traduire en algorithme). Cependant, ce nouvel ensemble d’algorithmes contient une contradiction, il ne peut donc pas être construit. Ainsi, l’ensemble des solutions humaines pour résoudre un problème ne peut pas être traduit sous forme d’algorithme.
* Le dernier argument avancé est le Dancing with Pixies, Reductio ad Absurbum. C’est un argument qui stipule qu’une IA est comme un compteur (une machine capable de compter à intervalle régulier) en plus évolué. Et que donc, si on considère qu’une IA va-t-être intelligente et surtout consciente, alors un compteur peut l’être aussi. Cet argument s’oppose plus à l’idée d’une IA consciente que d’une IA forte. Enfin, l’argument va même plus loin en démontrant qu’un caillou peut aussi être considéré comme un compteur, et que donc si une IA peut être consciente, alors un caillou aussi, ce qui est une hypothèse (cette hypothèse porte le nom de « panpsychism ») que l’auteur considère comme absurde, donc trivialement non valide, ce qui invalide la prémisse : « Une IA peut être consciente ».

## Positions critiquées par l’auteur

L’auteur critique l’idée qu’une IA puisse se rapprocher d’un humain dans sa compréhension de son environnement ou dans le fait de développer une conscience. Pour lui, il y a une différence fondamentale entre une machine et un humain, et peu importe les efforts mis dans le développement de l’IA, elle ne pourra jamais être une IA forte.

C’est un avis assez différent de la population en général. Il y a beaucoup d’exemple dans la culture populaire d’IA étant consciente ou capable de réelle intelligence (Terminator, Matrix, 2001 l’Odysée de l’espace). Et c’est une crainte partagée par une partie de la population.

## Avis personnel

Mon opinion est à l’opposé de celle présente dans ce texte. Je pense en effet qu’il peut être possible qu’un jour une IA puisse comprendre réellement un concept. Plus exactement, je pense que notre façon d’apprendre n’est pas différentes de celle d’une IA (donc nous ne somme finalement qu’une chambre chinoise très bien documentée) et que si l’on nous plonge dans un environnement totalement inconnu, nous allons faire des erreurs tant que nous ne pourrons pas juger des conséquences de nos actions.

Or c’est bien comme cela qu’une IA agit : elle donne des réponse et elle se calibre en fonction de si elle reçoit une récompense ou une punition.

Pour ce qui est de la conscience, je ne suis pas très réceptif à l’argument proposé dans le texte. En effet, en quoi est ce absurde de penser qu’une pierre avec des échange électriques et chimiques très bien coordonnés puisse devenir consciente ? Mais je ne pense pas que notre conscience ne soit que le fruit d’échange électriques et chimiques à cause d’un autre argument avancé par le philosophe Franck Jackson : La chambre de Marie. Il montre que Marie, une brillante scientifique ne peut voir que en noir et blanc, ne peut comprendre ce que ça fait de voir des couleurs, même si elle comprend tous les mécanismes physiques, biologiques, etc, elle ne pourra jamais comprendre ce que ça fait de voir une couleur, ce qui montre qu’il existe une différence entre les mécanismes physiques et la sensation en elle même.

Enfin, sur l’argumentation en elle même, je trouve que l’auteur affirme de manière convaincu mais plusieurs arguments se basent sur des prémisses qui font débat et qui ne sont donc pas admis comme par exemple le test de Turing est aujourd’hui contesté, et l’argument de Penrose est fortement débattu. Selon moi, nous ne savons pas ce qui peut être possible avec l’IA car nous n’avons pas encore une compréhension suffisante de notre propre intelligence ni du concept de conscience pour pouvoir prétendre juger une IA sur ces domaines.