Conscience : à l'origine des êtres vivants. La conscience est non-physique.

Michel Troublé – Ph.D – directeur de recherche robotique - intelligence artificielle.

La pensée dominante actuelle concernant la nature des êtres *vivants* et de la *cognition* dont ils sont munis, est que leurs fonctionnalités 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 serions donc 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 caractérise. Cette approche purement *computationnelle* de la vie n'est cependant pas scientifiquement fondée.

C'est l'analyse fonctionnelle de robots qui seraient pleinement *autonomes*, c'est-à-dire artificiellement *vivants* car capables d'assurer en toutes circonstances la pérennité de leur structure, qui va apporter les éléments formels qui justifieront cette assertion. Un robot est un système physico-chimique plus ou moins complexe associant des éléments mécaniques, électroniques et informatiques. Il est doté des éléments suivants : des *capteurs* qui mettent en relation le robot avec le milieu extérieur, des *actionneurs* qui assurent sa mobilité et pour finir un *contrôleur*, une structure de calcul qui établit des liens informationnels appropriés entre ces capteurs et ces actionneurs.

La robotique autonome

L'*autonomie* des robots est actuellement une question majeure en matière d'automatisme. Il ne s'agit pas moins en effet que de créer des structures physico-chimiques qui, à terme, seraient *artificiellement vivantes* et à ce titre dotées de capacités semblables à celles des êtres vivants en matière des décisions qu'ils doivent prendre pour assurer leur pérennité en toutes circonstances, ce qui les caractériserait.

Cette *autonomie* est notamment une demande récurrente en matière d'exploration spatiale compte tenu du fait que le temps de transmission d'une communication radio entre la Terre et un véhicule qui se déplace par exemple sur le sol de Mars, peut s'élever jusqu'à 20 minutes. Ce qui interdit par là même toute commande instantanée avec le véhicule qui doit alors se tirer seul d'affaire. Il serait donc souhaitable que ce robot d'exploration soit ainsi totalement *autonome* comme pourrait l'être un être humain en pareille circonstance.

En fonction de la nature des différents objets qu'un robot *autonome* percevrait avec ses capteurs, les actions qu'il effectuerait devraient donc être telles qu'elles devraient assurer la pérennité de sa structure, en évitant en particulier d'endommager les délicats et très coûteux appareils de mesure dont il est équipé.

Si par exemple le robot est de par sa construction sensible à toute élévation de température de sa structure, il faudrait que quelles que soient les températures T mesurées par le capteur thermométrique dont il est équipé, lesquelles températures correspondraient aux objets chauds rencontrés lors de ses déplacements, son module de contrôle ou cerveau artificiel soit capable de *créer* deux *catégories* distinctes *d'actions* :

> 'Fuir' ou 'continuer à se déplacer dans la même direction' afin que la température de sa structure reste par exemple toujours inférieure à environ 30°C. Ce qui correspondrait alors à la création des deux *catégories d'actions cohérentes* suivantes : {'fuir' si T>30°} mais {'avancer' si T<30°}.

Il faut noter que cette analyse fonctionnelle de la capacité d'un robot qui serait *autonome* s'applique pareillement à toute structure physico-chimique comme celle d'un technicien qui observant les mouvements de la colonne de mercure d'un thermomètre dont il est équipé déciderait de s'éloigner prudemment de tous les objets très chauds qu'il pourrait rencontrer.

L'indiscernabilité des objets du monde

Cette capacité essentielle qu'aurait donc le robot *autonome* de 'fuir' un objet chaud si la température de sa structure devenait supérieure à 30°C impliquerait alors logiquement que les différentes températures mesurées par le capteur thermique du robot au cours de ses divers déplacements seraient évidemment *discernables* les unes des autres par son contrôleur. Que le contrôleur sache ainsi, par

exemple, faire la différence entre 25°C qui est une température < à 30°C, et 45°C qui est au contraire une température supérieure à 30°C.

Dans le cas contraire, les diverses manœuvres du robot commandées par le contrôleur ne pourraient alors se faire qu'au *hasard* ce qui serait assurément contraire à la capacité attendue d'éviter toute approche destructrice avec les objets chauds de son environnement.

Que les différentes températures affichées par le capteur thermométrique du robot soient physiquement *discernables* par son contrôleur afin que le robot puisse se détourner systématiquement de tous les objets très chauds de son environnement, semble aller naturellement de soi.

Nous effectuons en effet naturellement une opération similaire lorsque le matin nous observons la température affichée par le thermomètre qui est accroché à l'extérieure de notre maison afin de savoir si nous devons conserver ou non un vêtement chaud pour sortir. Mais cette tâche qui est pour nous très naturelle est en fait paradoxalement *irréalisable* pour toute structure physique inanimée comme le robot.

On peut en effet démontrer mathématiquement à partir de la théorie formelle de la « reconnaissance des formes »[1] qui porte sur l'identification des *formes* d'objets à partir des propriétés qui les caractérisent, que les différentes températures mesurées par le capteur thermométrique du robot, soit les différentes positions du ménisque de mercure dans le tube capillaire, sont en l'occurrence *indiscernables* par son contrôleur qui commande le système de locomotion. C'est ce que, pour faire court, nous appellerons le 'théorème d'indiscernabilité'. Sans être une démonstration rigoureuse de ce théorème, le scénario suivant permet de beaucoup s'en approcher :

Supposons qu'un jardinier dispose de deux objets, une grosse pierre qui est *lourde* et un morceau de bois qui est *léger*.

- Dans un premier temps, le jardinier souhaite enfoncer un clou qui en dépassant d'une planche risque de le blesser. Pour ce faire, il utilise la pierre qui est *lourde* et non pas le morceau de bois qui est manifestement trop *léger* – c'est son expérience passée qui lui dicte cette conduite. La pierre est ainsi un objet *différent* du morceau de bois en ce qui concerne l'*action* d'enfoncer un clou.
- Dans un second temps, le même jardinier décide de repérer plusieurs endroits dans son jardin pour y disposer des plantes. Il utilise alors indifféremment la pierre ou le morceau de bois pour repérer des emplacements distincts dans le jardin. La pierre est donc maintenant un objet *semblable* au morceau de bois puisqu'ils peuvent avoir l'un et l'autre la même fonction de repérage.

Ce qu'il faut retenir de cette historiette, c'est que les *actions* physiques qui peuvent être associées aux deux objets 'pierre' et 'morceau de bois', dépendent essentiellement du « bon vouloir du jardinier ». En l'absence du jardinier qui est une structure physico-chimique déjà *vivante*, les deux objets sont en effet fondamentalement *indiscernables* en ce qui concerne les différentes *actions* qu'ils peuvent entraîner. C'est ce qu'implique le 'théorème d'indiscernabilité'.

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.

Ce 'théorème d'indiscernabilité' remet de ce fait totalement en question la représentation que nous nous faisons du monde ainsi que la façon dont nous agissons sur lui pour assurer notre pérennité, autrement dit pour être *vivant*.

Pour illustrer les conséquences considérables qu'implique ce 'théorème d'indiscernabilité' en ce qui concerne les *actions* que nous effectuons sur les objets du monde pour assurer la pérennité de notre existence, prenons l'exemple d'un simple thermostat dont le capteur est un thermomètre où le ménisque de mercure est, pour simplifier, soit en position 'haute' (on dit alors qu'il fait 'chaud' dans la pièce), soit en position 'basse' (on dit alors qu'il fait 'froid' dans la pièce). La position du ménisque étant par exemple observée par un dispositif électrique, un relais par exemple, qui commande la mise en marche ou l'arrêt d'un dispositif de chauffage.

Ce que nous dit en l'occurrence le 'théorème d'indiscernabilité', c'est qu'en l'absence d'un technicien, une structure physico-chimique *vivante*, qui aurait *préparé* le thermostat en connectant entre eux les différents composants d'une façon appropriée, les positions 'hautes' et 'basses' du ménisque de mercure

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

seraient physiquement *indiscernables* par son actionneur, ici un relais électrique. Il en résulterait donc que la fermeture et l'ouverture du relais qui commande le chauffage de la pièce ne pourraient donc se faire qu'au *hasard*. Et non pas comme cela est souhaité par celui qui achète le thermostat pour son usage personnel : « il fait 'froid', le système de chauffage doit être mis en marche », ou « il fait 'chaud', le système de chauffage doit être arrêté ».

Cet exemple du thermostat est à l'image du robot d'exploration du sol martien qui en absence d'un technicien, ne pourrait que se diriger ou se détourner au *hasard* de tous les objets qu'il rencontrerait. Réactions qui le conduiraient inéluctablement à sa destruction.

En l'absence d'un technicien, une structure physico-chimique naturellement *vivante*, qui contrôlerait le système de locomotion du robot d'exploration en temps réel, ce qui est physiquement impossible à cause du temps de transmission des signaux radio entre la Terre et la planète Mars, ce robot ne peut donc être qu'un automate plus ou moins performant qui ne serait 'autonome' que dans un domaine d'exploration dont les propriétés auraient été préalablement spécifiées par son constructeur. En dehors de ce domaine particulier qui aurait été strictement défini, le robot ne pourrait donc que réagir *aléatoirement* aux perceptions qu'il aurait d'un nouveau domaine. En conséquence de quoi, un robot d'exploration strictement *autonome* est donc physiquement irréalisable.

La sélection naturelle darwinienne

Il semble qu'il y ait néanmoins une façon pragmatique de rendre *autonome* ce robot d'exploration, c'est de s'inspirer du mécanisme mis en œuvre dans la « sélection naturelle darwinienne » qui selon les chercheurs expliquerait justement l'émergence des fonctions d'animation dont sont munies certaines structures physico-chimiques alors qualifiées de *vivantes* et de ce fait *autonomes*.

Ce mécanisme de « sélection naturelle darwinienne » aurait en effet semble-t-il cette vertu fondamentale en matière d'*autonomie* de ne pas nécessiter la formation de *catégories d'actions cohérentes* que nous avons vu être irréalisable pour un robot solitaire en raison du 'théorème d'indiscernabilité'.

On appelle ainsi « robotique évolutionniste » ce mécanisme qui ferait donc intervenir non pas un seul robot d'exploration mais plusieurs de ceux-ci. À cette fin, on devrait commencer par envoyer sur le sol martien un certain nombre de robots, et au retour de leur première mission, ceux qui n'auraient pas été endommagés pourraient alors se 'marier' en combinants leurs 'gènes' artificiels. D'où l'émergence d'une nouvelle flotille réduite de robots 'fils' ayant hérité de l'expérience acquise fortuitement par leurs 'parents'. Pratiquement, ces 'gènes' artificiels pourraient être constitués par les 'poids synaptiques' des réseaux de neurones artificiels qui composeraient les contrôleurs des robots de la flotille.

Au bout de plusieurs générations de robots ayant exploré le sol de la planète, il devrait donc théoriquement exister selon la théorie de la « sélection naturelle » un ou plusieurs robots qui auraient ainsi appris sans l'aide d'aucun technicien à éviter tous les objets chauds du sol martien qui auraient pu les détruire.*

Un ou plusieurs robots d'exploration seraient devenus ainsi spontanément *autonomes*, ce qui irait assurément à l'encontre de notre précédent propos où nous affirmions que l'*autonomie* d'un robot d'exploration solitaire était physiquement irréalisable car ne pouvant que réagir aléatoirement à la perception des objets de son environnement.

Des expériences de « robotique évolutionniste » menées dans plusieurs laboratoires montrent que ce mécanisme de « sélection naturelle » est tout à fait opératif. Mais l'analyse rationnelle du mécanisme de *duplication fonctionnelle* qui permet la naissance des robots 'fils' ayant hérité de leurs 'parents' grâce aux mixages des 'gènes' des robots qui seraient revenus indemnes de leurs explorations, montre que ce mécanisme est en réalité physiquement impossible en raison du 'théorème d'indiscernabilité'. Ce théorème qui, comme nous venons de le voir, rendait physiquement impossible la formation de *catégories d'actions cohérentes* qui devaient fonder l'*autonomie* d'un robot d'exploration solitaire.

C'est ce même mécanisme de duplication *fonctionnelle*, et non pas par *empreinte*, qui permet qu'on puisse visionner les informations gravées sur un DVD qu'on vient d'acheter. À l'*empreinte* proprement dite du disque, un simple matriçage, doit être en effet nécessairement associée la description précise du mécanisme de lecture des informations gravées. C'est ce mécanisme de lecture cohérente des informations qui est en fait physiquement impossible en raison du 'théorème d'indiscernabilité'.

Si ces expérimentations en laboratoire de « robotique évolutionniste » sont néanmoins tout à fait probantes, c'est que des techniciens sont intervenus d'une façon nécessairement très orientée dans l'implémentation des algorithmes relatifs au processus du 'mariage' des robots définis par leurs gènes.

Au même titre qu'un technicien devait nécessairement intervenir dans la mise en place appropriée des connexions électriques entre le capteur thermométrique et le relais d'un thermostat.

En absence d'un technicien, le mécanisme de la « sélection naturelle darwinienne » appliqué à la robotique n'a donc aucunement la capacité attendue de rendre *autonome* un ou plusieurs robots de la flotille d'exploration du sol martien.

L'analyse fonctionnelle du processus de la « sélection naturelle darwinienne » que nous savons être constamment à l'œuvre dans la nature en créant une multitude de nouvelles espèces, montre que celui-ci ne fait en réalité que sélectionner parmi les différentes *formes* possibles de systèmes matériels déjà *autonomes*, celles qui sont les mieux adaptées pour survivre aux contraintes environnementales. Ce processus de sélection des *formes* qui ne fait appel qu'à des mécanismes de reproduction par *empreintes* et non pas *fonctionnelles*, étant quant à lui tout à fait licite au regard du 'théorème d'indiscernabilité'.

La conscience

En définitive, la réalisation d'un robot *autonome*, artificiellement *vivant*, qui est le but que nous nous étions fixé, est donc physiquement impossible en vertu du 'théorème d'indiscernabilité'. Au mieux, on ne peut donc que construire un automate qui ne fera qu'obéir aux ordres limités que son créateur aura implantés dans sa mémoire.

Qu'un robot *autonome*, artificiellement *vivant*, soit en définitive physiquement irréalisable est sans doute quelque peu décevant pour les tenants d'une intelligence artificielle *forte* où les diverses fonctionnalités du vivant, et en particulier la conscience, doivent pouvoir toutes se réduire à des algorithmes. Mais ce qui est finalement tout à fait paradoxal au sens fort du terme, c'est qu'il existe pourtant sur Terre de nombreuses structures physico-chimiques qui sont naturellement *autonomes*, ce sont les êtres vivants !

Considérons ainsi un technicien, une structure physico-chimique *vivante*, qui observant les mouvements de la colonne de mercure d'un thermomètre qu'il tient entre ses mains décide de s'éloigner prudemment de tous les objets très chauds qui pourrait le brûler et de ce fait abréger sa vie.

La bonne réussite de cette opération d'évitement systématique des objets chauds implique donc que ce technicien doit être spécifiquement muni d'un « opérateur » dont la vertu essentielle est de pouvoir *différencier* les différents positions du ménisque de mercure du thermomètre qui sont entre elles physiquement *indiscernables* en vertu du 'théorème d'indiscernabilité'.

Expérimentalement, on constate en dernière analyse que c'est le 'plaisir' que ressent en l'occurrence le technicien à la réussite de l'opération d'évitement des objets chauds qui pourraient le détruire, qui serait à l'origine de cette singulière capacité que possèdent les vivants à *différencier* les objets du monde qui sont fondamentalement *indiscernables*. C'est également pour le 'plaisir' du technicien qui s'est donné beaucoup de mal à construire un robot d'exploration, qu'il est alors impératif que ce robot se détourne ultérieurement de tout type d'objet chaud qui provoquerait inéluctablement sa destruction.

Le 'plaisir' ressenti par le technicien résulte donc essentiellement des *actions* qui conduisent à la préservation de sa structure où des objets qu'il a construit. Ce 'plaisir' a pour origine son expérience passée d'être vivant.

C'est par exemple en visitant une fonderie dans laquelle le technicien commençait à avoir très chaud, ce qui l'incommodait sérieusement, que la solution technique particulière {se détourner de l'objet incandescent observé} créé par hasard par son cerveau-ordinateur à partir des données capteurs, à été spontanément activée.

L'éloignement rapide de l'objet chaud avait alors eu pour conséquences suivantes : d'une part de diminuer rapidement son malaise thermique, en faisant baisser sa température corporelle, et d'autre part l'apparition concomitante d'un 'plaisir' qui avait dans l'instant même marqué, étiqueté, l'action {se détourner de l'objet incandescent observé}. Cette action ainsi marquée par le 'plaisir', avait alors été mémorisée à jamais dans son cerveau-ordinateur.

D'où la réaction actuelle du technicien qui pour son 'plaisir' cherche à faire éviter au robot d'entrer en collision avec des objets chauds qui se trouvent sur sa trajectoire à l'instar de son expérience passée acquis lors de la visite d'une fonderie.

Empiriquement, la *conscience* munie des *qualités sensibles* extraordinairement prégnantes que sont le 'plaisir' et la 'douleur' serait l'opérateur recherché dont doivent être dotés tous les êtres vivants, car il possède le pouvoir unique de différencier pour 'son plaisir d'être' les objets du monde qui, fondamentalement, sont physiquement *indiscernables*.

Ce mode d'action de la conscience qui se réduirait à choisir des solutions techniques particulières parmi celles qui sont créées en aveugle par notre cerveau en tant qu'ordinateur, est en parfait accord avec les résultats paradoxaux des expériences du neurobiologiste Benjamin Libet : « la conscience oppose son veto aux solutions préalablement élaborées quelques 500 ms plus tôt par le cerveau-ordinateur. »[2]

Propriété essentielle, la *conscience* dont est muni naturellement le technicien, et plus généralement tous les êtres vivants est *irréductible* à toute interaction physicochimique neuronale du cerveau eu égard le 'théorème d'indiscernabilité'. Autrement dit, elle est de nature *non-physique* car étant en effet non-mesurable.

Si la conscience était *mesurable* elle pourrait être définie par des attributs duals complémentaires tels que chaud/froid, blanc/noir,... et le calcul des prédicats (ou descripteurs) appliqué au nouveau système {capteur + conscience} montrerait alors 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* serait donc bien de nature *non-physique*.

Cette hypothèse de la non-matérialité de la *conscience* n'est pas physiquement irrecevable comme on pourrait le penser de prime abord en s'appuyant sur le postulat communément 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*, et à ce titre, il résulte d'un processus de *catégorisations cohérentes* qui, comme nous l'avons vu, implique l'existence de la conscience, un « opérateur premier », conscience qui est elle nécessairement *irréductible* à toute interaction physique en raison du 'théorème d'indiscernabilité'. À ce titre, cette hypothèse de la *non-matérialité* de la conscience serait donc licite.

Les différents *états de conscience* qui nous ouvrent à la perception colorée et sensible d'un univers autrement clos sur lui-même, seraient de ce fait fondamentalement irréductibles aux états mentaux de notre cerveau contrairement à la théorie de « l'identité esprit-cerveau (IT) ».

Ces *états de conscience* seraient donc non pas le fruit au sens physico-chimique du terme d'états neuronaux spécifiques, mais seulement induits par ces états neuronaux. La *conscience* n'est pas comme la mémoire d'un ordinateur qui résulte de l'association d'une multitude de composants électroniques (semi-conducteurs) à base de silicium.

À ce titre, la *conscience* ne pourrait pas se réduire à un algorithme comme l'affirme le neuroscientifique Stanislas Dehaene (Collège de France) dans son livre 'Code de la conscience' où il est question d'un code qui décrit l'emplacement des différents neurones qui sont activés lorsque le sujet dit être *conscient*.

Dans l'expérimentation qui a été menée il ne s'agissait en effet que de rechercher s'il existait des corrélations entre les déclarations faites par un sujet à propos par exemple d'une 'douleur' qu'il ressentait, avec l'activité neuronale de certaines régions de son cerveau qui était visionnée grâce aux techniques de l'imagerie cérébrale ou d'électrodes implantées dans son cortex. Mais ces mesures ne donnaient aucunement accès à la *conscience* proprement dite, à sa capacité opérative qui fonde le *vivant*. Ces recherches n'infirment donc nullement la thèse que nous soutenons en matière d'irréductibilité de la *conscience* à toute interaction physique.

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