Conscience : la clé du vivant. Limites de l'intelligence artificielle.

Michel Troublé Ph. D – directeur de recherche

Résumé

La conscience qui nous ouvre à la représentation d'un monde autrement clos sur lui-même, est un donné fondamental de la nature, un opérateur essentiel à la genèse des structures vivantes et des processus cognitifs qui y sont associés. La conscience est la clé du vivant, en son absence aucune vie ne serait apparue sur la Terre ni sur aucune exoplanète.

La 'théorie computationnelle de l'esprit' où l'esprit humain fonctionnerait comme une machine informatique est tout à fait infondée.

Un robot construit à partir des seules mises en relations de composants techniques gérés par les lois physiques ne peut pas être fondamentalement autonome, auto-organisé, à l'image des êtres humains. Ce n'est qu'un automate plus ou moins efficient dans un environnement qui a été spécifiquement défini par son constructeur lequel, lui, est naturellement doté d'une conscience qui est irréductible à toute interaction physique.

Sommaire :

Introduction

Nombreux sont les chercheurs qui soutiennent que la vie artificielle qui s'inspire des systèmes vivants, est pour demain. Que des robots humanoïdes seront bientôt dotés de capacités semblables et même supérieures à celles des êtres humains en matière d'intelligence et de forces développées. À notre image, ces robots, *autonomes*, s'adapteront aux contraintes du moment en inventant de nouvelles stratégies. Ils seront *conscients* et auront des *émotions*. Afin d'assurer leur pérennité, leur « vie », ces robots seront capables de *décider* intelligemment des actions à entreprendre en réponse aux stimulations utiles ou nuisibles induites par l'infinie diversité des objets du monde qu'ils perçoivent.

Ces structures physico-chimiques qui seraient artificiellement *vivantes* et à ce titre dotées de capacités semblables à celles des êtres vivants en matière des décisions qu'ils prendraient

pour assurer leur pérennité en toutes circonstances seraient ainsi strictement *autonomes*. Elles ne recevraient aucune assistance humaine en ce qui concerne le choix des solutions possibles en matière des *actions* qu'ils devraient effectuer pour assurer la durabilité de leur structure et des fonctionnalités qui y sont attachées.

La compréhension des mécanismes de *décision* que devraient posséder ces robots pour être *autonomes*, artificiellement *vivants*, est donc essentielle. Les techniques de la robotique et de l'intelligence artificielle (réseau de neurones, logique floue, algorithmes génétiques,..) permettent d'effectuer une approche formelle de la logique de ces dits mécanismes tels que *connexionnisme*, *apprentissages par récompenses*, *structures dissipative*, *sélection évolutive*, que nous analyserons successivement.

Une définition du vivant. Le « théorème d'indiscernabilité »

Pour que soit assurée la durabilité d'un robot artificiellement *vivant*, son analyse fonctionnelle montre que pour chaque objet du monde, aux propriétés infiniment variées, que celui-ci perçoit à l'aide de ses différents capteurs, il faut que son contrôleur (son « cerveau ») sache *choisir* adéquatement les actions à accomplir par le truchement de ses actionneurs (système de locomotion, bras manipulateur).

C'est ainsi qu'un robot d'exploration terrestre sensible à toute élévation de température qui peut le détruire doit, pour être *autonome*, artificiellement *vivant*, effectuer impérativement les actions suivantes : *fuir* de la lave en fusion, *fuir* un feu de forêt, *fuir* une nappe de pétrole enflammée,... En d'autres termes, le contrôleur de ce robot *autonome* doit être capable de créer une *catégorie cohérente* des objets perçus par ses capteurs, en l'occurrence la *catégorie* {*fuir* tous les objets chauds}, alors qu'il n'existe aucune affinité physico-chimique particulière de ces objets avec le contrôleur qui pourrait expliquer ce type d'actions pérennisantes :

> La capacité de créer des *catégories cohérentes des objets perçus* est une définition opérationnelle de la vie artificielle. Par là même, c'est aussi la définition des *êtres vivants* qui doivent également assurer leur pérennité en toutes circonstances au regard des contraintes infiniment variables du milieu qui ne peuvent que les détruire.

Du point de vue du mécanisme de *décision* dont le contrôleur du robot doit être ainsi doté, cette capacité de créer des *catégories cohérentes* implique logiquement que les différents objets perçus par le système soient *discernables* les uns des autres. Dans le cas contraire, ces *décisions* ne pourraient être en effet que prises au hasard ce qui serait assurément antinomique de la capacité attendue de former des *catégories cohérentes* qui fondent son *autonomie*.

Que les objets du monde perçus par les capteurs du robot soient physiquement *discernables* afin que celui-ci puisse, par exemple, *fuir* systématiquement les objets chauds et *aller* au contraire vers des sources d'énergie supposées froides, semble aller naturellement de soi. Mais cela pose en réalité un problème épistémologique majeur :

En s'appuyant en effet sur la théorie formelle de la « Reconnaissance des formes[1] » qui porte sur l'identification des formes d'objets à partir de leurs paramètres caractéristiques afin de prendre des décisions dépendant des catégories attribuées à ces formes, on peut démontrer la propriété essentielle suivante :

> **Les différentes formes des objets qui sont perçues par un système physique (ou physico-chimique) lors d'un processus de mesure, sont physiquement *indiscernables* par sa partie opérative ou actionneur (système de locomotion, bras manipulateur).**

[1] Satosi Watanabe - *Pattern recognition, human and mechanical*, John Wiley & Son, 1985.

Appliqué à notre robot d'exploration terrestre, cela implique que d'une façon tout à fait paradoxale eu égard la façon dont nous percevons le monde, des objets tels que de la 'lave en fusion' ou un 'bloc de 'glace' qui sont perçus par le capteur du robot (une caméra, par exemple) sont en fait physiquement *indiscernables* du point de vue de son contrôleur qui est le 'cerveau' du robot.

Ce « théorème d'indiscernabilité », pour faire court, s'applique à tous les niveaux de la matérialité – macroscopiques ou microscopiques/quantiques –, indépendamment des lois physiques qui régissent ces domaines. Il peut être établi[2] en analysant la nature des liaisons physiques qui doivent être mises en place entre le *capteur* et l'*afficheur* d'un appareil de mesure dont la fonction essentielle est de déterminer les propriétés spécifiques des objets du monde avec lesquels un système physico-chimique interagit.

Cet état *d'indiscernabilité*, primordial, des entités matérielles a été jusqu'alors complètement ignoré des chercheurs pour lesquels la *discernabilité* des objets macroscopiques ou microscopiques perçus/mesurés, allait de soi et qu'il était de ce fait aucunement nécessaire de se poser la question du bien-fondé d'une telle affirmation. Le 'théorème d'indiscernabilité' est à rapprocher de l'analyse de la vérité des propositions dans le domaine du langage qu'avait fait le logicien Ludwig Wittgenstein[3] :

- Dans son ouvrage le « Tractatus Logico-philosophicus », Ludwig Wittgenstein affirme, en déclinant l'ensemble de toutes les propositions possibles construites à partir de descripteurs élémentaires, qu'on ne peut pas faire la différence entre un état d'une chose, et un état complètement différent de cette chose comme le fait pour une pierre d'être lourde ou légère. Autrement dit, que les différents états de cette pierre dont nous devisons communément sont strictement *indiscernables*, à l'instar des objets 'lave en fusion' et 'bloc de glace' perçus par le robot d'exploration.

Puisque les *formes* infiniment variables des objets du monde perçues par le robot au cours de ses déplacements dans l'espace, sont ainsi strictement *indiscernables*, il en résulte qu'aucune *catégorisation cohérente* de ces formes ne peut être effectuée, sauf aléatoirement ; ce qui, statistiquement, est infiniment improbable du fait que les formes perçues varient en permanence lorsque le temps s'écoule. À ce titre, un robot qui serait construit à partir des seules mises en relations (mécaniques, électroniques, informatiques, chimiques,...) de composants techniques gérés par les lois physiques, ne pourrait pas être *autonome*, artificiellement *vivant*, dans un environnement multiforme, perpétuellement variable.

Le connexionnisme - les réseaux de neurones

Nombreux sont les chercheurs dans les domaines de la robotique et de l'intelligence artificielle, pour lesquels il est définitivement acquis qu'il existe des 'réseaux connexionnistes' à *auto-apprentissage* qui sont donc, par définition même du processus d'*apprentissage*, capables d'*auto-classer* dans une même *catégorie* les différentes formes d'objets perçues par un robot. À ce titre, le contrôleur d'un robot d'exploration doit avoir naturellement – sans aucune assistance humaine – la capacité de créer des *catégories cohérentes* d'actions telle que {*fuir* tous les objets chauds}, ce qui doit rendre ce robot totalement *autonome* à l'instar des *êtres vivants*.

Ces 'réseaux connexionnistes' sont composés de neurones artificiels (ou formels) qui s'inspirent du fonctionnement des neurones biologiques. Ces neurones artificiels possèdent

[2] Voir Annexe

[3] Ludwig Wittgenstein – *Tractatus Logico-philosophicus* - articles 5.101, 5.135, 5.15, 5.151

ainsi plusieurs entrées (‘dendrites’ des neurones biologiques) et une seule sortie (‘axone’ des neurones biologiques).

Comme preuves expérimentales du bien-fondé de l’apprentissage *non-supervisé* des réseaux de neurones artificiels, sont très souvent cités les dispositifs informatiques suivants : ‘l’Informon d’Uttley’, les ‘cartes auto-organisatrices de Kohonen’, les ‘réseaux de Hebb’, ainsi que la récente technique d’*auto-apprentissage* du ‘deep learning’ (‘apprentissage profond’) qui fait appel à des ‘réseaux connexionnistes’ comportant un très grand nombre de couches de neurones artificiels.

Compte tenu de l’extraordinaire efficience que possède la machine ‘deep learning’ de *reconnaître*, de *classer*, sans que celle-ci ait fait l’objet d’un apprentissage supervisé par un ‘professeur’, les objets qui sont soumis à sa rétine d’entrée (une caméra, par exemple), le ‘deep learning’ serait ce traitement de l’information longtemps attendu qui fonderait naturellement notre *cognition* et par là même celle des futurs robots *autonomes*. La *cognition* étant comprise comme l’ensemble des processus mentaux qui permettent aux êtres humains d’acquérir de la connaissance à partir de la perception des différents objets du monde.

Mais qu’en est-il réellement de cette capacité d’apprentissage *non-supervisé* que posséderaient ces différents réseaux que nous venons de mentionner compte tenu des arguments négatifs que nous avons développé à leur encontre à propos d’un robot d’exploration terrestre qui serait *autonome* :

— ‘L’Informon d’Uttley’ est un réseau de neurones qui est cité par le physicien Henri Atlan[4] comme étant assurément *auto-organisé*, à *auto-apprentissage*. À ce titre, eu égard l’existence de ce processus d’*auto-apprentissage*, il fait l’hypothèse que ce réseau pourrait constituer un modèle plausible de l’activité mentale du cerveau. Pour les roboticiens, cela pourrait donc être aussi l’élément de base d’un contrôleur doté d’une intelligence artificielle véritablement *autonome*.

Ce réseau est ainsi capable selon H. Atlan d’*apprendre* puis de *reconnaître* sans l’aide d’un opérateur les différents objets (pommes et oranges, par exemple) qu'on lui présente successivement sous la forme de deux ensembles dissemblables : un ensemble A constitué de plus de pommes que d’oranges, un ensemble B constitué de plus d’oranges que de pommes.

Mais l’analyse fonctionnelle de l’Informon montre que l’*auto-apprentissage* qui est supposé se développer spontanément ne peut en fait se réaliser que dans la mesure où l’opérateur, insuffisamment attentif dans l’établissement du protocole expérimental, *prépare* soigneusement l’expérience. Il forme ainsi deux ensembles distincts A et B d’apprentissage dans lesquels prédominent respectivement les pommes et les oranges. En l’absence de cette *préparation* précise des ensembles d’apprentissage, l’expérience et le calcul montrent que la *reconnaissance* ultérieure des pommes et oranges en deux catégories distinctes ne peut être qu’*aléatoire*.

L’*auto-apprentissage* de ce réseau n’est donc pas logiquement fondé. L’opérateur en charge de ce réseau doit préparer l’expérience en fonction des calculs statistiques que lui-même a implantés dans les modules du réseau.

— Les ‘cartes auto-adaptatives/auto-organisatrices’ de Kohonen[5] sont des réseaux de neurones développés par le physicien Teuvo Kohonen. Ces cartes sont également très souvent mentionnées par les chercheurs comme étant fondées sur des méthodes d'apprentissages *non-supervisés*. C’est ainsi que le neurobiologiste Gerald Edelman[6] se réfère à ce type de ‘cartes auto-adaptatives’ pour justifier l’existence dans le cerveau des êtres

[4] Henri Atlan – *L'organisation biologique et la théorie de l'information* – Hermann

[5] Kohonen – Algorithme *de Kohonen : classification et analyse exploratoire des données* – CNRS Samos Université Paris1

[6] Gérard Edelman – *Biologie de la conscience*, p. 109 – Odile Jacob

humains de ce qu'il appelle les 'cartes neuronales', lesquelles cartes expliqueraient le développement de processus d'*auto-organisations* qui fondent la *cognition.*

Ces 'cartes auto-adaptatives' seraient en effet capables de regrouper spontanément dans trois zones distinctes de leur organe de sortie (un écran vidéo, par exemple) chacun des trois éléments fondateurs de phrases courtes – *sujet*, *verbe*, *complément* – telle que « le singe aime les bananes » qui sont successivement présentées à son organe d'entrée ou rétine (une caméra vidéo, par exemple).

L'analyse fonctionnelle de ce dispositif montre cependant que c'est un technicien, et non pas un dispositif physique (mécanique, électronique), qui en observant l'écran de sortie du réseau, déclare qu'il existe des regroupements distincts des éléments *sujets*, *verbes* et *compléments* dans trois zones différentes de cet écran. Alors qu'en en réalité ces éléments sont physiquement *indiscernables* en raison du 'théorème d'indiscernabilité' (cf. § Une définition du vivant – le « théorème d'indiscernabilité »).

Ces regroupements, ou catégories, en trois zones distinctes de l'écran n'ont en fait d'existence que dans l'esprit de l'opérateur. Sans son intervention, ce réseau ne fait que transposer les relations d'*ordre* du domaine objet (ensemble des phrases) au domaine réseau (écran de sortie) sans pour autant créer de *catégories*. L'apprentissage de ces 'cartes auto-adaptatives' implique donc, là aussi, la supervision d'un opérateur sans lequel aucune *action cohérente* ne peut se produire.

L'opérateur a en effet la capacité empirique de pouvoir regrouper, parce que cela lui fait *plaisir*, les images relatives aux sujets, aux verbes et aux compléments qui apparaissent en différents endroits de l'écran vidéo de sortie alors que ces images sont physiquement *indiscernables* et qu'à ce titre elles n'ont pas d'existence physique intrinsèque en tant que sources d'*actions* différenciées.

– 'Les réseaux de Hebb' sont également cités comme étant des réseaux capables d'*apprendre* d'une façon *non-supervisée*. Partant de l'idée, tirée de l'observation du fonctionnement des neurones de notre cerveau, que deux neurones en activité au même moment créent ou *renforcent* leurs connexions, on propose successivement au réseau dont les poids synaptiques ont été réglés adéquatement, différents objets appartenant à une base d'apprentissage.

Si pour un objet donné, la sortie du réseau est conforme à la valeur prédéterminée inscrite dans la base, l'algorithme passe directement à l'instruction suivante et un autre objet peut alors être présenté au réseau ; si au contraire, l'objet qui est soumis au réseau induit une valeur de sortie qui n'est pas conforme à celle qui est inscrite dans la base, l'algorithme corrige alors automatiquement les poids synaptiques du réseau suivant la loi de *renforcement* de Hebb, puis on passe aux objets suivants. On réitère éventuellement cette phase d'apprentissage, qui est dite *non-supervisée*, jusqu'à ce que, pour chaque objet présenté, la sortie correspondante soit conforme à la valeur qui est inscrite dans la base.

L'*apprentissage* d'un réseau de Hebb est qualifié de *non-supervisé* parce qu'il est ainsi laissé libre de converger vers n'importe quel état final lorsqu'on lui présente un objet donné. Alors que pour un apprentissage *supervisé* classique, on impose au contraire une valeur déterminée à la sortie du réseau pour chaque objet nouveau qui lui est présenté. Cette désignation est cependant tout à fait impropre, car en réalité le réseau a été attentivement *préparé* par un opérateur qui a implémenté un algorithme tel que si pour un objet donné l'état final du réseau diffère de celui qui correspond à la base d'apprentissage, une instruction adéquate corrige automatiquement les poids synaptiques du réseau suivant la loi de *renforcement* de Hebb jusqu'à ce que l'état final soit identique à celui de la base.

Autrement dit, le réseau de Hebb est intégralement *supervisé* par un opérateur extérieur.

— Le ‘deep-learning’ est une méthode d’*auto-apprentissage* implémenté sur une machine informatique comportant un très grand nombre de neurones artificiels distribués en couches multiples (jusqu’à quelques centaines de couches).

Après avoir fait l’objet d’une longue phase d’apprentissage dite *non-supervisée*, durant laquelle lui ont été présentées une multitude d’images comprenant toutes sortes d’objets, dont des chats, cette machine est alors dit-on capable de découvrir, par elle-même, le *concept* {chat}. Ceci par le seul fait que parmi les N neurones de sortie, seul le neurone Nc se trouve spontanément activé lorsqu’on montre un véritable chat à sa ‘rétine’ (caméra).

Mais à l’instar des images *sujets*, *verbes* et *compléments* qui sont affichées sur l’écran de sortie des ‘cartes auto-adaptatives’ de Kohonen, les N sorties de cette machine, dont la sortie Nc qu’un technicien signale comme étant activée, sont en fait physiquement *indiscernables* du point d’un quelconque dispositif matériel produisant une *action*.

Dire que la machine ‘Deep-learning’ a découvert le *concept* de {chat} est donc tout à fait inapproprié. C’est seulement le technicien qui observant simultanément le chat assis devant la machine et la sortie Nc qui est activée ne fait qu’alors évoquer ses propres connaissances en matière de chat.

Cette hypothèse forte, jamais discutée, que des ‘réseaux connexionnistes’ peuvent naturellement s’*auto-organiser*, autrement dit créer des *catégories cohérentes* en l’absence de tout opérateur humain, est donc totalement infondée.

Apprentissage par récompenses

En complément des différents *réseaux connexionnistes* dans lesquels sont interconnectés un grand nombre de composants élémentaires comme les neurones artificiels, il faut également mentionner ces autres mécanismes de contrôle qui s’appuient quant à eux sur ce qu’il est convenu d’appeler des « apprentissages par récompenses (Q learning algorithms) ». Ces mécanismes ont, semble-t-il, cette vertu essentielle en matière de *cognition* et de *vie* artificielle, de permettre la réalisation de robots *autonomes* qui, sans aucune assistance humaine, sont par exemple capables d’avancer non pas en zigzag, mais toujours aller en marche avant.

Cette technique de contrôle qui est inspirée par la Nature (*biomimétisme*) est la suivante : le contrôleur du robot reçoit en l’occurrence une ‘récompense’ numérique *positive* toutes les fois que le robot avance et une ‘récompense’ *négative* quant le robot recule.

Prenons ainsi l’exemple d’un petit robot mobile à quatre roues qui est motorisé de la façon suivante : un bras articulé à deux degrés de liberté constitué de deux parties mobiles dont l’une se termine par un grappin, fait avancer ou reculer le véhicule en ligne droite sur une petite distance lorsque le grappin accroche le sol. La tâche que le chercheur assigne au robot sous la forme d’un algorithme d’*apprentissage* implémenté dans le contrôleur, est qu’à terme celui-ci se déplace uniquement en marche avant.

Pour ce faire, le contrôleur du robot reçoit une ‘récompense’ numérique positive toutes les fois que le robot avance et une ‘récompense’ négative quant le robot recule. Toutes ces ‘récompenses’ successives sont additionnées entre elles et les mouvements du robot qui sont conservées sont alors ceux qui maximisent la somme des récompenses.

L’expérience montre que quelle que soit la nature du sol (rugueux, sec, humide), le robot finit toujours par ne plus se déplacer qu’en marche avant. Et le technicien d’en conclure : le robot a appris seul à se déplacer uniquement en marche avant alors qu’aucune information sur la nature de son environnement ne lui avait été fournie. Le robot aurait ainsi *auto-appris* à se déplacer préférentiellement en marche avant dans un environnement dont il n’avait a priori aucune connaissance.

Mais cette analyse est totalement infondée, ce robot n’est qu’un *automate* qui n’a obéit qu’aux consignes numériques précises de l’algorithme qui avait été implémenté initialement dans son

contrôleur par un technicien. Le déplacement final en marche avant était prévisible. Sans technicien, aucun mécanisme de traitement de l'information ne permet en effet de calculer les 'récompenses' qui sont fonctions des distances positives ou négatives variables parcourues par le robot, puisqu'en raison du 'théorème d'indiscernabilité des objets' les différents descripteurs des états du robot (positions variables du bras articulé, distances variables parcourues par le robot) sont strictement indiscernables en vertu du 'théorème d'indiscernabiité'.

Contrairement aux apparences, l'*apprentissage* du petit robot est donc totalement *supervisé* par le technicien, en l'absence duquel les déplacements de ce robot seraient strictement aléatoires quelle que soit la durée de l'expérimentation.

Ce mécanisme « d'apprentissage par récompenses et punitions » qui est à la base même de l'éducation des êtres humains, ne permet donc pas la réalisation de robots pleinement *autonomes*, artificiellement *vivants*.

Structures dissipatives

Il existe des objets étranges étudiés par Illya Prigogine (prix Nobel de physique), ce sont les « structures dissipatives, des systèmes ouverts loin de l'équilibre thermodynamique », qui auraient la propriété essentielle de s'*auto-organiser*, donc de créer des *catégories cohérentes* en l'absence de tout opérateur humain ce qui irait manifestement à l'encontre de ce que nous venons de soutenir à propos des 'réseaux connexionnistes' en matière d'*auto-organisation*. Un système étant dit dissipatif lorsqu'il échange de l'énergie ou de la matière avec son environnement.

L'affaire est d'importance, car l'existence de tels objets expliquent pour certains chercheurs l'apparition 'automatique' des êtres vivants et de leur cognition.

En tant que « structure dissipative », l'objet 'tourbillons (ou cellules) de Bénard', est très souvent cité par les biophysiciens. On peut aisément réaliser un tel objet dit *auto-organisé* en chauffant de la paraffine dans un récipient cylindrique jusqu'à ce que celle-ci soit entièrement fondue. Puis après quelques minutes on arrête le chauffage. Lorsque la paraffine est figée, ce qui est un état correspondant à une véritable photographie du phénomène, on découvre alors que le récipient est occupé par des cellules de convection hexagonales. Ce sont les célèbres 'tourbillons de Bénard' qui évoquent étonnamment des structures créées par des êtres vivants comme les rayons d'une ruche d'abeilles constitués par des cellules hexagonales de cire. Ce qui pourraient justifier leur qualification de structures auto-organisées.

Mais il ne s'agit pas là d'un objet *auto-organisé* comme cela est toujours affirmé, qui serait la réponse attendue à la problématique de l'*autonomie* des structures artificiellement ou naturellement vivantes. Ce n'est qu'un objet *ordonné* en ce sens qu'il est le fruit obligé des interactions stéréotypées entre éléments eu égard les lois physiques. Si la source de chaleur change en effet de position par rapport au récipient, les conditions initiales étant modifiées les cellules de Bénard vont disparaître instantanément, l'objet va 'mourir'.

Pour que l'objet 'cellules de Bénard' soit véritablement *auto-organisé*, artificiellement *vivant*, et non pas seulement *ordonné*, il faudrait que suite au déplacement fortuit de la source de chaleur par rapport au récipient contenant la paraffine, l'objet en question serait lui-même capable d'effectuer une action correctrice telle que la source de chaleur se trouverait de nouveau en bonne position par rapport au récipient de façon que soit assuré la pérennité des cellules. Mais ce mécanisme impliquerait que l'objet soit lui-même capable de *discerner*, afin de pouvoir agir, les différentes positions de la source chauffage par rapport au récipient. Ce que nous savons être impossible en raison du 'théorème d'indiscernabilité'.

La même chose peut être dite pour la très spectaculaire 'réaction oscillante de Belousov-Zhabotinsky' qui se produit dans une solution d'ions bromate acidifiée par l'acide citrique

laquelle change périodiquement de couleur avec une grande régularité. Mais c'est pareillement une structure fluide *ordonnée* et non pas *auto-organisée*.

Quelque soit leur complexité, les « structures dissipatives » sont *ordonnées* et non pas *auto-organisées*. Soutenir que des structures dissipatives sont à la base de processus d'*auto-organisation* est complètement infondé.

Sélection évolutive

Pour d'autres chercheurs en robotique, une façon pragmatique de rendre un robot *autonome*, artificiellement *vivant*, ce n'est pas d'analyser la nature de tous les évènements, auxquels un robot *solitaire* peut être confronté, ce qui est une tâche insurmontable étant donné que ces événements sont en nombre infini. C'est de plutôt s'inspirer du mécanisme mis en œuvre dans la « sélection naturelle darwinienne » qui expliquerait l'émergence des fonctions d'animation dont sont munies certaines structures physico-chimiques alors qualifiées de *vivantes*, donc éminemment *autonomes*. C'est ce qu'il est convenu d'appeler la « robotique évolutionniste ».

Rappelons que la *sélection évolutive darwinienne* est essentiellement basée sur la transmission entre générations successives d'êtres *vivants*, des propriétés pérennisantes fortuites engrammées dans des *gènes*. À ce titre, le mécanisme de la *sélection naturelle* est donc essentiellement fondé sur le processus de *reproduction fonctionnelle* « mère → fils ».

Pour réaliser une expérience de « robotique évolutionniste », on commence par constituer une flotille composée de plusieurs robots qui sont chacun dotés d'un *réseau de neurones artificiels* en tant que contrôleur, dont les *poids synaptiques* (valeurs des forces de liaisons physico-chimiques qui existent entre les neurones) initialement tous différents (aléatoirement distribués), sont les *gènes artificiels* de ces robots.

On soumet alors tous les robots de la flotille à une tâche donnée : par exemple atteindre une cible en un temps minimal tandis que le domaine d'essais est jonché de gros obstacles qui peuvent bloquer leurs déplacements.

Après un premier essai, on conserve le robot qui est déclaré vainqueur de l'épreuve car ayant évité, fortuitement, les obstacles qu'il a rencontrés. Puis on recopie, en faisant volontairement quelques erreurs (des *mutations* fortuites), ses différents *poids synaptiques* (valeurs et positions) vers les réseaux de neurones des robots 'perdants' qui n'ont pas atteint la cible. On effectue plusieurs fois cette opération en gardant à chaque fois le robot gagnant.

Finalement, on constate qu'il existe alors un (ou plusieurs) robot qui a atteint la cible en évitant tous les obstacles rencontrés. Autrement dit, ce robot a créé, seul semble-t-il, la fonction qui le rend *autonome* dans l'environnement prescrit, soit {éviter tous les obstacles pour atteindre la cible}.

Mais ce qu'aucun chercheur n'a semble-t-il vu jusqu'alors, ni même le physicien von Neumann avec sa théorie des systèmes *auto-reproducteurs*[7]**,** c'est que le processus de *reproduction fonctionnelle* « mère → fils », celui-là même qui permettrait la recopie des *poids synaptiques* (valeurs et positions) du robot 'gagnant' – lesquels mémorisent les relations pérennisantes entre les *objets perçus* et les *actions* qui en ont résultées – vers les réseaux de neurones des robots 'perdants', est logiquement interdit en raison du 'théorème d'indiscernabilité'.

Par contre, une reproduction de type 'empreinte', comme le matriçage des différents creux et bosses des pistes d'un dvd, est quant à elle tout à fait licite eu égard le 'théorème d'indiscernabilité' (cf. § Une définition du vivant – le « théorème d'indiscernabilité »).

En raison de l'interdiction qui porte ainsi sur la faisabilité de la *reproduction fonctionnelle* « mère → fils », la recopie des poids synaptiques du robot 'gagnant' vers les réseaux de

[7] John von Neumann – *Theory of self-reproducing automata* **–** University of Illinois Pre**ss (**1966)

neurones des robots ‘perdants’ ne peut donc être qu’*aléatoire*. Ce qui, par là même, infirme le processus de *sélection évolutive* en tant que mécanisme permettant la création incrémentale de *catégories cohérentes* qui fonderait l’*autonomie* du robot.

De récentes expérimentations de *robotique évolutionniste* portant sur des générations successives de robots, semblent cependant montrer qu’un certain nombre de ces robots peuvent spontanément s’*auto-organiser*, c'est-à-dire devenir *autonomes*. Alors que la chose est impossible pour des robots solitaires en raison du ‘théorème d’indiscernabilité’). Par robots *solitaires*, il faut entendre des robots qui n’interagissent avec aucun autre robot et qu’à ce titre la formation de *catégories cohérentes* qui devraient les rendre *autonomes* ne dépend que de leur seule capacité.

Mais si certains robots sont effectivement devenus *autonomes*, sachant par exemple s’éloigner de toutes les sources de chaleur qui pourraient les détruire, c’est parce que leurs concepteurs, insuffisamment vigilants, ont, en tant qu'êtres *vivants*, injecté leur propre capacité en matière d’*autonomie*. Ils ont ainsi, inconsciemment, *préparé* les divers mécanismes de *recopie fonctionnelle* « mère → fils » en sélectionnant les différents domaines mémoires autrement *indiscernables* (domaines mémoires correspondant aux différents *poids synaptiques* des réseaux de neurones constituant le contrôleur du robot).

En ne pouvant s’appuyer que sur des reproductions de type ‘empreinte’ – les *reproductions fonctionnelles* étant ainsi interdites –, la « sélection naturelle darwinienne » que nous savons être constamment à l’oeuvre dans la nature en créant une multitude de nouvelles espèces, ne ferait que sélectionner parmi les différentes formes possibles de systèmes matériels déjà *autonomes*, *vivants*, celles qui seraient les mieux adaptées pour survivre aux contraintes environnementales. Cette théorie, sans donc expliquer le *vivant*, ayant cependant l’immense mérite d’expliquer l’apparition naturelle de toutes les *formes du vivant* qui sont apparues sur Terre.

En conclusion, ni le *connexionnisme*, ni le mécanisme de *sélection évolutive*, ni les *structures dissipatives*, ni l’*apprentissage par récompenses et punitions*, ne sont donc des réponses possibles à la question de l’autocréation de *catégories cohérentes* qui fondent les systèmes *autonomes*, *artificiellement vivants*. Un robot construit à partir des seules mises en relations (mécaniques, électroniques, informatiques, chimiques,...) de composants techniques gérés par les lois physiques, ne peut donc pas être *autonome*, *artificiellement vivant*, dans un environnement protéiforme, infiniment changeant. Ce robot reste un automate plus ou moins efficient dans un environnement qui a été spécifiquement défini par son constructeur.

« Et pourtant, elle tourne… »

Empiriquement, on sait cependant que pour assurer sa propre survie un technicien – un système physico-chimique complexe – est lui tout à fait capable de fuir naturellement différents objets chauds qui pourraient le détruire comme de la lave en fusion, un feu de forêt, une nappe de pétrole enflammée, et de créer ainsi une *catégorie cohérente* des objets avec lesquels il va interagir, soit {*fuir* tous les objets chauds}.

Grâce à cette capacité fonctionnelle, d’ailleurs inintelligible en vertu du ‘théorème d’indiscernabilité’ qui s’applique indistinctement à toute structure physico-chimique, ce même technicien peut alors *superviser* le contrôleur de notre robot d’exploration en mettant en place des liaisons *cohérentes* entre son capteur et son actionneur afin que ce robot puisse en l’occurrence *fuir* automatiquement tous les objets chauds qui pourraient le détruire.

Un robot qui serait de la sorte préparé par un technicien serait *autonome* tant que celui-ci superviserait son contrôleur. Et c’est ainsi qu’une voiture, un dispositif physico-chimique complexe, devient *autonome*, artificiellement *vivante*, à partir du moment où un conducteur prend les commandes.

En l’absence du technicien, le robot ne serait plus qu’un automate efficient dans un monde limité aux seuls objets hostiles spécifiés par ce dernier. Avec des contraintes environnementales

inédites, il faudrait que de nouveau le technicien intervienne en créant dans le contrôleur du robot des *catégorisations cohérentes* étendues à ces nouvelles contraintes.

Devant cet obstacle physiquement insurmontable de l'autocréation de *catégories cohérentes* qui fonde l'état de *vie* artificiel du robot, il faut alors se poser la question principielle suivante :

Qu'est-ce qui différencie alors le contrôleur du robot du cerveau d'un technicien, sachant que le technicien peut, lui, régler le contrôleur du robot à sa convenance afin créer des *catégories cohérentes* alors que ces *catégorisations* sont physiquement impossibles étant donné le 'théorème de l'indiscernabilité ?

La conscience

Il existe une réponse possible, expérimentale, à cette situation paradoxale : ce qui différencie le contrôleur d'un robot du cerveau d'un technicien, c'est la *conscience* que possède ce technicien. Cette faculté que tous nous possédons qui nous ouvre à la vision colorée et sensible du monde avec lequel nous interagissons.

L'expérience montre en effet que la *conscience* que possède le technicien a cette propriété singulière de discriminer des objets comme de la 'lave en fusion' et un 'morceau de glace' qui, eu égard le 'théorème d'indiscernabilité' sont pourtant physiquement *indiscernables*.

Cette discrimination des objets du monde par le technicien vient essentiellement de ce que cela « *fait mal* » ou « *fait plaisir* » à ce dernier suivant l'état, actuel ou mémorisé dans son système nerveux, de sa structure physique. C'est ainsi qu'ayant eu une rage de dents insoutenable dans sa jeunesse – avant tout *apprentissage* de ses parents en la matière –, le technicien avait rapidement consulté par *hasard* un dentiste au lieu d'aller chez son fleuriste... deux destinations, en tant qu'objets, qui étaient pourtant physiquement *indiscernables* au même titre que les différents objets du monde qui sont perçus par le robot.

La *conscience* n'est donc pas un épiphénomène, un phénomène accessoire qui en l'occurrence aurait gratuitement accompagné l'irritation mécanique de la dent du technicien. Empiriquement opératif, la *conscience* choisit les solutions techniques – élaborées puis mémorisées dans le cerveau du sujet – qui sont porteuses de *plaisir*, ou de son équivalent opératif une *diminution de la douleur*, car ayant dans le passé assuré *fortuitement* la pérennité du sujet, son état de vie. En l'occurrence, le *plaisir* ou la *diminution de la douleur* qui 'étiquette' la solution *dentiste* résulterait ainsi de la diminution rapide de la douleur dentaire que le technicien avait éprouvée dans le passé lorsque, *fortuitement*, il avait été soigné par un dentiste.

Bien qu'étant empiriquement opérative, la *conscience* est formellement *irréductible* à toute interaction physico-chimique neuronale en raison du 'théorème d'indiscernabilité'. Elle est donc de nature *non-physique* ; si la *conscience* était en effet de nature matérielle, elle serait de ce fait réductible à des interactions physico-chimiques particulières et à ce titre elle pourrait être définie par des attributs duals complémentaires tels que chaud/froid, blanc/noir,… comme toute entité physique soumise à une mesure. Dans ces conditions, le calcul des prédicats (ou descripteurs) appliqué au nouveau système {capteur + conscience} montrerait que l'ajout de cette *conscience* au système capteur dont les états étaient originellement *indiscernables*, ne ferait qu'augmenter le nombre des états perçus par le système capteur sans pour autant réduire l'état d'*indiscernabilité* de ses états. La *conscience* est donc bien de nature *non-physique*.

Cette hypothèse relative à la nature non-matérielle de la *conscience* n'est pas physiquement irrecevable comme on pourrait le penser de prime abord en s'appuyant sur le postulat adopté par la communauté scientifique que la 'dimension matérielle' est la seule 'dimension' possible de l'univers. Ladite 'dimension matérielle' n'est pas en effet un 'objet' qu'on peut observer par le truchement de mesures. Ce n'est qu'un *concept* qui résulte d'un

processus de *catégorisations cohérentes* qui, comme nous l'avons montré, implique l'existence de la *conscience* qui est, elle, nécessairement irréductible à toute interaction physique en raison du 'théorème d'indiscernabilité'. Cette hypothèse de la non-matérialité de la *conscience* serait donc fondée.

Le rôle de la *conscience* serait donc essentiellement de *choisir* les solutions techniques qui assurent la pérennité du sujet avec lequel elle est associée, parmi toutes celles – physiquement *indiscernables* – qui sont élaborées spontanément au cours des interactions physico-chimiques entre les neurones de notre cerveau-ordinateur. La *conscience* n'élaborerait ainsi aucune solution technique, elle ne posséderait aucune connaissance a priori sur les objets du monde.

La nature des processus de *décision* qui se développent dans le cerveau du technicien par l'entremise de la *conscience* dont il est empiriquement muni, sont en accord avec les résultats paradoxaux des expériences du neurobiologiste Benjamin Libet[8] :

> […] la conscience peut opposer son 'veto' aux solutions présentées (unconscious brain activity) qui ont été préalablement élaborées quelque 500 millisecondes plutôt par le cerveau-ordinateur [des solutions techniques issues d'interactions obligées entre des entités possédant entre eux une affinité physico-chimique spécifique].

En définitive, pour pouvoir être véritablement *autonome*, *artificiellement vivant* – ce qui était notre questionnement initial – un robot construit à partir des seules mises en relations (mécaniques, électroniques, informatiques, chimiques,...) de composants techniques, devrait être, lui aussi, muni d'une *conscience*.

Plus généralement, il en résulte alors que l'émergence des êtres vivants impliquerait qu'ils soient chacun munis d'une *conscience*. La *conscience* serait ainsi la clé du *vivant*. L'analyse de l'animation de la bactérie E. coli, un être vivant élémentaire, va illustrer cette thèse.

Contrairement à ce qui est souvent allégué, la « biologie synthétique » ne répond pas à cette question fondamentale qu'est la création des structures vivantes à partir de l'inerte. Avec une première tentative de construire de novo un système vivant[9], les ingénieurs biologistes n'ont en effet que combiné, synthétisé, des organites déjà fonctionnels tels que les *ribosomes*. Des composants élémentaires présents dans les cellules de tout organisme, qui ont le rôle essentiel de déchiffrer le code ARN qui induit par le biais de *recopie fonctionnelle* (cf. § Sélection évolutive) la synthèse des protéines et à ce titre, possédant déjà la capacité physiquement paradoxale de discerner, à fin d'actions pérennisantes, des entités autrement *indiscernables*.

De la bactérie à l'homme

Pour qu'une structure physico-chimique soit vivante ou artificiellement *vivante* comme le serait un robot *autonome*, il faut donc qu'elle soit directement ou indirectement (un technicien supervise continûment un robot) dotée d'une *conscience* qui, fondamentalement, assure la pérennité de la structures avec laquelle elle est associée.

Ce qui permet l'animation d'une bactérie comme E. coli, en tant que structure physicochimique élémentaire éminemment *vivante*, va illustrer ce propos.

La fonction nommée « chimiotactique » que possède la bactérie est fondamentale, elle lui permet de se diriger préférentiellement vers des zones où il y a une forte concentration de molécules de nutriment (du glucose, par exemple) nécessaire à son dynamisme, mais aussi en l'éloignant de zones où il y a des molécules comme le phénol qui altéreraient sa structure[10].

---

[8] Libet Benjamin – *Unconscious cerebral initiative and the role of conscious will in voluntary action. Neurophysiology of Consciousness,* pp. 269-306 – Contemporary Neuroscientists 1993.

[9] J. Craig Venter – *Le vivant sur mesure* -2014

[10] Sept 25 Biochemical Networks – *Chemotaxis and Motility in E. coli*.

Pour faire court, la chaîne « chimiotactique » est composée des éléments suivants : des *capteurs membranaires* MCP (protéines) qui, en particulier, calculent le *gradient* du glucose dans le milieu (variations locales de la concentration), d'un *flagelle* lié à un moteur moléculaire qui peut tourner dans le sens direct ou rétrograde (rotation rétrograde : la bactérie se déplace en ligne droite ; rotation directe : la bactérie culbute et change ainsi de direction), d'une *protéine de contrôle* CheY qui détermine le sens de rotation du flagelle en fonction du gradient du glucose dans le milieu mesuré par les *capteurs* membranaires MCP.

L'analyse fonctionnelle de la *fonction chimiotactique* de la bactérie montre que pour être *animée*, son flagelle doit tourner dans le sens *rétrograde* toutes les fois que le *gradient* du glucose dans le milieu est positif et dans le sens *direct* lorsqu'au contraire il est négatif.

En effet, le sens *rétrograde* (sens inverse des aiguilles d'une montre) de rotation du flagelle conduit la bactérie à naturellement poursuivre son déplacement rectiligne et donc à se diriger vers des zones où il existe de plus en plus de molécules de glucose, puisque la variation mesurée de la concentration du glucose dans le milieu au cours du déplacement de la bactérie – ou *gradient* du glucose – est positive. Quant au sens *direct* de rotation du flagelle, il conduit la bactérie à culbuter sur elle-même (par ébouriffement des filaments qui constituent le flagelle) d'où un possible changement ultérieur de direction à explorer.

Mais en vertu du 'théorème d'indiscernabilité' qui interdit que la protéine de contrôle CheY puisse faire la distinction entre les valeurs positives ou négatives du gradient des molécules de glucose mesuré par les capteurs MCP, il en résulte que le processus « chimiotactique » de recherche des molécules de glucose ne peut, logiquement, être qu'*aléatoire*, et à ce titre totalement inefficient.

Les protéines capteurs MCP n'ont en effet aucune affinité physico-chimique particulière avec la protéine de contrôle CheY, qui pourrait conduire spontanément à l'émergence d'actions différenciées comme celles qui sont nécessaires pour assurer la capture efficace des molécules de glucose. Si contre toute attente, de telles affinités pouvaient néanmoins exister, toutes les actions résultantes (rotations directes ou rétrogrades du flagelle) seraient alors nécessairement récurrentes, stéréotypées, et de ce fait logiquement incompatibles avec la formation de *catégories cohérentes* qui fondent l'*animation* de la bactérie. En conclusion, la bactérie ne peut donc pas être *autonome, vivante* !!

Il y a cependant une solution empirique possible à cette situation paradoxale où la bactérie est pourtant une structure physique *autonome*, *vivante*, alors qu'aucun technicien ne vient la superviser comme cela était possible avec le robot : le contrôleur de la bactérie, la protéine CheY, doit être spécifiquement muni d'une *conscience* qui va permettre la création in situ de *catégories cohérentes* qui fondent son animation.

C'est parce que la bactérie *« aurait mal* lorsque son état énergétique (nombre de molécules de glucose disponibles) est très faible, et qu'au contraire elle « *aurait du plaisir* » lorsque son état énergétique est élevé, que la bactérie pourrait capturer d'une façon efficiente les molécules de glucose dispersées dans le milieu. En l'absence d'une *conscience*, la capture des molécules de glucose ne pourrait être que fortuite puisque pour le contrôleur, la protéine CheY, les différentes valeurs du gradient du glucose seraient *indiscernables* eu égard le 'théorème d'indiscernabilité' (cf. § Une définition du vivant – le « théorème d'indiscernabilité »).

Hypothèse raisonnée, la *conscience* qui, en raison du *théorème d'indiscernabilité,* est logiquement irréductible à toute interaction physicochimique entre molécules, serait induite par la *forme* spécifique de la protéine CheY, au moins en ce qui concerne la bactérie E. coli pour laquelle cette protéine CheY à un rôle essentiel de contrôle de la fonction chimiotactique.

À ce titre, certaines protéines (ou pseudo-protéines) possédant une *forme* adéquate pourraient être les premières structures vivantes apparues sur Terre car étant munies d'une *conscience* et possédant naturellement des capteurs (sites spécifiques) ainsi qu'une certaine motilité par déformation de leur configuration globulaire. Les *prions* (*protéines* « malformées » qui ne se reproduisent pas mais qui

provoquent la « déformation » de protéines saines avec lesquelles elles sont en contact (cf. la maladie de *Creutzfeldt-Jakob*) ne seraient-ils pas une illustration de cette hypothèse ?

Une *conscience* ne serait donc associée à la structure physico-chimique de la bactérie qu'à la condition qu'il existe certaines configurations matérielles spécifiques de ses constituants, en l'occurrence au niveau de la protéine CheY de la chaîne chimiotactique. Ainsi, bien qu'étant essentielle à l'émergence de tous les êtres vivants, de la *conscience* ne serait pas nécessairement présente dans tout l'univers comme le suggère le philosophe David Chalmers[11] :

> « [...] la conscience serait universelle. On la trouverait partout dans l'univers, des particules élémentaires jusqu'aux astres et galaxies. Dans le domaine de la biologie terrestre, elle serait de même présente de la bactérie jusqu'à l'homme. »

La *conscience* étant ainsi empiriquement localisée au niveau des protéines qui constituent l'élément de base de toutes cellules vivantes, en particulier des *neurones* qui sont les unités fonctionnelles élémentaires du système nerveux qui fonde la cognition, on peut alors formuler l'hypothèse : la *conscience* qui fonde l'animation des êtres humains, résulterait de la fusion à la fois *spatiale* et *temporelle* d'un grand nombre de *consciences* élémentaires dont leurs cerveaux seraient munis.

Nature opérative de la conscience

Le rôle essentiel de la *conscience* étant considéré, une question essentielle se pose alors : les interactions présumées entre l'opérateur *conscience* – irréductible à toute interaction physique en raison du 'théorème d'indiscernabilité' – et des structures physico-chimiques neuronales du cerveau du technicien ou de la protéine CheY de la bactérie, sont-elles licites ?

Hypothèse raisonnée, les interactions présumées entre l'opérateur *conscience* et les structures physico-chimiques du cerveau du technicien ou de la protéine CheY de la bactérie, doivent toutes se résoudre au 'niveau quantique' par la réduction *orientée* de l'état de superposition des « fonctions d'onde » (ondes de probabilité) qui représentent les états quantiques de ces structures physico-chimiques.

Point essentiel, lors de la *réduction de la fonction d'onde* ou transition quantique, qui conduit à l'émergence d'un état singulier (dit *standard*), celui que nous observons, il ne se produit qu'un simple réaménagement des énergies déjà existantes et de ce fait il y a globalement « conservation de l'impulsion-énergie»[12], cette loi physique fondamentale entre toutes.

De ce fait, à travers le processus de réduction de la fonction d'onde, bien que l'opérateur *conscience* soit strictement irréductible à toute interaction physique et échappant de ce fait à toute mesure physique, pourrait néanmoins contrôler les structures physico-chimiques du cerveau du technicien ou de la protéine CheY de la bactérie.

La capacité opérative de l'opérateur *conscience* serait donc licite eu égard les lois physiques contrairement à ce qu'affirme le philosophe des sciences Daniel Dennett dans son ouvrage *La conscience expliquée*[13] :

> Comment Casper le gentil fantôme [histoire d'enfant], peut-il à la fois passer à travers les murs et attraper une serviette qui tombe ? Comment la substance mentale [la *conscience*] peut-elle à *la fois* échapper à toute mesure physique et contrôler le corps ? Un fantôme dans la machine ne nous est d'aucune aide pour nos théories s'il ne peut mouvoir des choses autour de lui — comme un esprit frappeur bruyant qui peut renverser une lampe ou claquer une porte. Mais toute chose qui peut mouvoir une chose physique est elle-même une chose physique.

---

[11] David Chalmers – *L'esprit conscient, à la recherche d'une théorie fondamentale* - Ithaque 2010

[12] O.C. de Beauregard – *Le second principe de la science du temps*, p. 98 - Edition du seuil, Paris.

[13] Daniel Dennett – *La conscience expliquée*, p. 52 - Odile Jacob 1993.

L’existence

Les *consciences* ne feraient donc qu’effectuer des *choix* – fondés sur le *plaisir* d’exister – qui assureraient la pérennité des structures physico-chimiques auxquelles elles seraient associées. Elles ne seraient porteuses d’aucune connaissance a priori sur le monde, elles seraient « brutes », sans objet, comme la *douleur*, le *plaisir*, le *rouge*, le *salé*... La *conscience* ne serait ainsi jamais *conscience de quelque chose* comme le supposait Husserl[14].

De sorte qu’on peut alors conjecturer que la *conscience* des choses du monde serait semblable pour tous les *êtres vivants*, depuis la bactérie jusqu’à l’homme.

La fabuleuse diversité des *actions* mécaniques menées par l’homme, comparée aux activités élémentaires des bactéries, ne résulterait finalement que du fantastique accroissement du nombre des diverses solutions matérielles potentielles générées par son système nerveux central grâce aux propriétés de *généralisation* et d’*associativité* des réseaux de neurones. La bactérie, quant à elle, ne disposant que de quelques protéines associées en réseaux pour calculer les solutions mécaniques possibles devant lui permettre d’assurer la pérennité de sa structure. La richesse plus ou moins grande des *actions* menées par un être vivant ne seraient donc pas significative en matière de ce que « ressent » cet être vivant.

« L'existence » d’un être vivant résulterait essentiellement de la *représentation* du monde par le truchement des *qualités sensibles* (plaisir, douleur, couleur, son,..) dont la nature serait commune à tous. À ce titre, il y aurait une « existence bactérienne » au même titre qu'il y a une « existence humaine ».

Ce qui distinguerait « l'existence d'une bactérie » de celle de l'homme, c'est que la première serait très élémentaire car se composant de seulement quelques descripteurs proprio- ou extéroceptifs du monde « observés » par les *qualités sensibles*. Alors que « l’existence de l’homme » serait d'une extrême complexité, faisant intervenir un très grand nombre de descripteurs de ce monde en vertu de l'extraordinaire puissance de calcul et d'associativité des réseaux nerveux du cerveau de l'homme.

« L'existence » serait, à des degrés divers, partagée par tout ce qui vit, les processus cognitifs seraient de même nature pour tous les êtres vivants.

Le transhumanisme

Pour la pensée *transhumaniste*, la fonctionnalité des êtres vivants et de la *cognition* dont ils sont munis doivent toutes pouvoir se réduire à des ‘algorithmes’. C’est-à-dire à des ensembles de règles opératoires, d’instructions, s’appliquant au développement d’interactions physico-chimiques plus ou moins complexes à l’exemple des gaz oxygène et hydrogène qui se combinent pour former de l’eau.

À ce titre, nous ne serions que des *machines*, certes très perfectionnées, qui se seraient spontanément *auto-construites* en ayant la capacité physiquement paradoxale eu égard la dégradation entropique de l’univers, d’assurer en toutes circonstances la pérennité de leur structure, ce qui les caractériserait. Au cours du temps, ces machines *vivantes* se seraient dotées d’une intelligence plus ou moins développée grâce à des systèmes de traitement de l’information *auto-apprenants*.

La *conscience* que nous possédons tous naturellement qui a cette vertu essentielle de nous ouvrir à la perception colorée et sensible d’un univers autrement clos sur lui-même dans sa minéralité, ne serait de ce fait qu’un *épiphénomène* qui ne jouerait aucun rôle fonctionnel dans l’édification des êtres vivants.

---

[14] Edmund Husserl – *Une idée de la phénoménologie de Husserl : l’intentionnalité – Situations I, p. 32* - Paris, Gallimard, 1947

La pensée *transhumaniste* ne s'intéresse ainsi au *vivant* qu'en termes mécaniques ou computationnels. Soit essentiellement : réparer, améliorer, ces structures dites *vivantes* comme on construit des fusées, des ordinateurs, aux performances techniques de plus en plus grandes. Et puisque ces divers processus techniques seront certainement maîtrisés dans le futur, nous devrions alors un jour être capables de nous libérer des limitations physiques et mentales acquises au cours de notre évolution. C'est pourquoi les êtres humains pourraient devenir beaucoup plus intelligents et même quasi immortels.

Le but ultime de la mutation vers le transhumain étant d'éliminer tous les mécanismes de la *douleur* aussi bien physiques que mentaux, le téléchargement de l'esprit vers un super-ordinateur pourrait être ainsi la solution finale à tous nos tourments. Il suffirait pour ce faire de copier sur une 'clé' de grande capacité les différents états des neurones de notre cerveau pour ensuite les reporter sur la mémoire d'un super-ordinateur.

La thèse *transhumaniste* est ainsi un rêve pour certains mais en revanche un épouvantable cauchemar pour d'autres. Mais cette thèse est-elle fondée ?

Nous avons montré que loin d'être un épiphénomène, la *conscience* s'avère être la clé de l'existence sur Terre des êtres vivants et de leurs cognitions. Ce sont les *consciences* dont sont munis les êtres *vivants* qui déterminent la façon dont nous agissons sur le monde afin d'assurer notre pérennité, autrement dit pour être *vivant*. Loin d'être un épiphénomène, la *conscience* est un terme essentiel sans lequel aucune vie ne serait apparue sur Terre.

Cette *conscience* qui a la capacité physiquement paradoxale de faire des *choix* parmi des objets du monde, qui sont quant à eux physiquement *indiscernables*, implique logiquement que cette *conscience* est strictement irréductible à tout processus physique. Et c'est ainsi que la *conscience* ne peut pas résulter d'un algorithme et ceci quelle qu'en soit sa complexité. Ce qui va manifestement à l'encontre de la thèse transhumaniste puisqu'il est alors strictement impossible de réparer ou créer une *conscience* comme on construit un ordinateur ou qu'on greffe un morceau de tissu cardiaque reconstitué sur un coeur défaillant.

De sorte que même si les progrès de la science sont tels qu'on puisse un jour construire des structures physico-chimiques adéquates qui, expérimentalement, s'avéreraient posséder la capacité d'induire de la *conscience*, il n'empêche qu'étant donné l'irréductibilité fondamentale de la *conscience* à toute interaction physique on ne pourra jamais contrôler la capacité physiquement paradoxale que possède la *conscience* de sélectionner « pour son plaisir d'être en vie » des solutions techniques qui sont par ailleurs physiquement *indiscernables*.

Par ailleurs, pour qu'un être humain devienne plus intelligent, il faudrait qu'il soit capable de créer un grand nombre de nouveaux *concepts* qui, par définition, sont des *catégorisations cohérentes* des objets du monde qu'un être *vivant* perçoit avec ses capteurs afin d'agir d'une façon appropriée qui assure sa pérennité. Comme par exemple les *catégories d'actions cohérentes* ou *concepts* {'fuir' pour toutes températures *supérieures* à 30°} et {'avancer' pour toutes températures *inférieures* à 30°} qui doivent permettre d'éviter 'intelligemment' tous les objets chauds.

Mais pour créer de nouvelles *catégories cohérentes* il ne suffirait pas de seulement augmenter la capacité de sa mémoire comme le suggère la thèse *transhumaniste*. Il faudrait aussi que soient totalement maîtrisées les propriétés opératives de la *conscience*. Faute de quoi, les informations traitées par le cerveau, en tant qu'ordinateur, ne pourraient être en effet chargées dans la mémoire que d'une façon *aléatoire*, puisque ces informations seraient alors strictement *indiscernables* eu égard le 'théorème d'indiscernabilité', et donc sans aucune possibilité ultérieure de lectures ordonnées.

La thèse *transhumaniste* qui consiste à affirmer que la seule augmentation de la capacité de calcul de notre cerveau devrait permettre d'augmenter d'une façon illimitée l'intelligence des êtres humains est donc infondée.

La seule chose qui un jour sera sans doute réalisable, sera de modifier ou même de supprimer, en agissant mécaniquement ou chimiquement sur notre cerveau, l'*induction* de certaines 'qualités sensibles' ou constituants de la *conscience*, que sont le 'plaisir', la 'douleur', la 'couleur', le 'son',... Mais sans pour autant maîtriser la capacité de décision de la *conscience* qui fonde l'état de vie et par là même l'existence de tous les objets qui sont construits par les êtres *vivants*

À ce titre, la proposition des *transhumanistes* de construire un cerveau artificiel capable de *créer*, comme nous savons naturellement le faire, à partir des seules interactions mécaniques, informatiques, chimiques,... entre des composants techniques, est tout aussi infondé. Toute création implique en effet une *conscience*. Les solutions à un problème posé par un être humain que peut produire une machine dotée d'une intelligence artificielle ne sont que des combinaisons *aléatoires* produites à partir de processus préalablement implémentés par des opérateurs humains. Seuls ces opérateurs humains munis de leur *conscience* peuvent ensuite décider, afin d'assurer leur pérennité, de la pertinence de certaines des solutions techniques opérantes possibles proposées par la machine.

La *créativité* ce n'est pas tant en effet le pouvoir d'élaborer des *formes* nouvelles (cf. § Sélection évolutive) que d'associer ces *formes* d'une façon *cohérente* – alors qu'elles sont physiquement *indiscernables* – afin que les *actions* qui en résultent assurent la pérennité de structures physico-chimiques alors dites *vivantes* face aux sollicitations généralement dégradantes de leurs environnements. Les structures inertes ne *souffrent* pas d'être détruites, un rocher n'entreprend aucune action spécifique afin ne pas être cassé en deux morceaux.

Les machines informatiques ne font ainsi qu'élaborer en aveugle des objets numériques issues d'interactions obligées qui se développent spontanément eu égard les lois physiques. C'est la *conscience* dont est muni un sujet qui choisit un objet particulier parmi tous ceux élaborés par la machine, parce que cet objet technique permet, in fine, d'assurer sa pérennité eu égard les contraintes infiniment variables du milieu. La machine étant quant à elle indifférente à la pérennité de sa structure.

C'est ainsi que la machine 'Deep learning' n'a pas inventé, créé, le concept de {chat}, pas plus qu'une 'carte auto-adaptative Kohonen' les catégories *sujet*, *verbe* et *complément* (cf. § Le connexionnisme). Ce sont les opérateurs en charge de ces expérimentations qui ont créés ces différents concepts grâce qu'a la capacité que possède leur *conscience* de sélectionner adéquatement les objets – physiquement *indiscernables* – qui doivent assurer leur pérennité.

De même qu'il n'y a aucune *création* véritable de la part de la machine de moulage numérique qui produit les éléments emboîtables LEGO. C'est seulement un jeune enfant qui en associant ultérieurement pour son « plaisir » plusieurs éléments d'une boite qu'on lui a offerte, va réellement *créer* ce qu'il nommera une 'voiture' car pouvant rouler sur une table. Alors que ces différents éléments comme une roue et une plaque sont en fait strictement *indiscernables* par tout dispositif qui devrait les assembler. Les boites de LEGO ne sont en l'occurrence que des réservoirs de formes variées qui n'ont aucune valeur opérative en matière de *vivant*, d'*autonomie*.

Un autre exemple significatif : comme nous l'avons précédemment analysé (cf. § De la bactérie à l'homme), un flagelle associé à un moteur rotatif moléculaire permet à une bactérie comme E. coli de se diriger 'intelligemment' vers les sources de glucose dont elle se nourrit. Mais cet extraordinaire moteur moléculaire composé de plusieurs protéines disposées en anneaux ne résulte que d'interactions *obligées* entre protéines eu égard les lois physiques. À ce titre, ce n'est donc pas là aussi une *création* de la nature inerte en matière de vie.

La *créativité* ce n'est pas tant en effet l'émergence de *formes* nouvelles comme ce dit moteur moléculaire, que d'associer ces *formes* d'une façon *cohérente* – alors qu'elles sont physiquement *indiscernables* – afin que les *actions* qui en résultent assurent la pérennité d'une structure physico-chimique comme la bactérie E. coli alors face aux sollicitations

toujours dégradante de son environnement. La « *créativité* de la bactérie » c'est que ce moteur moléculaire soit associé d'une façon *cohérente* à un capteur sensible au glucose – une protéine membranaire MCP, afin que cette bactérie puisse capturer efficacement des molécules de glucose. C'est la *conscience* qui comme nous l'avons montré doit être nécessairement associée à la bactérie, qui va choisir pour son « plaisir d'être en vie » les liaisons physico-chmiques adéquates entre le capteur membranaire MCP et le moteur moléculaire associé à la flagelle.

Quant au transfert hypothétique de notre esprit vers un super-ordinateur en recopiant simplement les différents états d'activation ou de non-activation des neurones du cerveau assimilé à une machine informatique c'est-à-dire en ignorant l'existence de la *conscience*, est tout aussi irréalisable. On ne ferait que construire un super-automate à l'instar de « l'Automaton » du neurobiologiste Wilder Penfield[15]. Un être humain privé de *conscience*, donc de toute *sensibilité*, à cause d'un dysfonctionnement majeur ou d'une opération chirurgicale, et ayant de ce fait totalement perdu la capacité de *créer*, de s'*adapter* à un environnement différent de celui qu'il avait connu lorsqu'il était en bonne santé.

Il n'est pas contestable que les formidables progrès techniques en matière de biologie et de neurosciences laissent à penser que la plupart des maladies vont disparaître et que toutes ou presque les différentes parties du corps pourront être réparées ou remplacées. Mais cette approche purement *computationnelle* de la vie que proposent les tenants du *transhumanisme* qui conduirait à une augmentation infinie de nos capacités mentales et à d'élimination de tous les mécanismes de la *douleur*, n'est donc pas scientifiquement fondée.

Conclusion

La construction d'un robot *autonome*, artificiellement *vivant*, constitué d'éléments techniques résultant d'interactions physico-chimiques, est formellement irréalisable. On peut seulement construire des robots *automates* dotés d'outils dont les performances peuvent largement dépasser celles dont sont capables les êtres vivants en matière de force développée, de capacité de calcul et de mémorisation. Mais ces robots demeurent incapables de s'adapter aux contraintes infiniment variables de leur environnement, de créer de nouveaux outils comme savent le faire tous les *êtres vivants* que ces robots autonomes devraient imiter.

Partageant le même critère d'existence que les robots *autonomes* devraient posséder, à savoir assurer à tout prix la pérennité de leur structure, les êtres vivants ne devraient donc pas exister !
Il s'avère que les structures *vivantes* accompagnées de leurs capacités cognitives, doivent leur existence à la *conscience* munie de *qualités sensibles* comme le *plaisir* ou la *douleur*, dont ils sont empiriquement dotés. La *conscience* ayant la capacité essentielle de sélectionner les objets du monde – tous physiquement *indiscernables* – avec lesquels ces structures interagissent afin d'assurer à tout prix leur pérennité. La *conscience* s'avère être la clé de l'existence sur Terre des êtres vivants et de leurs cognitions.

La 'théorie computationnelle de l'esprit' où l'esprit humain fonctionnerait comme une machine informatique à l'instar de la thèse transhumaniste où les fonctionnalité des êtres vivants et de leurs cognitions doivent toutes pouvoir se réduire à des 'algorithmes' ce qui entraînerait le possible transfert de notre 'esprit' vers un indestructible super-ordinateur, est de ce fait totalement infondée. Il en résulte que les machines informatiques ne possèdent aucun pouvoir de *création*. Toute *création* implique une *conscience*.

Puisque les *actions* que nous effectuons sur les objets du monde ne dérivent que des *choix* effectués par nos *conscience* munies de *qualités sensibles*, c'est que ces *actions* qui émanent de

[15] Wilder Penfield – *The Electrode, the Brain and the Mind* - Z. Neurol. 201, 297-309 (1972) Springer Verlag 1972

nos pensées sont toutes fondamentalement « irrationnelles ». En ce sens que les *actions* qui sont sélectionnées parmi toutes celles possibles issues d'interactions physiques entre neurones, ne résultent en aucun cas d'*opérations logiques* fondées sur les lois de la physique. Ces *actions* pérennisantes sont en effet essentiellement fondées sur le *plaisir* que perçoit la *conscience* à leur déploiement, et non pas résultant de calculs spontanés se développant dans les réseaux de neurones de notre cerveau en tant qu'ordinateur.

La *conscience* munie de ses *qualités sensibles* qui choisit – et non pas construit – des solutions techniques particulières élaborées par un cerveau plus ou moins performant, ne posséderait ainsi aucune connaissance a priori sur le monde. À ce titre, la nature du *ressenti* apporté par la *conscience* pourrait être qualitativement semblable pour tout ce qui vit. La *conscience d'exister*, le « vivre dans la certitude du monde » selon Husserl, serait alors, à des degrés divers en fonction de la richesse de la description technique du monde, naturellement partagée par tous les êtres vivants.

- Annexe -

'théorème d'indiscernabilité'

Pour effectuer la mesure d'un observable **O** (température, poids, longueur,...) sur un objet **A**, il faut que cet objet interagisse avec un dispositif technique donné $\mathbf{A_p}$ qui a la particularité de se mettre dans un état final spécifique unique **E** lorsque l'interaction est complète.

Considérons un dispositif de mesure thermométrique constitué des éléments suivants : un *capteur* (un thermomètre), un *afficheur* (un écran sur lequel on peut lire les résultas de la mesure), des *liaisons physiques* entre le capteur et l'afficheur. Le monde sur lequel portent les mesures thermométriques est supposé constitué des seuls deux objets **A** et **B** dont le premier est, par exemple, de la lave en fusion, il est *chaud*, et le second un morceau de glace, il est *froid*. Les mesures sont ainsi relatives au seul observable **P**, la *température*.

Généralement, si **P** est le nombre d'observables, il y a $\mathbf{N} = 2^{\mathbf{P}}$ propriétés possibles. En l'occurrence, avec le seul observable *température*, soit **P** = 1, il en résulte qu'il existe $\mathbf{N} = 2^{\mathbf{P}} = 2^1 = 2$ propriétés différentes. À savoir, les propriétés 'chaud' et 'froid' qui caractérisent respectivement les objets **A** et **B**.

À un état donné (*chaud* ou *froid*) de l'entité **A** ou **B** qui fait l'objet de la mesure correspond un *état unique* du capteur thermométrique. Ainsi, avec un thermomètre en tant qu'élément sensible $\mathbf{A_p}$ à l'observable *température*, cet état est représenté par la longueur de la colonne de mercure qui est fonction de la température de l'objet sur lequel porte la mesure.

Deux cellules photoélectriques $\mathbf{C_A}$ et $\mathbf{C_B}$ – seulement sensibles à la forme spécifique du ménisque de mercure dans le tube capillaire – sont positionnées en deux points du capillaire qui correspondent aux deux positions possibles atteintes par le ménisque suivant que c'est l'objets **A** ou **B** qui fait l'objet de la mesure .

Lorsque l'objet **A** se trouve devant le bulbe du thermomètre et que la mesure est complète (la colonne de mercure est stabilisée, les positions transitoires sont ignorées), seule la cellule photoélectrique $\mathbf{C_A}$ est activée, soit un signal de sortie $\mathbf{S_A} = 1$, avec $\mathbf{S_B} = 0$. Pour l'objet **B**, c'est seulement la cellule $\mathbf{C_B}$ qui est activée, soit un signal de sortie $\mathbf{S_B} = 1$, avec $\mathbf{S_A} = \mathbf{0}$.

La question qui se pose est la suivante : quelles sont les liaisons physiques qui peuvent être établies entre les deux sorties **S_A** et **S_B** du capteur et l'entrée de l'afficheur (écran de lecture) afin de prendre en compte la totalité des informations qui sont issues du capteur.

Logiquement, il y a 3 et seulement 3 combinaisons physiques possibles entre les deux sorties **S_A** et **S_B**, soit : **S_A**, **S_B**, **{S_A *ou* S_B}**. La combinaison {**S_A** ***et*** **S_B**}, au moyen de l'opérateur « ***et*** », étant logiquement toujours égale à 0 puisque les cellules photoélectriques **C_A** et **C_B** ne peuvent pas être simultanément activées lors de la présentation des objets **A** ou **B** devant le thermomètre est ignorée.

> En toute généralité, il y a **M** = 2^N-1 combinaisons possibles *opérantes* établies à partir des **N** propriétés possibles (ce sont les **N** lignes qui résultent des différentes combinaisons possibles de 0 et de 1 d'un tableau possédant **N** colonnes).

En l'occurrence, le fait qu'il existe **N** = 2 propriétés *chaud* et *froid* correspondant respectivement aux objets **A** et **B**, entraîne qu'il y a bien **M** = 2^N**-1** = 2^2-1 = 3 combinaisons possibles, soit : **S_A**, **S_B**, **{S_A *ou* S_B}**

En considération du contexte expérimental, un opérateur établit alors d'une façon exhaustive les 3 liaisons possibles suivantes entre le capteur et l'afficheur :

- une liaison **L_1**, attachée à la sortie **S_A**.
- une liaison **L_2**, attachée à la sortie **S_B**.
- une liaison **L_3**, attachée à la sortie composite **{S_A *ou* S_B}** – aussi licite que les liaisons **L_1** et **L_2**

Les 3 liaisons **L_1**, **L_2**, **L_3** entre le capteur et l'afficheur étant établies, on réalise alors les tâches suivantes :

(1) l'objet **A** est placé devant le bulbe du capteur thermométrique : seule la cellule photoélectrique **C_A** est alors activée, d'où **S_A** = 1 et par là même **{S_A *ou* S_B}** = 1**.** Les liaisons **L_1** et **L_3** sont donc simultanément activées, ce qu'indique l'afficheur qui mémorise aussi ce résultat.

(2) l'objet **B** est placé devant le bulbe du capteur thermométrique : seule la cellule photoélectrique **C_B** est activée, d'où **S_B** = 1 et par là même **{S_A *ou* S_B}** = 1. Les liaisons **L_2** et **L_3** sont simultanément activées, ce qu'indique l'afficheur qui mémorise aussi ce résultat.

Finalement, l'écran de l'afficheur se présente ainsi à l'opérateur :

objet **A**	**L_1**	**L_3**
objet **B**	**L_2**	**L_3**

N'ayant aucune connaissance a priori des objets que nous avons nommés initialement **A** et **B**, l'opérateur doit en conclure à la lecture de la deuxième colonne du tableau, que **A** = **L_1** et **B** = **L_2** et qu'à ce titre ces deux objets sont *différents*.

Mais la lecture de la troisième colonne du même tableau lui indique aussi que **A** = **L_3** et **B** = **L_3** ce qui signifie que ces deux objets sont aussi *identiques*.

Étant à la fois *différents* et *identiques*, les objets A et B perçus par un *capteur* sont physiquement *indiscernables* du point de vue de l'*actionneur* auquel ce capteur est associé.