# L'homme versus machine

15. Juin 2015, Cedres réflexions – Espaces des Terreaux

**Défense de l'homme**, par *Wulfram Gerstner* (EPFL)

Je pense donc je suis !
Cela me plait de commencer ma défense de l'homme avec cette citation de René Descartes publiée en 1637 dans son 'Discours de la méthode'.
Je pense donc je suis ! Cogito ergo sum, cette citation me plait parce qu'elle met bien l'accent sur une spécificité de l'homme : l'homme est quelque chose assez spéciale, merveilleux même, parce que c'est une chose qui sait réfléchir, une 'res cogitans' en Latin.
Pour moi, c'est cela la grande différence avec un animal ou avec une machine, la possibilité de l'homme d'avoir connaissance de soi-même, ou plus précisément :
**La conscience de soi.**

La conscience de soi a, pour moi, deux aspects :
**Premier aspect : l'action**.
J'ai connaissance de mes actions, je peux observer mes actions, je me sens **responsable** de mes actions, parce que je les perçois comme les miennes. Si toi, **tu bouges ton bras**, c'est toi qui le fais. Si moi, je bouge mon bras, c'est bien moi qui le fais et c'est bien mon bras à moi.
Je ne sais pas si quelqu'un dans la salle a deux voitures, imaginons une **Renault Espace familiale, et une Porsche**. Si vous pressez la pédale de la Renault Espace elle accélère tout gentiment. Si vous pressez la pédale de la Porsche, cela fait vroum et vous partez – et dans les deux cas vous sentez bien que c'est vous qui êtes à l'origine de l'accélération. Alors premier aspect – vous êtes responsables de vos actions.

**Deuxième aspect de la conscience de soi : la réflexion mentale**.

Quand je réfléchis à quelque chose, **je sais que je pense**, je sais que je réfléchis. Le contenu de mes pensées m'est connu, je peux observer non pas seulement mon corps qui bouge, mais en plus je peux observer mes pensées !

Je trouve cela assez surprenant :

Mes pensées peuvent avoir comme sujet non pas seulement les choses externes, des choses que j'observe dans le monde ---

Mais **mes pensées peuvent avoir comme sujet la pensée même** !

Si vous me dites que cela est presque circulaire, je répondrai, mais oui, c'est circulaire, ou plus précisément autoréférentiel.

Alors, c'est ce deuxième aspect de la connaissance de soi, qui m'intéresse surtout : **La réflexion, la pensée réfléchie.**

Le docteur François Fleuret viens de vous parler du **Test de Turing**. François Fleuret est un collègue que je connais depuis longtemps, que j'estime énormément pour son honnêteté, sa clarté, sa précision intellectuelle. Et on s'est rencontré il y a quelques semaines pour une première discussion. Il faut dire que pour moi c'est un vrai plaisir de discuter avec lui – surtout parce qu'on est **d'accord sur beaucoup de choses, mais pas toutes**.

Cher collègue, Cher ami, François, tu as mentionné dans ton discours le Test de Turing. Et tu as expliqué de façon claire la question de départ :

Est-ce qu'un observateur qui peut poser des questions sur n'importe quel sujet avec son clavier de correspondance et qui reçoit des réponses écrites, est-ce qu'un tel observateur peux distinguer si les réponses sont écrits par un homme ou par un ordinateur ?

Effectivement, c'est une manière très élégante de juger la performance d'une machine, de **tester expérimentalement l'intelligence artificielle**.

Mais, il faut l'admettre, **ce test rate un point essentiel**.
Pour moi, ce test s'inscrit dans une tradition de psychologie expérimentale qu'on appelle le '**behavioralism' en anglais ou 'comportementalisme'** en français. C'est une approche expérimentale où on décide de se concentrer sur le **comportement observable**, par exemple d'un animal.
Par exemple, on observe qu'un **chien** apprend de s'approcher chaque fois qu'il entend le **bruit caractéristique de la boîte à nourriture**. Ensuite on peut mesurer le temps d'approche, et alors on peut quantifier le succès d'apprentissage par une réduction du temps d'approche en fonction du nombre de fois qu'il a entendu ce bruit caractéristique.

Et bien sur, il y avait **à l'époque** de très bonnes raisons de faire comme cela. Comme cela on évite d'attribuer des états mentaux à un animal quand on ne sait pas si l'animal en question a vraiment un tel état mental. Par exemple, notre chien a-t-il vraiment **compris** que le bruit veut dire nourriture, ou alors est-ce que c'est plutôt un **réflexe** qu'il a appris ?
C'est quoi la différence ?
Nous savons tous que quand on est chez le médecin, le docteur peut faire bouger notre jambe en touchant une zone de réflexe avec son petit marteau – la jambe bouge sans que nous ayons initié le mouvement. C'est cela un réflexe.
Et il y a beaucoup de **réflexes qu'on a appris**. Si vous êtes au volant d'une **voiture**, vous commencez automatiquement à freiner quand la voiture devant ralentit. Effectivement, vous n'avez normalement pas le temps de réfléchir. Ou les **stars du tennis**, Roger Federer, Stan Wawrinka : ils répondent à une balle rapide avec des réflexes qu'ils ont appris sur beaucoup d'années d'entrainement.
Alors, je ne peux pas savoir si le chien a appris un réflexe quand il s'approche ou s'il comprend le sens de ce bruit. Et si je ne sais pas l'état mental d'un animal, c'est plus sûr de se baser seulement sur les

observations du comportement. C'est cela l'approche scientifique, c'est cela la procédure expérimentale du comportementalisme.

Je disais tout à l'heure que c'était une bonne stratégie expérimentale à l'époque. Par contre, aujourd'hui avec **l'imagerie médicale**, on peut observer pas seulement le comportement mais aussi, au moins partiellement les bases **des états du cerveau**. Si on peut observer l'état du cerveau, il faut utiliser ces connaissances pour l'interprétation des expériences !

**Revenons sur les machines** :
Je suis d'accord avec Allan Turing qu'il faut éviter d'attribuer des états de pensées à des machines, quand on ne sait pas si cela existe.
Par exemple :

- Vous connaissez des **thermostats**. La température de beaucoup de chambres est, en hiver, réglée par un **thermostat**. On pourrait bien entendu dire : Actuellement, le thermostat réalise qu'il fait trop froid et pour cette raison-là **il décide** d'ouvrir la vanne du chauffage. Maintenant il pense qu'il fait trop chaud et il ferme le chauffage. Mais, en fait, on sait bien que notre pauvre thermostat ne pense pas du tout. C'est un simple détecteur lié à un bouton de réglage. **Il faut se méfier.** On a tous tendance à attribuer des états d'esprit à des objets – ce matin **ma voiture** ne VOULAIT pas démarrer, elle en avait marre et cetera.
- Vous avez déjà vu dans les jardins, sur les gazons, ces **robots tondeuses** ? La première fois qu'on voit une de ces machines, on est tenté de dire – ah, elle est très intelligente, maintenant elle voit qu'elle est au bord du jardin et elle décide de faire demi-tour.

Maintenant elle choisit un autre angle, pour arriver au coin du jardin. Mais en fait, elle ne pense rien – **elle fait ce qu'elle fait. C'est tout**.

- **Alors, méfions-nous d'attribuer une pensée à une machine quelconque**.

Fin de parenthèse – **retournons au test de Turing**.

Supposons qu'on a une machine parfaite. A chaque question que notre observateur pose la machine donne exactement la même réponse que l'homme dans l'autre chambre. On peut questionner la machine pendant 10 minutes ou 2 heures, c'est impossible de trouver une différence entre les réponses d'un humain ou celles de la machine. Alors, la machine a passé le test du Turing.

**Mais attention !**

Est-ce que cela veut dire que la machine peut penser ?

Pas du tout. On commettrait la même erreur qu'on ferait si on disait que notre thermostat peut penser. Il ne faut pas attribuer une pensée là où on ne l'a pas observée !

L'avantage de la machine (contrairement au chien) c'est qu'on peut comprendre exactement sa manière de fonctionner. On peut ouvrir la boîte et observer. Alors elle fonctionne comment?

**Un ordinateur suit une recette**. Il lit une question, mot par mot, caractère par caractère, il la transforme dans son langage de machine, il compare avec les instructions, cherche des contenus de sa mémoire, prend la prochaine instruction et de suite.

Il y a un célèbre exemple dite **'la chambre chinoise'** du Professeur John Searle de Berkeley. Pour comprendre le fonctionnement d'un ordinateur il

propose d'imaginer une grande administration dans un immeuble avec des centaines de chambres, et avec des employés qui, malheureusement ne parlent pas du tout la langue locale. Imaginons un groupe de chinois comme employés. **Les seules choses qu'ils savent lire sont les numéros**. Alors, dans chaque chambre, il y a une pile de papier. Sur chaque papier il y a un mot en français (que les employés ne comprennent pas) plus un numéro qui indique dans quelle chambre ils doivent apporter le papier. Arrivé là, ils posent le papier qu'ils ont amené à coté de la pile. Maintenant il y a **deux papiers visibles**, celui en haut de la pile et celui qu'ils ont amené. **Sur un grand tableau au mur, il y a une table d'instructions. Si Papier 1 montre le mot 19, et le papier 2 le mot 1457, ensuite remplacer le mot sur papier 2 par le mot 66, mettre le papier amené sur la pile, et chercher le papier de la chambre 18**. En fait, avec une telle procédure on peut résoudre n'importe de quel problème qu'on sait résoudre avec un ordinateur ultramoderne. Cela occupera une centaine d'employés qui viennent et vont, cela sera lent, mais si la procédure qu'on appelle **l'algorithme** est bien conçu, on arrivera à la bonne réponse.

Alors, en principe, notre équipe de chinois sera capable de reproduire - pour chaque question posée pendant le test de Turing - la même réponse que notre ordinateur, qui selon notre hypothèse de départ passera le test de Turing. Il faut juste supposer que l'algorithme employé par l'équipe chinoise sera identique à celui utilisé par l'ordinateur.

**OR, évidemment notre équipe chinoise ne comprend rien du sens** des questions posées pendant le test de Turing. Les chinois suivent seulement des instructions simples : va dans la chambre 19 et prends le papier. Dépose le papier dans la chambre 137 et cetera. A aucun moment la procédure ne 'pense', ni les employés. Ils exécutent une procédure, un

algorithme, mais ils ne réfléchissent pas à la question posée par l'observateur dans le test de Turing.

Cet exemple-là illustre bien que **la machine ne pense pas**. L'intelligence apparente du système est cachée dans **la conception de l'algorithme**, de la procédure, du programme. Mais la machine **exécute seulement ce que le programmeur a prévu**.

**Conclusion :**

**il n'y a aucune trace d'une conscience de soi dans la machine.**

**Par contre, l'homme possède une conscience de soi, et c'est pour cette raison-là qu'il surpasse la machine.**

Quels sont **les objections possibles à ce raisonnement?**

Il y a peut-être entre vous quelques-uns qui se disent.

Mais aujourd'hui il y a des **machines qui apprennent**, alors ce n'est pas vrai que le programmeur qui a conçu l'algorithme a tout prévu.

Alors, oui, ce constat est correct. Aujourd'hui il y a des **machines qui apprennent.**

Par exemple, un groupe de chercheurs à Londres a récemment conçu une machine qui sait apprendre à jouer des jeux vidéo qu'on jouait il y a une trentaine d'années sur les ordinateurs Atari. La machine regarde la scène du jeu sur l'écran, elle commence à utiliser les boutons un peu au hasard et reçoit son score de succès comme n'importe de quel joueur humain. A chaque jeu, la machine essaie une autre variante de jouer et si le score est meilleur elle accepte cette variante. Et, vu que la machine ne doit pas dormir et n'a pas non plus de parents qui disent que jouer est une perte de

temps, elle peut s'entrainer de longues heures – et à la fin elle est bien meilleure que les meilleurs joueurs professionnels.

Mais, notons que notre machine ne sait toujours pas ce qu'elle fait, elle n'a toujours pas de conscience de soi. On est toujours dans notre maison chinoise, sauf, disons, qu'il y a des **papiers de différentes couleurs**. Chaque couleur correspond à une certaine stratégie dans le jeu vidéo. Alors, quelquefois un employé reçoit sur un papier blanc l'instruction d'échanger une pile de papier de **couleur bleue** par une pile de **papiers verts**. C'est cet **échange de pile d'instructions, qui correspond à l'apprentissage**. Mais, ni les employés, ni la procédure elle-même ne comprennent ce qui se passe. La va-et-vient dans la maison chinoise qui simule le processus d'apprentissage par un algorithme n'a toujours pas de conscience de soi.

**Deuxième objection – il y a des gens qui diraient qu'évidemment les machines actuelles sont encore trop simples.** Mais ils affirmeront qu'une fois que les machines sont plus complexes elles auront la conscience de soi.

Peut-être, je ne veux pas exclure cette hypothèse avec une certitude de 100 pourcent, mais de mon avis, c'est très peu probable, et ceci pour une raison fondamentale. Avec un ordinateur on peut **simuler** des processus physiques mais la simulation d'un processus n'est pas identique au processus-même.

Par exemple, les physiciens savent simuler **la transition de l'eau vers la glace**. Dans l'état liquide les molécules sont mobiles, mais après une baisse de température – simulée dans l'ordinateur en utilisant les lois de la

physique,  ils s'accrochent l'un à l'autre et la matière se stabilise dans l'état de glace.

Les simulations de l'eau peuvent **prédire avec une exactitude étonnante** les propriétés de l'eau et de la glace pour chaque température. **Mais attention,  cela reste une simulation**. L'ordinateur ne devient pas humide quand l'eau simulée est liquide et pas non plus froid que l'eau simulée a des propriétés de la glace. Une simulation sur ordinateur sort des chiffres que les physiciens savent interpréter comme propriété de la matière, mais cela ne devient pas la matière.

Alors, je pourrais imaginer **une simulation de la conscience** sur ordinateur comme **un programme spécial. Ce programme travaillera sur un niveau plus haut et observera** tous les autres programmes de l'ordinateur et sortira des informations de type : maintenant l'ordinateur joue un jeu vidéo et s'occupe du traitement d'une image sur l'écran du jeu, maintenant il additionne plusieurs variables pour calculer la meilleure action, maintenant il choisit la prochaine action et cetera.

Ce programme observateur simule bien quelques aspects de la conscience, par exemple l'aspect observateur, l'aspect de l'attention sur une chose à la fois, l'aspect qu'on peut parler de soi-même. Mais attention : cela reste une **simulation de la conscience** et ce n'est pas la même chose qu'avoir de la conscience.

**Conclusion :**

L'homme est quand-même différent d'une machine. Il a l'esprit, il a la conscience de soi. La conscience reste, même aujourd'hui un mystère pour la science. L'homme reste un être à part, parce qu'il est conscient de lui-même.