Vie et intelligence artificielles, et discussion autour de la possibilité d'une conscience artificielle

Granier Arno
encadré par S.Cormier

6 mai 2018

Table des matières

Introduction 3

I. La vie. 4

1. **Vie organique, vie artificielle et complexité.** 4
2. **La vie comme le comportement complexe émergent de l'autoréplication et de l'évolution d'une population d'agents simples.** 6
3. **La vie comme l'autopoïèse.** 12

II. L'Intelligence. 15

III. Possibilités d'une conscience dans une machine. 20

1. **Problemes corps/esprit et impacts sur les possibilites d'une conscience dans une machine.** 20
2. **D'autres difficultes dans la definition de l'esprit et de la conscience.** 23
3. **Approches du problème.** 25

Références **28**

Introduction

Les problématiques de l'intelligence, de la vie et de la conscience dans les machines animent depuis quelques décennies la communauté scientifique et les médias, et les visions scientifiques de la question tout comme les articles alarmistes et les œuvres de fictions traitant de ces sujets abondent. Dans ce contexte et pour pouvoir discuter de ces sujets sérieusement, il me semble essentiel de savoir précisément ce que l'on entend par les termes *vie artificielle*, *intelligence artificielle* et *conscience artificielle*. *Est-il possible de construire une machine vivante ? Intelligente ? Consciente ? Si oui, comment ?*. Ces questions sont simples à poser, mais leur réponses sont loin d'être évidentes et dépendent bien évidemment de l'interprétation (car il s'agit bien là d'interprétation) que l'on a des termes *vie, intelligence, conscience*. Ainsi, en 1957 H.Simon énonce que, "pour résumer simplement, je dois dire qu'il existe aujourd'hui dans le monde des machines qui pensent, qui apprennent et qui créent" tandis que, 50 ans plus tard, aucune machine n'a encore passé le test de Turing, que le mathématicien a proposé afin de déterminer si une machine était intelligente ou non. Cette quête d'une définition précise des ces concepts nous amènera donc à nous poser des questions millénaires telles que *Qu'est-ce que la vie ? Et d'où vient-elle ?* , *Qu'est-ce que l'intelligence ?* ou encore *Quelle est la nature de l'esprit ?*, et à analyser le rapport de ces questions avec un concept contemporain qui est celui de l'*artificialisation*. Toutes ces questions en entrainant inévitablement d'autres dès que l'on commence à les examiner de plus près.

Pour aborder ces questions nous allons tout d'abord parler de vie artificielle, nous présenterons deux définitions des critères qui font la vie et quelques résultats surprenants de cette jeune discipline. Nous tenterons ensuite de définir l'intelligence à travers l'usage qui en est fait dans le domaine de l'intelligence artificielle, notamment en présentant les travaux de formalisation de M.Hutter sur l'intelligence généralisée. Nous finirons par nous poser la question de la nature de l'esprit et de la pensée et nous étudierons l'impact de la réponse à cette question sur les possibilités d'une machine consciente.

La philosophie de l'esprit est une discipline vivace où les consensus sont rares (et pour cause une question en philosophie de l'esprit qui fait consensus trop longtemps a souvent tendance à changer de statut, de "question en philosophie de l'esprit" à "connaissance en sciences de l'esprit - sciences cognitives, neurosciences, psychologie, etc.") et les débats animés courants. Cet essai se veut être une introduction aux approches de ces questions qui me paraissent les plus prometteuses et ne doit pas être pris pour ce qu'il n'est pas, à savoir un rapport de l'intégralité des approches qui existent autour de ces questions.

Première partie .
La vie.

Dans cette partie, nous allons tenter d'apporter des éléments de réflexion sur le concept de vie, en orientant nos propos vers une analyse des fondements de l'étude scientifique de la vie artificielle.

1. Vie organique, vie artificielle et complexité.

Pour pouvoir juger de la réussite ou même de la possibilité de simuler ou d'instancier la vie dans un médium artificiel, il faut d'abord *savoir* ce qu'est la vie, c'est-à-dire en avoir posé une définition claire. Dans cette quête d'une définition de la vie, nous pouvons au moins avoir un point de repère : les entités que nous savons vivantes, c'est-à-dire, d'après les biologistes, les animaux (*Animalia*), les plantes (*Plantae*), les champignons (*Fungi*) et plusieurs catégories d'espèces unicellulaires (*Protozoaire*, *Bactérie*, *Archaea*), qui devront nécessairement être englobés par une définition de la vie se voulant en accord avec la biologie moderne. Mais, dans l'Histoire, relativement peu de biologistes se sont sérieusement intéressés à la définition même de la vie, préférant occulter ce problème pour se concentrer sur l'étude des mécanismes (ou des métabolismes, devrais-je dire) à l'œuvre dans des entités posées comme vivantes, avec en fond, l'espoir que le recensement méthodique de ces mécanismes mène à une compréhension de la vie. Actuellement, il n'y a donc pas de consensus quant à une définition précise de la vie proposée par les biologistes. Malgré tout, il y a deux processus à inclure dans une définition de la vie sur lesquels une majorité tombent d'accord : l'entité vivante extrait de l'énergie de son environnement, et évolue et s'adapte à son environnement à l'aide d'un mécanisme reproductif.

On peut remarquer qu'aucun de ces deux points ne force la vie à n'être qu'organique, et ne décrit, finalement, que des comportements complexes d'une entité. Un changement de point de vue sur la vie, historiquement initié par John Von Neumann et la cybernétique de Ashby, s'impose alors, de la hiérarchie de l'ordre biologique du vivant vers une hiérarchie en termes de complexité d'un automate peu importe son medium d'incarnation. Pour être plus précis, ce que propose la cybernétique c'est une nouvelle conceptualisation des machines par la théorie de l'information, et une nouvelle compréhension d'un organisme vivant comme une réalisation complexe de ce nouveau concept de machine. D'après cette définition, on a donc une différence de complexité entre le vivant et le non-vivant, et non pas une différence ontologique.

La grande complexité du vivant biologique provient d'un historique de processus évolutifs, et c'est ainsi naturellement que Von Neumann émet l'hypothèse comme quoi cette complexité pourra être atteinte dans un automate artificiel lorsqu'il sera capable de s'auto-répliquer et d'évoluer (ces deux propriétés pouvant être résumées dans le terme

auto-reproduire) : ce qu'il appelle briser la barrière de la complexité (*complexity barrier*). Nous reviendrons sur ces notions dans la suite de cette partie. Bien sûr, l'évolution d'une telle entité dépendrait de son *milieu de vie*, et produire des entités vivantes artificielles ressemblants à la vie organique en passant par l'évolution implique nécessairement que ces entités évoluent dans le même environnement que nous, ou dans un environnement semblable, ou pour le moins que l'ensemble d'actions que le monde a sur l'entité et que l'entité peut avoir sur le monde soient semblables à ceux de notre environnement. Mais cette vision naturocentrée de la vie est-elle nécessaire ? Je ne pense pas que ce soit le cas, et je pense qu'il serait bénéfique de s'extraire de ce réflexe de juger du caractère vivant d'une entité en la comparant nécessairement aux organismes biologiques que nous connaissons.

Dans la quête de la vie dans un automate, une des grandes questions qui se pose est la suivante : Est-ce qu'un tel type d'automate modèlise et simule simplement la vie ? Ou alors, est-il vivant, c'est-à-dire instancie-t-il la vie ? C'est une question centrale qui n'est pas sans rappeller la problématique parallèle des p-zombies de Chalmers au sujet de la conscience, que nous abordons dans la partie III. Dans l'approche d'une définition que nous avons esquissée, la vie est définie comme les principes qu'elle respecte et les fonctions qu'elle accomplit, et non pas comme une propriété de la matière elle-même : c'est ce qui permet à la vie de s'extraire du médium organique. Et dès lors que notre définition de la vie s'exprime en termes de principes, de processus, de fonctions, alors une machine qui reproduirait ces principes, processus ou fonctions serait alors plus qu'une simple simulation de vie : elle serait vivante. [1] C'est de ce changement de point de vue sur la vie et de ces idées de Von Neumann et de la cybernétique qu'émergera la discipline scientifique que nous allons placer au cœur de cette partie : ALife, ou l'étude de la vie artificielle. Cette nouvelle discipline, comme très bien dit dans (Johnston, 2008),

> [...] est nécessairement positionnée dans l'espace qu'elle créée entre la biologie moléculaire (en tant que forme la plus contemporaine des sciences de la vie) et l'histoire des objets techniques.

On peut même aller plus loin et dire que cette discipline de la vie artificielle rend plus floue la barrière ente Nature et Technique, voire, en impose une redéfinition conjointe dans laquelle certaines de leurs différences sont remises en cause. Au cœur de cette redéfinition, un double changement de point de vue sur le vivant : les machines, de par leurs propriétés et leurs comportements, sont considérées comme vivantes et les organismes vivants biologiques sont considérés comme des machines dotées de certaines propriétés spécifiques.

1. Certains pourraient être en désaccord avec cette définition de la vie en tant qu'ensemble de principes et de fonctions : "Il y a plus que ça dans la vie !". C'est une position tout à fait défendable, qui relève du vitalisme, mais cette intuition que "Il y a plus que ça dans la vie !" peut aussi venir d'une confusion sur certains concepts. J'aimerais rappeler qu'il convient bien ici de séparer les concepts de vie de ceux d'intelligence, de pensée ou de conscience : notre "expérience" de la vie n'est pas nécessairement la définition de la vie ! Nous réduisons plus facilement la vie unicellulaire à une série de processus et de fonctions que la vie humaine, les deux sont pourtant, d'après les biologistes, des manifestations de la vie.

2. La vie comme le comportement complexe émergent de l'autoréplication et de l'évolution d'une population d'agents simples.

Dans les travaux de C. Langton, on retrouve l'idée comme quoi une entité vivante ne devrait pas être conçue comme une seule entité complexe, mais comme un ensemble d'entité-composantes simples en interaction. La complexité dans le comportement de cette entité vivante étant alors attribué aux interactions nombreuses, parallèles et non-linéaires entre ces composantes simples. On peut dire que la qualité ou la propriété de vie de l'entité constituée émerge des entités constituantes. Des mots de Langton (comme cité dans (Johnston, 2008)),

> Animé les machines ce n'est pas 'apporter' la vie à une machine ; il s'agit plutot d'organiser une population de machines de telle façon que leur dynamique interactive est 'en vie'

Dans le débat entre le vitalisme, qui considère que la vie provient d'une force non-physique qui sépare les entités vivantes des non-vivantes, et la thèse mécaniste du réel, qui réduit la vie à une série de lien cause-conséquence déterministes, cette troisième idée, représentative de ce qu'on appelle l'émergentisme, se positionne en tant que troisième voie (bien que souvent affiliée à des idées mécanistes).

La théorie de l'émergence, émergentisme ou encore théorie de l'auto-organisation est une théorie explicative des propriétés des systèmes complexes qui connait une grande popularité dans la littérature moderne, dont on peut attribuer l'origine à G.H. Lewes qui est le premier à établir la différence entre les effets émergents et les effets résultants dans (Lewes, 1877). Cette théorie énonce que les propriétés d'un système complexe *émergent* de l'organisation des entités fondamentales qui le composent. Le terme émerger signifie ici que l'organisation de ces entités fondamentales est la cause des propriétés ou des comportements du système complexe mais que les propriétés ou comportements émergents ne peuvent pas être réduits aux propriétés et comportements des entités fondamentales, c'est-à-dire qu'une connaissance complète des propriétés et comportements des entités fondamentales n'est pas suffisante pour prédire ou expliquer les propriétés ou comportements du système complexe : on dit que la théorie de l'émergence est non réductionniste. Cependant, cette théorie ne fait pas reposer cette émergence sur des forces ou des lois autres que les lois de la matière, et est en accord avec le monisme matérialiste, qui propose que le seul monde d'existence des entités soit le monde de la matière.

Pour préciser sa définition de la vie comme comportement complexe émergent de l'autoréplication et de l'évolution d'une population d'agents simples, Langton s'approprie les termes *génotype* et *phénotype* en les extrayant de leur utilisation purement biologique. Pour Langton, un génotype est un ensemble de règles locales qui vont définir le comportement d'un (ou d'un partie d'un) agent simple "constituant", tandis que le phénotype est la résultante de l'*exécution* du génotype et des interactions parallèles et non-linéaires qu'une telle exécution entraine. Cette définition a une conséquence méthodologique importante pour Langton : étant donné la nature non-linéaire des interactions

entre les agents simples, une connaissance parfaite du génotype n'est pas suffisante pour prédire le phénotype produit, et une étude du phénotype ne permet pas de déterminer le génotype qui l'a produit. Pour être plus précis, Langton pense que le seul algorithme capable de trouver des génotypes associés à certains phénotypes, c'est l'algorithme de la sélection naturelle, c'est-à-dire, par essai et erreur ("l'algorithme naturel" comme l'appelle Dennett). Cela pose certains problèmes : doit-on remettre en question le *dogme génétique* qui assure aujourd'hui pouvoir trouver des gènes codant pour certains aspects du phénotype ? Eh bien, il semble bien compliqué, lorsqu'on voit que l'on arrive à rendre une sourie fluorescente par modification génétique, de s'opposer totalement à cette prétention. Mais, pour moi, cela n'est pas contradictoire avec la définition de la vie de Langton pour deux raisons : la première est que, dans beaucoup de cas, la découverte d'un *gène du X* a en fait été faite par un algorithme de recherche assez similaire à un algorithme d'essais et erreurs, et deuxièmement, la non-linéarité des interactions entre les éléments codés par le génotype n'est, pour moi, pas un argument fort pour dire que la compréhension des propriétés du phénotype ne peuvent pas être déduite des propriétés du génotype. En effet, la compréhension du lien entre ces interactions et leur résultante n'est qu'une question de complexité, et la complexité est de mieux en mieux traitée par les sciences modernes, par exemple en mathématiques où la formalisation de la théorie de l'auto-organisation est un sujet populaire. Aussi, il me semble que cette impossibilité de déterminer un génotype à partir d'un phénotype et inversement ne doit pas être vue comme une impossibilité théorique mais bien comme une impossibilité technique.

Pour Langton, les caractéristiques principales du comportement vivant sont la réplication (ici, l'autoréplication) et l'évolution. Concentrons-nous d'abord sur le principe d'autoréplication, et revenons à l'origine du concept. Dans (Von Neumann & Burks, 1996), Von Neumann établit deux critères nécessaires que devra posséder un automate s'il veut pouvoir s'autorépliquer. Premièrement, cet automate devra posséder un *schéma de construction*, qui, lorsqu'il est activé, permettra de créer une copie de l'automate. Deuxièmement, ce *schéma de construction* devra être passé à la copie, pour que la copie soit capable elle-même de se répliquer. Dans les années 1940, Von Neumann, pour mettre en pratique ces deux principes, invente le concept d'automate cellulaire et, dans la foulée, en construit un capable de s'autorépliquer. Tout d'abord, qu'est-ce qu'un automate cellulaire ? Il s'agit d'une nouvelle manière d'envisager la computation de manière parallèle et faisant la part belle au principe d'émergence de comportement complexes à partir de règles simples. Un automate cellulaire est une grille de cellules qui sont chacune dans un état particulier, cet état évoluant en fonction de règles simples prédéfinies. Un automate cellulaire peut donc être définie par un ensemble d'états possible pour les cellules, et un ensemble de règles gouvernant les changements d'états des agents au court du temps. Ce qui est intéressant avec ces automates cellulaires c'est qu'ils peuvent exhiber des comportements complexes, qui ne peuvent alors être que la conséquence des interactions parallèles et non-linéaires dans le système, et c'est pourquoi les automates cellulaires sont souvent considérés comme un outil de travail aussi bien pratique que conceptuel dans le cadre d'une théorie de l'auto-organisation. De nombreux résultats mathématiques sur les automates cellulaires remarquables ont été démontrés, notamment que certains types d'automates cellulaires sont turing-complets, mais je ne m'étendrai pas sur les proprié-

tés mathématiques et logiques des automates cellulaires dans cet essai, pour ceux qui souhaitent en savoir plus, je recommande (Kari, 2005) comme introduction.
L'automate cellulaire capable de s'autorépliquer de Von Neumann, que l'on appelle aujourd'hui *réplicateur de Von Neumann* ou *constructeur universel de von Neumann*, est un automate à 29 états et chaque cellule calcule son état à l'étape de temps t par rapport à l'état de ses voisins à l'état de temps t-1. Ce réplicateur de Von Neumann peut être décomposé en trois parties distinctes : un *schéma de construction* contenant les instructions pour construire un individu, une *machine de lecture* capable de lire un schéma de construction et de construire l'individu correspondant et une *machine de copie* capable de copier un schéma de construction. Un réplicateur de Von Neumann procède alors à une autoréplication par ces 2 étapes : d'abord, la machine de lecture est utilisée pour lire et construire la machine spécifiée par le schéma de construction ; puis la machine de copie est utilisée pour créer une copie du schéma de construction, et cette copie est placée dans la machine nouvellement construite. Il peut d'ailleurs être intéressant de constater que cet automate autorépliquant utilise le même principe que les organismes biologiques pour se répliquer, à savoir l'utilisation d'un *schéma de construction* utilisé pour construire un nouvel individu, nouvel individu dans lequel est instancié une copie de ce *schéma de construction*. Dans l'automate, il s'agit d'une série de cellules de la grille dans une configuration d'état particulière, et dans la Nature, il s'agit de l'ADN. Il est d'autant plus remarquable de constater que le design de cet automate autorépliquant précède la découverte de l'ADN, ce qui en dit long sur l'intérêt d'aborder la vie naturelle avec un angle de vue algorithmique (machinistique, si je puis dire).

Nous allons maintenant nous intéresser à la deuxième caractéristique du comportement vivant selon Langton : l'évolution. On parle ici d'une évolution au sens darwinien du terme, c'est-à-dire un processus créateur de diversité inter-individuelle et intergénérationnelle au sein d'une population, par les mécanismes de sélection naturelle, de dérive génétique et de transmission héréditaire des caractères acquis. Depuis le réplicateur de Von Neumann, de nombreux automates cellulaires autorépliquants plus simples ont été construits et, parmi ces automates autorépliquants, les boucles de Langton ont tout particulièrement servies de base pour des tentatives d'y instancier des mécanismes évolutifs. Dans sa version de base, une boucle de Langton est un automate cellulaire simplement autorépliquant dont les mécanismes sont expliqués plus en détails dans (Langton, 1984). Pour faire court, une boucle est constituée d'une structure externe (son "phénotype") et d'une structure interne se déplaçant dans le sens inverse des aiguilles d'une montre dans la structure externe, et permettant la modification de la structure de la cellule et son autoréplication (son "génotype"). Dans la version classique, chaque boucle possède 7 gènes codant une croissance sans changer de direction puis 2 gènes codant un tournant à gauche. Après avoir suivi 4 fois ces instructions, l'excroissance de la cellule ainsi formée rentre en collision avec elle-même, provoquant la séparation de cette excroissance et ainsi la création d'une nouvelle cellule distincte : la boucle s'est auto-répliquée ! (voir figure 1)

Mais ce qui nous intéresse ici est la *version évolutive* de la boucle de Langton créée par H.Sayama. En éludant les détails techniques que vous pouvez retrouver dans (Sayama, 1998), Sayama permet la mutation d'un individu (d'une boucle) en introduisant le

Figure 1. Self-replication of an evoloop.

Fig. 1. **Les étapes d'autoréplication d'une boucle de Langton ((Sayama, 2004))**

concept de collision : quand deux boucles rentrent en collision, cela peut générer soit : une fusion des deux boucles et une interaction de leur génome, provoquant ainsi une mutation, soit une *attaque* d'une boucle sur l'autre, provoquant ainsi l'apparition d'un agent *dissolvant* qui détruit la structure des boucles, provoquant leur *mort*. De manière très simple, Sayama a donc introduit dans cet automate cellulaire les concepts de reproduction et de sélection naturelle ! Des études ultérieures comme (Salzberg, Antony, & Sayama, 2004) rapportent qu'une très grande diversité d'espèce peut être observée dans un système evoloop après quelques milliers d'itérations, chaque espèce étant capable de s'autorépliquer, que toutes ces espèces n'ont pas les mêmes chances de survies et qu'elles ne s'organisent pas de la même façon en *colonies* de la même espèce (certaines sont indépendantes, d'autres ont plutôt tendance à se regrouper) (voir figure 2).

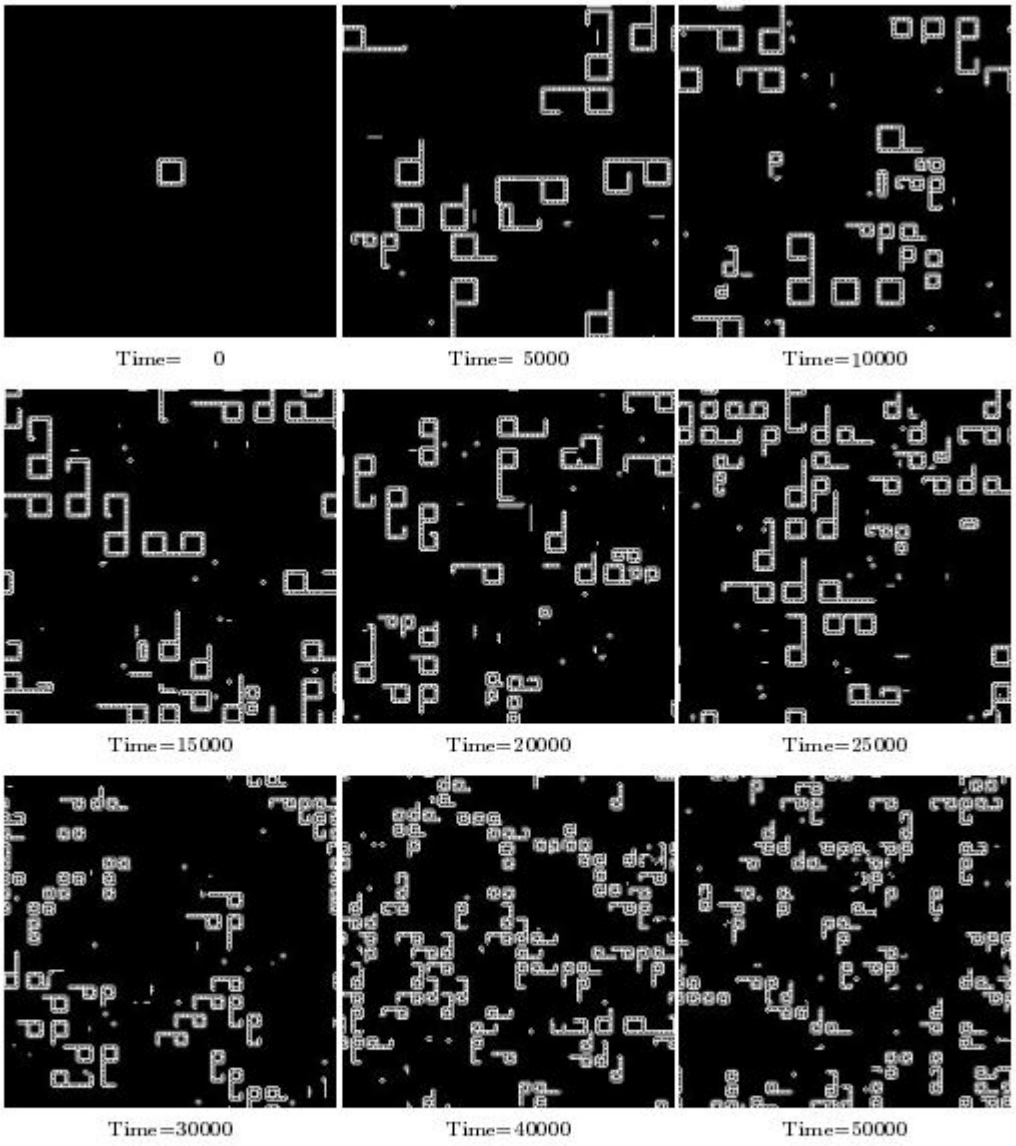

Figure 6.15: Temporal development of configuration in the evolutionary process of 2-*evoloops*. The ancestor is of species 13. The space is of 200×200 sites with periodic boundary conditions. As time proceeds, smaller loops emerge and dominate larger ones. The whole population gradually evolves toward smaller species, and finally the space becomes filled with the loops of species 4 which is strongest in this world.

Fig. 2. **Un aperçu du monde d'evoloop ((Sayama, 1998))**

Un autre exemple de tentative d'instancier des mécanismes évolutifs dans un médium artificiel est présenté par Thomas Ray dans (Ray, 1993). Dans cet article Ray présente un monde virtuel nommé Tierra dans lequel des organismes digitaux naissent, évoluent et meurent. Ce monde est entièrement défini par l'ensemble d'instructions exécutables par le processeur, la mémoire, et le système d'exploitation. Les organismes digitaux qui y évoluent sont composés par un ensemble d'instructions informatiques, l'espace dans lequel évoluent ces organismes est la mémoire de l'ordinateur, et le processeur qui analyse les instructions constitutives d'un individu et fait émerger son phénotype est considéré comme la source d'énergie utilisée par les individus. Chaque individu, au bout d'un certain temps, va *mourir*, c'est-à-dire que les instructions qui le composent vont être effacées de la mémoire. Avant cela, cet individu aura eu le temps d'exécuter une ou plusieurs fois ses instructions par l'intermédiaire du processeur. Ainsi, si les instructions de l'individu lui permettent de s'autoreproduire, c'est-à-dire de créer une copie de lui-même dans la mémoire de l'ordinateur, cet individu laissera des *descendants* avant de mourir. Les individus sont donc en compétition pour accaparer le temps de computation du processeur : en effet, plus leur ensemble d'instructions est exécuté, plus leur descendance sera nombreuse. Enfin, Ray modélise le nécessaire concept de dérive génétique en introduisant un processus qui inverserait aléatoirement certains 0 et certains 1 dans le code en binaire des instructions d'un organisme lors de sa création. Avec tous ces éléments en place, il ne reste plus qu'à lancer la simulation de Tierra en y introduisant un premier organisme capable d'autoréplication que Ray nomme l'Ancêtre. Dès que l'ancêtre est placé dans Tierra, il commence à s'auto-répliquer, puis la version originale de l'ancêtre meurt mais ses descendants s'autorépliquent à leur tour, etc .., chaque autoréplication présentant donc potentiellement des mutations. La première fois que T. Ray a laissé Tierra tourner en autonomie pendant une nuit, les résultats produits étonnèrent tout le monde, y compris Ray lui-même ! Dans cet univers digital s'était développé, de ses propres mots, "une incroyable ménagerie de créatures digitales". Parmi cette "ménagerie", on retrouve des organismes comparables à des "parasites", qui ne possédaient pas dans leurs instructions de code pour s'autorépliquer, mais qui avait la capacité d'utiliser des parties des instructions de leurs voisins, et ainsi pouvaient utiliser le code d'autoréplication des descendants de l'ancêtre. Du fait de leur rapidité d'exécution, ces "parasites" ont rapidement envahi le monde de Tierra, puisqu'ils pouvaient faire exécuter leur code plus de fois que les descendants de l'ancêtre, qui eux étaient plus lents. Mais, au fur et à mesure que la proportion de descendants de l'ancêtre diminuait par rapport à celle des parasites, le code contenant l'information nécessaire pour s'autorépliquer devenait de plus en plus rare et les parasites, ayant de la difficulté à trouver le code nécessaire à s'autorépliquer, commencèrent à décliner. Dans le même temps, les descendants de l'ancêtre restant, profitant du déclin des parasites, eurent accès à plus de mémoire et de temps de processeur et ont donc recommencé à croître rapidement. Cette croissance, provoquant une réinsertion massive de l'information génétique nécessaire à l'autoréplication, permit aux parasites de recroître, etc, etc .. En fait, il a été montré que les dynamiques cycliques des parasites et des descendants de l'ancêtre suivent un modèle de type Lokta-Volterra, qui est un modèle de dynamique de proies-prédateurs ou d'hôtes-parasites bien connu en modélisation de la biologie. De nombreuses autres caractéristiques de l'évolution prenant

place dans Tierra sont très intéressantes mais par souci de brièveté je ne les présenterai pas ici, je vous invite cependant à commencer à les découvrir dans (Ray, 1993).

Ces exemples invitent à penser la vie non pas exclusivement comme elle est présente sur Terre, mais bien comme elle pourrait être (par exemple, dans des médiums artificiels).

3. La vie comme l'autopoïèse.

Parallèlement aux approches de la vie artificielle basées sur les concepts d'autoréplication et d'évolution, se développe une école de pensée parallèle, portée par H.Maturna et F.Varela et inspirée par Merleau-Ponty, qui propose une redéfinition des critères pour qu'un système, un automate, une machine soit considérée vivante. Pour Maturna et Varela, l'autoréplication et l'évolution sont des processus qui présupposent l'existence d'une entité vivante, et ne peuvent ainsi pas être au centre de la définition de la vie. A la place, Maturna et Varela propose dans (Varela, Maturana, & Uribe, 1974) et dans d'autres essais, que le centre de la définition de la vie devrait reposer sur l'autonomie et l'individualité de l'entité vivante. Plus spécifiquement, Varela définie une entité vivante comme une machine autopoïetique, c'est-à-dire comme une machine

> organisée comme un réseau de processus de production de composants qui (a) régénèrent continuellement par leurs transformations et leurs interactions le réseau qui les a produits, et qui (b) constituent le système en tant qu'unité concrète dans l'espace où il existe, en spécifiant le domaine topologique où il se réalise comme réseau. Il s'ensuit qu'une machine autopoïetique engendre et spécifie continuellement sa propre organisation. Elle accomplit ce processus incessant de remplacement de ses composants, parce qu'elle est continuellement soumise à des perturbations externes, et constamment forcée de compenser ces perturbations. Ainsi, une machine autopoïetique est un système à relations stables dont l'invariant fondamental est sa propre organisation (le réseau de relations qui la définit) ((DUMOUCHEL, Bourgine, & VARELA, 1989))

et affirment que l'autopoïèse est nécessaire et suffisante pour caractériser l'organisation des systèmes vivants. Ainsi, Varela distingue les machines autopoïetiques vivantes des machines allopoïetiques non-vivantes par leur capacité à recréer et maintenir leur propre organisation et donc leur identité. Par exemple, tant qu'un animal répond à ses besoins vitaux, le réseau autopoïetique qui le caractérise va permettre de maintenir les *éléments* de l'animal, c'est-à-dire à ses cellules, ses tissus, etc .., qui à leur tour vont permettre à l'animal de continuer à interagir avec le monde en tant que lui-même. Au contraire, une voiture, qui est une machine allopoïetique, est composée d'éléments provenant de processus extérieurs à la voiture, et est incapable par ses propres opérations, de maintenir son identité. Varela précise également qu'une définition de la vie doit nécessairement prendre en compte la nature située des interactions d'une entité vivante avec son monde, et ne pas tomber dans le piège de vouloir définir la vie uniquement par des "abstractions désincarnées". Afin de proposer une réorientation de la recherche dans le domaine de la vie artificielle, Varela et et Bourgine proposent dans (Bourgine & Varela, 1992) deux

critères de définition d'un système autonome (qui est ici un synonyme pour machine autopoïetique) :

— **Hypothèse 1 :** "Tout système autonome est opérationnellement fermé" ;
— **Hypothèse 2 :** "Tout système autonome se comporte comme une machine abductive avec une capacité herméneutique acquise proche de l'unité sur la trajectoire de ses états", ou, dit plus simplement, tout système autonome se comporte comme une machine capable de trouver des comportements viables avec une très grande efficacité.

Bien que Maturna et Varela soient opposés à l'approche computationnelle/informationnelle de l'étude de la vie, leur approche partage une idée avec celle de Langton : une entité vivante est une machine complexe, machine autopoïetique pour Maturna et Varela, machine autorépliquante et évolutive pour Langton. Ainsi, Maturna et Varela ne s'opposent pas à la proposition comme quoi la vie peut être instanciée dans un médium artificiel (ce que l'on pourrait appeler l'*Alife forte*), et Varela propose même un modèle d'automate cellulaire où on peut observer l'apparition d'un agent autopoïetique. Cet automate cellulaire est décrit par quatre états possibles des cellules : les éléments de base, les liens, les liens attachés et une catalyse. Les règles d'interaction sont les suivantes : (1) Deux éléments de base peuvent former un lien en présence d'une catalyse, (2) Un lien peut s'attacher à deux ou plus autres liens, devant ainsi un lien attaché, (3) Un lien ou un lien attaché peut se séparer en deux éléments de base. Comme on peut le voir sur la figure, on observe la formation d'un agent autopoïetique dans cet automate cellulaire (voir figure 3).

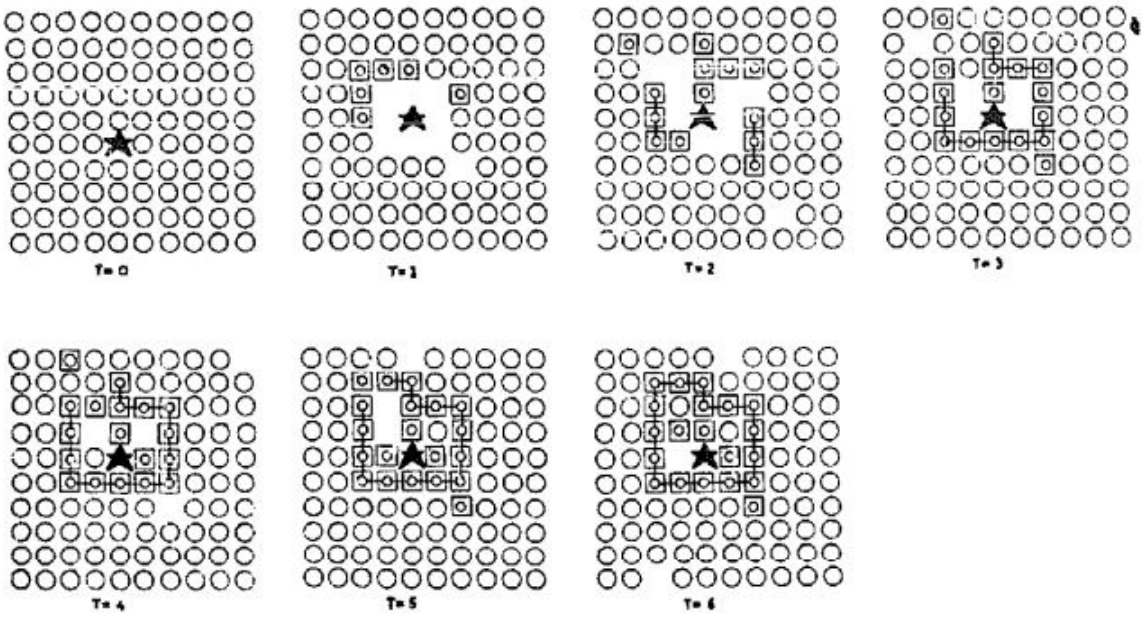

Fig. 1. The first seven instants (0→6) of one computer run, showing the spontaneous generation of an autopoietic unity. Interactions between substrate ○ and catalyst ★ produce chains of bonded links ⊡, which eventually enclose the catalyst, thus closing a network of interactions which constitutes an autopoietic unity within this universe.

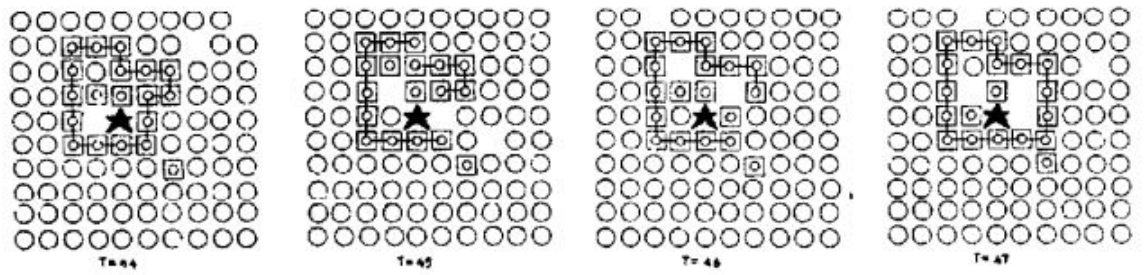

Fig. 2. Four successive instants (44–47) along the same computer run (Fig. 1), showing compensation in the boundary broken by spontaneous decay of links. Ongoing production of links re-establishes the unity under changes of form and turnover of components.

Fig. 3. **Un aperçu de l'automate cellulaire de Varela ((Varela et al., 1974))**

Cette définition de la vie comme l'autopoïèse implique cependant certaines choses qui dérangent notre idée intuitive de ce qu'est la vie. Par exemple, les systèmes immunitaires sont considérés par Varela lui-même comme des exemples de systèmes autopoïetiques et seraient donc vivants. Peut-être plus perturbant encore, certains systèmes sociaux peuvent être considérés comme autopoïetiques et donc vivants. Au vu de ces implications contre-intuitives, certains auteurs comme F. Guattari appelle à une reformulation de cette définition de la vie, en la rapprochant d'une approche évolutionnaire.

Deuxième partie . L'Intelligence.

Maintenant que nous avons présenté quelques éléments de définition de la vie notamment à travers l'étude de la vie artificielle, nous allons nous tourner vers une dimension plus cognitive, en tentant de discuter d'une définition de l'esprit et de l'intelligence, et encore une fois nous le ferons à travers le prisme des tentatives d'instancier ces concepts dans un médium artificiel, c'est-à-dire ici à travers le prisme de l'intelligence artificielle.

L'intelligence artificielle est certainement un des domaines scientifiques qui a le plus progressé depuis ces dernières années, et les avancées dans les domaines de la vision (reconnaissance d'objet), de la motricité (Boston Dynamics), des jeux (AlphaGo), de la navigation dans un environnement bruyant (voiture autonome), de l'analyse du langage (traduction, etc..) ne sont plus à présenter. Certaines avancées récentes dans le domaine des modèles génératifs dont le représentant le plus connu est Deep Dream de Google laisse entrevoir une possibilité d'avoir des machines imaginatives, comme par exemple dans (Radford, Metz, & Chintala, 2015), où une machine est capable de produire de nouvelles images représentant des chambres après qu'on lui ait dans un premier temps fourni quelques images de chambres à analyser. Ces résultats sont, à n'en pas douter, très impressionnants. Mais sont-ils représentatifs de l'intelligence dans une machine ? De l'incarnation d'un esprit dans une machine ?

Tout comme cela a été le cas dans notre aperçu de la vie artificielle, la question de l'intelligence artificielle va rapidement nous amener à nous poser une question plus fondamentale : *Qu'est-ce que l'intelligence ?*. Cette question a une triple importance pour le domaine de l'intelligence artificielle : (1) Statuer de la possibilité ou non d'avoir une intelligence dans un médium artificiel et donc justifier l'existence même de l'intelligence artificielle : *Les machines peuvent-elles être intelligentes ?*, (2) orienter la recherche : *Comment s'y prendre pour construire des machines intelligentes ?* et (3) décider si une machine est intelligente ou non, c'est-à-dire statuer de la réussite de l'intelligence artificielle : *A-t-on réussi à construire des machines intelligentes ?*. Un aperçu des tentatives de définition de ce concept d'intelligence sera l'objet de cette partie.

Cependant nous allons rapidement voir qu'une telle tentative de définition de l'intelligence est assez limitante, et nous nous tournerons plutôt vers des questions telles que *Qu'est-ce que l'esprit ?*, *Qu'est-ce que la conscience ?* et *Une machine peut-elle avoir une conscience ?*, peut-être plus fondamentales encore. La question de l'intelligence et celle de l'esprit et la conscience sont bien deux questions distinctes, puisqu'une machine exhibant des comportements intelligents pourrait très bien ne pas être dotée d'un esprit ou d'une conscience. En fait, même si l'on disposait d'un programme informatique agissant en tout point parfaitement comme un humain, il y aurait encore des questions philosophiques difficiles auxquelles il faudrait répondre, notamment *Est-ce que ce programme a un esprit, une conscience ?*. C'est de cette observation que Searle distingue la position de l'intelligence artificielle "forte" : "Un système de symbole physique peut

avoir un esprit et des états mentaux" de celle de l'intelligence artificielle "faible" : "Un système de symbole physique peut agir intelligemment" (comme exprimé dans (Searle, 2008)). Aujourd'hui, le débat philosophique s'oriente plutôt sur des problématiques liées à l'intelligence artificielle "forte", tandis que la possibilité d'une intelligence "faible" est admise par une grande partie de la communauté. Dans la troisième et dernière partie, nous aborderons ces sujets d'esprit et de conscience, et nous essaierons en particulier de répondre à la question *Une machine peut-elle avoir une conscience ?*.

Nous allons tout d'abord résumer brièvement les grandes approches dans la définition de l'intelligence qui peuvent être prisent par des adeptes de l'intelligence artificielle, puis nous aborderons le bien connu test de Turing, et enfin nous présenterons les travaux peut-être moins connus de M.Hutter dans lesquels il propose une définition formelle et universelle de l'intelligence.

Dans le domaine de l'intelligence artificielle on peut parfois se retrouver confronté au terme d'*agent intelligent*. Un *agent* peut être défini, d'après (Russell & Norvig, 2016) comme une entité qui interagit avec son environnement et qui possède une *mesure de ses performances*. Cet agent est parfois dit *intelligent* s'il maximise cette mesure de performance, en se basant sur ses propres interprétations et représentations du monde. C'est-à-dire, en quelque sorte, l'intelligence comme la rationalité. Cependant, cette définition ne parvient pas à s'accorder avec le sens commun de ce qui est intelligent et de ce qui ne l'est pas : en effet si l'on prend cette définition, un simple outil de mesure comme un thermomètre est intelligent ! Il semble alors clair que, soit le terme *intelligent* a été détourné, soit il est nécessaire d'affiner cette définition. Une autre approche de l'intelligence, est celle de l'intelligence comme la capacité à résoudre des problèmes complexes. C'est-à-dire qu'on va chercher à construire une machine intelligente *pour répondre à un certain problème*. Cette approche de l'intelligence, notamment portée par M.Minksy, est caractéristique de ce qu'on pourrait appeler l'intelligence artificielle classique. Une mise en pratique de cette définition pourrait être, par exemple, la construction d'une machine dans le but d'être performante à un certain jeu de plateau. Dans ce domaine, les machines sont en fait aujourd'hui capables de battre n'importe quel humain, comme ce fut le cas pour les échecs avec Deep Blue contre G.Kasparov, et plus récemment pour le Go avec Lee Sedol contre AlphaGo. Nous ne nous attarderons pas dans cet essai sur les techniques employées pour construire ces machines ni sur l'historique des rencontres, mais remarquons tout de même que dans ces deux matchs, les machines ont été, à certains moments, considérées par les commentateurs comme "jouant avec un certain style", ou produisant des coups "innovants et imaginatifs". Et si le fonctionnement de Deep Blue et de la plupart des machines joueuses d'échecs est très différent de celui d'un expert humain des échecs, le fonctionnement de AlphaGo et particulièrement des dernières versions comme AlphaGo Zero peut, lui, être rapproché de celui d'un expert du jeu de Go (en effet AlphaGo Zero ne fonctionne pas avec un algorithme de type *brute-force* qui teste toutes les possibilités de coup, mais apprend plutôt des concepts tactiques et stratégiques et des patterns de coup). Pour une partie de la communauté scientifique, il est tout à fait fondé de dire que ce genre de machines sont intelligentes mais d'une façon différente de l'humain, elles auraient une sorte d'*intelligence experte*, très développée,

mais n'existant que dans un seul environnement. Une autre partie de la communauté scientifique, au contraire, pense que l'intelligence serait mieux décrite par la capacité d'adaptation d'un individu à son environnement. Ainsi l'intelligence serait la capacité d'un individu à réaliser des objectifs complexes dans un large panel d'environnements, en déployant des comportements adaptés à cet environnement particulier. Cette définition de l'intelligence est à mettre en parallèle avec l'algorithme de la sélection naturelle, qui sélectionne par la survie les plus adaptés, et donc, d'après cette définition, les plus intelligents. Malgré cela, cette définition n'assimile pas l'intelligence à la survie, la survie est simplement une marque d'intelligence, dans le sens où elle démontre que l'individu a été capable de s'adapter pour répondre à *un de* ses objectifs : survivre.

Dans les années 1950, A.Turing, qui trouvait ces débats sur la nature de l'intelligence infertiles ou trop abstraits, proposa, afin de contourner le problèmes, un test d'intelligence pour les machines : le test de Turing, aussi connu sous le nom de *jeu de l'imitation*. Dans ce test, une machine est intelligente si un juge humain est incapable de faire la différence entre une conversation par messages écrits avec cette machine et avec un autre humain. Le test de Turing, bien que très novateur pour son époque et restant encore aujourd'hui une référence, a souffert de beaucoup de critiques. Par exemple, Searle propose que le test de Turing ne soit pas *suffisant* pour déterminer si une machine est intelligente ou non. Son argument est qu'une machine disposant d'une table de correspondance faisant correspondre toute questions (ou toute phrases) à une réponse adéquate pourrait passer le test de Turing sans véritablement faire preuve d'intelligence, juste en cherchant dans une base de données. Bien sûr, une telle machine semble impossible à construire en pratique de par la mémoire qu'elle nécessiterait, mais rien que le fait qu'une *machine théorique* non-intelligente puisse passer le test de Turing remet grandement en cause l'efficacité du test. Pour pallier à ce problème, certaines versions dérivées du test de Turing ont vu le jour, comme celle proposée par Dowe dans (Dowe & Hajek, 1997) où il propose une extension non pas comportementale mais computationnelle au test de Turing, en argumentant que la compression de données et l'apprentissage inductifs sont également des éléments nécessaires à l'intelligence, introduisant donc une limite dans la mémoire d'une machine pour qu'elle soit recevable dans son test de Turing modifié et réglant ainsi le *problème de la table de correspondance*. Une autre critique du test de Turing propose que le test de Turing ne soit pas *nécessaire* pour établir qu'une machine est intelligente, avec comme argument principal que le test de Turing est trop anthropocentré. En effet, passer le test de Turing nécessite surtout que la machine possède un modèle de la connaissance et de la façon de penser humaine, rendant ainsi le test de Turing un test d'humanité et non un test d'intelligence. Et en effet, en pratique, une machine voulant passer le test de Turing devra, par exemple, limiter sa capacité à faire des calculs rapidement pour pouvoir *paraitre humaine* : on s'éloigne de l'objectif du test. Ainsi, on voit que malgré la place centrale du test de Turing dans les tests d'intelligence d'une machine, il souffre de nombreuses critiques fondamentales assez difficilement contournables.

Dans (Legg & Hutter, 2007), M.Hutter et S.Legg s'attaquent au problème de la définition de l'intelligence générale dans un contexte non-anthropocentrée, et propose une définition formel de l'intelligence en générale. Dans un premier temps, Hutter et Legg analysent un corpus de définitions de l'intelligence proposé par des experts, et en ex-

traient les caractéristiques communes, afin d'aboutir à une définition de l'intelligence indépendante de l'auteur et condensant toutes les propriétés principales qu'il convient d'attribuer à ce concept. Hutter et Legg arrivent ainsi à cette définition informelle de l'intelligence :

> L'intelligence mesure la capacité d'un agent à atteindre ses buts dans un large panel d'environnements.

Ils s'emploient alors à formaliser cette définition dans un cadre théorique général et bien posé.

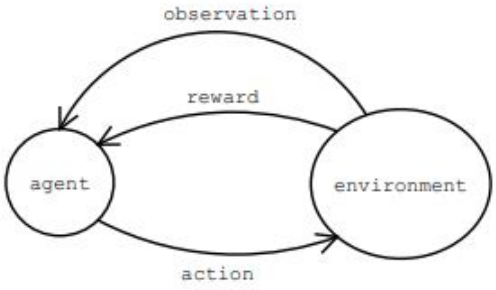

Figure 1: The agent and the environment interact by sending action, observation and reward signals to each other.

Fig. 4. **Le système agent-environnement ((Legg & Hutter, 2007))**

Dans leur définition, on retrouve trois composantes principales qui doivent être modélisées : un agent, des environnements et des objectifs. L'agent et l'environnement sont deux entités qui interagissent entre eux : de l'agent vers l'environnement par l'intermédiaire des *actions* et de l'environnement vers l'agent par l'intermédiaire des *observations*. Afin de modéliser les objectifs de l'agent, l'environnement est également capable de communiquer à l'agent des *récompenses*, qui sont des mesures de la *qualité* de la situation dans lequel l'agent est actuellement. L'objectif de l'agent est alors de maximiser ses récompenses. Hutter et Legg soutiennent que cette définition des objectifs de l'agent comme la maximisation d'une mesure de succès n'est pas limitante car

> Si l'agent veut réussir dans son environnement, c'est-à-dire, recevoir beaucoup de récompenses, il doit apprendre la structure de son environnement et en particulier ce qu'il doit faire pour recevoir des récompenses.

Nous allons présenter ici rapidement la formalisation de Hutter et Legg, ou en tout cas tenter d'en donner l'intuition, mais nous vous invitons à consulter (Legg & Hutter, 2007) pour plus de précisions sur la démarche. Les actions entreprises par l'agent seront notées a_i où a_i est donc la ième action, de même les observations seront notées o_i et les récompenses r_i. L'historique du comportement du système peut ainsi être résumé par une chaine d'actions, d'observations et de récompenses de la forme $o_1 r_1 a_1 o_2 r_2 a_2 \ldots o_i r_i a_i \ldots$. L'agent, noté π, et les environnements, notés $(\mu_i)_{i \in \mathbb{N}}$, sont des fonctions qui prennent en entrée l'historique du comportement du système et décide respectivement de la prochaine action et de la prochaine récompense et observation en fonction de cet historique.

Par exemple, en langage mathématique, on va noter $\pi(a_3|o_1r_1a_1o_2r_2)$ la probabilité que l'agent effectue l'action a_3 sachant qu'*il s'est déjà passé* $o_1r_1a_1o_2r_2$. Pour continuer, il semble nécessaire d'établir une mesure de la réussite de l'agent π dans un environnement μ. Hutter et Legg choisissent de prendre pour cette mesure l'espérance de la somme des récompenses, c'est-à-dire,

$$V_\mu^\pi = E(\sum_{i=1}^{+\infty} r_i)$$

De plus, comme leur définition de l'intelligence se veut la plus large et englobante possible, il va être nécessaire de tester l'agent sur un grand nombre d'environnements : en fait, Hutter et Legg proposent de tester l'agent sur tous les environnements possibles, avec la restriction qu'ils doivent être possible à simuler informatiquement. Maintenant, la prochaine et dernière étape est d'assembler toutes les mesures de performances sur chacun des environnements en une seule mesure, qui sera une *mesure d'intelligence*. Comme il y a une infinité d'environnements, il est nécessaire, afin de garder une mesure finie, d'attribuer une importance plus grande à certains environnements par rapport à d'autres. Pour faire cela, Hutter et Legg instancient le principe du rasoir d'Ockham (privilégier la simplicité) en accordant une importance décroissante aux environnements par rapport à leur complexité algorithmique de Kolmogorov, qui mesure de la complexité d'un environnement comme la longueur du plus petit programme en binaire capable de produire cet environnement, c'est-à-dire que la complexité d'un environnement μ, $K(\mu)$ est définie par

$$K(\mu) = \min_p \{l(p) : \mathcal{U}(p) = \mu\}$$

où p représente un programme en binaire, $l(p)$ la longueur de ce programme et $\mathcal{U}$ une machine de computation appelée machine de référence.

Ainsi, au final, Hutter et Legg définissent l'intelligence universelle d'un agent π par

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$$

où E est l'ensemble des environnements possibles.

Cette définition formelle de l'intelligence générale possède plusieurs qualités qui sont importantes pour une définition de l'intelligence, notamment (toujours selon (Legg & Hutter, 2007)) :

1. **Valide** - on peut raisonnablement appeler ce que décrit cette définition comme étant *l'intelligence*, étant donné que les auteurs sont partis d'un corpus de définition de l'intelligence par des experts
2. **Informative** - $\Upsilon(\pi)$ est une mesure de l'intelligence sous la forme d'un score, d'un nombre, ce qui permet par exemple de faire des comparaisons entre les agents
3. **Couvre un large panel d'agents** - Cette définition de l'intelligence est capable de déterminer l'intelligence d'un agent aléatoire comme d'un agent super-intelligent (AIXI), et de tous les agents entre ces deux extrêmes

4. **Générale et non-biaisée par la tâche** - En testant un agent sur tous les environnements *calculables* possibles, Υ est une mesure la plus générale possible, et n'est donc pas biaisée quant à la *tâche* demandée à l'agent
5. **Non-anthropocentrique**.

Ainsi, cette définition de l'intelligence semble répondre de manière raisonnable à la question *Qu'est-ce que l'intelligence ?* et apporte en tout cas un cadre conceptuel intéressant pour penser l'intelligence de manière objective et non-anthropocentrée.

Troisième partie . Possibilités d'une conscience dans une machine.

Comme nous l'avons vu, il y a de bons arguments en faveur de l'idée selon laquelle la vie et l'intelligence pourraient être instanciées dans des médiums artificiels. Mais qu'en est-il de la conscience ? Pour certains, la conscience est une barrière infranchissable et restera à jamais incompréhensible et *non-artificialisable*, pour d'autres, la conscience sera le dernier défi qui se présentera au processus d'artificialisation des propriétés cognitives (au sens large), mais ce défi est *a priori* relevable, tandis que pour d'autres encore, la conscience artificielle ne sera qu'une étape de plus dans ce processus d'artificialisation, similaire à toutes les autres (vision, motricité, intelligence, apprentissage, évolution, etc..). Dans cette partie nous allons encore une fois accorder une grande importance à la formulation d'une définition de la conscience, et nous étudierons la possibilité théorique d'instancier une conscience dans une machine mais aussi les approches pratiques actuelles du problème. Le ton dans cette dernière partie sera plus réservé que dans les deux premières, pour la simple raison que le domaine de la conscience artificielle n'a pas encore produit de résultats très significatifs.

1. Problemes corps/esprit et impacts sur les possibilites d'une conscience dans une machine.

Lorsqu'on compare naïvement les propriétés physiques du corps et les états mentaux attribués à notre esprit, on note de grandes différences, par exemple : les lois physiques de la matière sont connues et observables par tous (si l'on dispose du bon matériel), mais les états mentaux sont privés, personnels, on ne peut pas *ressentir* les états mentaux de quelqu'un d'autre. Ce genre d'observation pose le problème de la nature, de l'essence de l'esprit et de la relation entre le corps et l'esprit, qui a toujours été un problème central en philosophie de l'esprit. Grossièrement, les trois grandes approches de ce problème sont les suivantes : (1) Monisme matérialiste : l'esprit et le corps appartiennent à la

même et unique catégorie d'entités : les entités matérielles (2) Dualisme : l'esprit et le corps, le mental et le matériel sont deux choses ontologiquement totalement séparées. (3) Monisme idéaliste : les états physiques sont des états mentaux, et le monde physique tel que nous l'expérimentons est le produit de notre expérience mentale collective. Cette partie se veut être une introduction rapide à quelques points de vue sur le problème corps/esprit et a pour but de statuer de la possibilité ou non d'une conscience dans une machine en fonction de si l'on accepte telle ou telle théorie, et non pas de répertorier scrupuleusement toutes les approches de ce problème.

Nous allons laisser de côté l'hypothèse idéaliste, qui semble difficilement articulable avec les thèmes et le point de vue scientifique de cet essai. Nous allons donc discuter du dualisme cartésien et émettre certaines critiques de cette forme de dualisme, puis nous aborderons la vision du monisme matérialiste sur l'esprit, et enfin nous présenterons le point de vue de D.Chalmers et l'hypothèse panpsychique, qui rassemble différenciation ontologique entre mental et physique et possibilité d'une conscience dans une machine.

Dans son discours de la méthode ((Descartes & Gilson, 1987)), Descartes propose qu'il est possible de douter du monde matériel et donc notamment de l'existence de notre corps, mais qu'il est impossible de douter de l'existence de notre esprit, car la pensée est une prémisse indispensable au doute. Ainsi, Descartes conclut qu'il est possible que *mon* corps n'existe pas, mais qu'il est impossible que *mon* esprit n'existe pas, et il en conclut que l'esprit et le corps sont deux entités ontologiquement séparées ! Mais la construction logique nécessaire à arriver à cette conclusion a souffert de certaines critiques, et surtout, accepter l'hypothèse du dualisme cartésien entraine un certain nombre de problèmes plutôt épineux. Notamment, la question de la relation de l'esprit et du corps et plus particulièrement la façon dont l'esprit fait bouger le corps est un problème du dualisme qu'il semble bien difficile de résoudre. En effet, les lois de la physique nous apprennent que, dans le monde matériel où évolue mon corps, le mouvement d'un objet matériel a nécessairement pour cause un autre mouvement, une perturbation de la matière (c'est ce qu'on appelle la *complétude causale* du domaine des états physiques). Mais alors comment expliquer que l'esprit, qui évolue dans un monde immatériel, puisse agir sur le corps, qui évolue dans le monde matériel ? Depuis le 17$^{\text{ième}}$ siècle et la formulation du dualisme cartésien, certain philosophe ont essayé de répondre à cette question, mais sans grand succès par rapport à des critères scientifiques modernes puisque les réponses à cette question, de Descartes à l'Harmonie préétablie de Leibniz en passant par l'Occasionnalisme de Malebranche, impliquent l'intervention de *Dieu*. En ce qui concerne notre problème de la possibilité d'un esprit dans une machine, les partisans d'un dualisme classique répondraient tout simplement que c'est impossible, puisque si l'esprit et le matériel sont deux mondes de réalités différents, il n'y aucune raison qu'une construction physique d'abord inerte (sans esprit) puisse être doté d'un esprit à partir d'une certain degré de complexité ou en assemblant certains composants *sans esprits* d'une certaine façon.

Au contraire, si l'on se place dans le cadre du monisme matérialisme, l'esprit est une conséquence de phénomènes physiques. Cette hypothèse semble déjà être plus clémente quant à sa réponse à la question *Peut-il y avoir une conscience dans une machine ?*, mais encore faut-il savoir de quel type de matérialisme parle-t-on. En effet, en temps qu'hy-

pothèse de base la plus populaire dans la philosophie et les sciences de l'esprit modernes, le matérialisme se voit subdivisé en une multitude de théories, qu'il est impossible de présenter dans leur intégralité ici, aussi nous nous limiterons aux théories de type identité cerveau-esprit et au fonctionnalisme. La théorie de l'identité cerveau-esprit énonce que les états mentaux sont *identiques* aux états physiques du cerveau, c'est-à-dire que pour chaque type d'état mental M, il existe un type d'état physique du cerveau P tel que M et P soit identiques. Cette hypothèse, dans ses formes les plus faibles, n'est pas en contradiction avec la possibilité d'un esprit dans une machine, mais de manière générale, l'hypothèse de l'identité cerveau-esprit ne permet pas à un esprit de prendre place dans une machine, à moins d'être capable de reproduire exactement les états physiques du cerveau dans une machine. Le physicien R.Penrose affirme qu'il sera impossible d'instancier une conscience dans une machine parce que la conscience est un phénomène lié à la structure même des neurones. Une des formes de matérialisme qui est plus favorable à la possibilité d'un esprit dans une machine est le fonctionnalisme, et le fonctionnalisme a également le bon goût de résoudre le problème des "versions multiples" énoncés par Putman et considéré comme une limite très handicapante des hypothèses de type identité cerveau-esprit [2]. Le fonctionnalisme énonce qu'il ne faut pas définir les états mentaux par des propriétés de la matière, mais plutôt par la fonction de ces états mentaux au sein du mental ou au sein de l'organisme, la matière n'étant que la base permettant la réalisation de cette fonction. Cette hypothèse ne fait aucun doute sur le fait qu'un esprit artificiel est possible, et étend le domaine des entités possédant un esprit à au moins tous les systèmes *fonctionnant comme* le système nerveux humain. En effet, on peut par exemple imaginer créer des neurones artificiels qui reproduisent la fonction des neurones biologiques et les agencer (c'est-à-dire les connecter) de la même façon que les neurones dans un cerveau biologique (notons que ce n'est ici qu'un exemple qui s'inspire de la manière de faire de la Nature, mais il est possible que réaliser les fonctions d'un cerveau puisse se faire autrement). Certaines expériences de pensée que l'on peut faire à partir du fonctionnalisme sont assez dérangeantes pour notre intuition concernant la nature de l'esprit. Par exemple, si l'on réunissait autant d'humains que de neurones dans le système nerveux humain et que l'on munissait chaque individu d'un appareil capable de simuler le comportement d'un neurone (capter, traiter et renvoyer des informations) et qu'on agencait physiquement les individus de la même manière que dans le système nerveux humain, alors le fonctionnaliste devrait s'attendre à ce qu'une forme de conscience émerge de ce *système*.

Malgré la prédominance du monisme matérialiste sur la scène moderne de la philosophie de l'esprit, certaines formes de simili-dualisme sont toujours présentes. Par exemple, D.Chalmers voit la conscience comme une loi fondamentale de la Nature au même titre que le temps et l'espace, et ajoute également que cette loi est à caractère universelle, héritant ainsi du *panpsychisme* qui considère que certaines entités fondamentales du monde (comme les atomes) ont des états mentaux conscients (ou des précurseurs d'états

2. Putman énonce dans (Putnam, 1980) qu'il est plausible que des formes de vie soient dans le même état mental sans nécessairement avoir des cerveaux dans la même configuration physico-chimique unique, ce qui remet en question l'idée selon laquelle chaque état mental peut être associé à un unique état physico-chimique du cerveau

mentaux conscients dans le cas d'un *panprotopsychisme* comme décrit par Russell ou plus récemment Peter Unger)[3]. Cette hypothèse panpsychique s'éloigne du dualisme classique et ce parce que, malgré le fait qu'elle suppose l'existence de deux modes d'existence des entités, ces deux modes d'existence sont présents en même temps dans chaque entité, ainsi, la thèse du panpsychisme peut être vu comme

> La thèse que tout est (ou au moins certains choses sont) fondamentalement physique(s) et fondamentalement mental(es) ((Chalmers, 2015))

. C'est une idée qui peut sembler contre-intuitive de prime abord, qui découle en partie de l'idée de penser la conscience en terme de traitement de l'information (lorsque ce traitement est complexe, comme chez l'homme, on a une *conscience complexe*, et lorsqu'il est plus simple, une *conscience simple*). Cette approche de la conscience en temps qu'intégration de l'information a d'ailleurs fait l'objet d'une formalisation mathématique par G.Tononi sous le nom de *théorie de l'information intégrée*, aboutissant à une mesure mathématique corrélée à la conscience. Dans son étude de la conscience, Chalmers a été amené à proposer le principe d'invariance organisationnelle, qui énonce que deux systèmes physiques avec précisément la même organisation fonctionnelle auront la même expérience consciente, car les expériences conscientes surviennent naturellement sur les systèmes physiques. Ce principe est assimilable au fonctionnalisme à ceci près qu'il n'est ici pas matérialiste mais dualiste ou plutôt *panpsychique*. Ce principe d'invariance organisationnelle permet de répondre positivement à la question *une machine peut-elle avoir un esprit, une conscience ?*, en effet, reproduire les caractéristiques fonctionnelles du cerveau humain dans une machine serait une condition suffisante pour que cette machine ait un esprit, de la même manière que pour le matérialisme fonctionnaliste.

2. D'autres difficultes dans la definition de l'esprit et de la conscience.

Depuis quelques années, on observe une explosion du nombre de travaux sur la conscience, aussi bien en sciences qu'en philosophie, et les points de vue présentés et les thèses défendues sont nombreuses. Nous allons dans ce paragraphe donner quelques éléments complémentaires par rapport à la section précédente, qui portait surtout sur l'essence de l'esprit. Dans (D. C. Dennett, 2017), D.Dennett propose la théorie des versions multiples et fait un argument contre l'idée du *théâtre cartésien*. L'idée du *théâtre cartésien* est celle selon laquelle il y aurait dans notre cerveau un *observateur* à qui sont présentés les évènements conscients et qui serait responsable de notre flux de conscience, c'est-à-dire de notre expérience du monde comme une suite ininterrompu de sensations, d'informations et de pensées. Cette idée ne convient pas à D.Dennett, qui cherche à expliquer la conscience par des procédés naturalistes et réductionnistes, et il se trouve que le cerveau, siège de la conscience dans une approches naturalisante, est un système massivement parallèle qui ne possède pas de *contrôleur central*. Dennett propose ainsi

3. ce qui a des implications éthiques pour le moins intéressantes mais que je n'aborderais pas dans cet essai

d'abandonner l'idée du théâtre cartésien et formule sa théorie des versions multiples. Cette théorie énonce qu'il existe non pas une seule, mais de multiples versions subjectives de notre expérience qui sont *vécues* par un sujet en même temps. Ce qui fait que l'on va avoir *conscience* d'une des versions parmi les autres, c'est qu'elle possède une plus grande utilité et sera donc amené à subsister dans la mémoire. Pour D.Dennett, un être conscient est un être capable de produire cette multitude de versions de notre expérience et de les utiliser. En contradiction (parfois directement et parfois indirectement) avec cette vision réductionniste de la conscience, on retrouve l'idée des *Qualia* et du *Problème difficile de la conscience* comme posé par Chalmers. Les *Qualia* sont les qualités ressenties de nos expériences conscientes, c'est-à-dire, les propriétés d'une expérience qui font que cela "fait un certain effet" pour vous de vivre cette expérience. Pour donner une intuition de ce que sont ces *qualia*, nous allons présenter l'expérience de pensée de la *chambre de Marie* de Frank Jackson

> Marie est une brillante scientifique qui est forcée, peu importe pour quelle raison, d'étudier le monde depuis une chambre noire et blanche par le moyen d'un écran de télévision en noir et blanc. Elle se spécialise dans la neurophysiologie de la vision et nous supposerons qu'elle acquiert toutes les informations physiques qu'il y a à recueillir sur ce qui se passe quand on voit des tomates mûres ou le ciel, et quand nous utilisons des termes comme « rouge », « bleu », etc. Par exemple, elle découvre quelle combinaison de longueurs d'onde provenant du ciel stimule la rétine, et comment exactement cela produit, via le système nerveux central, la contraction des cordes vocales et l'expulsion d'air des poumons qui aboutissent à la prononciation de la phrase : « Le ciel est bleu ». [...] Que se produira-t-il quand Marie sortira de sa chambre noire et blanche ou si on lui donne un écran de télévision couleur ? Apprendra-t-elle quelque chose, ou non ? ((Broad, 2014), traduction de Wikibooks)

Si l'on pense que Marie apprendra bien quelque chose en voyant disons une rose rouge pour la première fois, alors on admet l'existence des *qualia*, et le *quale rouge* s'imposera à l'expérience consciente que Marie fait de la rose rouge. T. Nagel propose dans (Nagel, 1974), dans le même esprit, de définir la conscience comme "l'effet que cela fait" d'être une certaine entité consciente dans une certaine situation. Nagel ajoute que pour savoir "l'effet que cela fait" d'être une chauve-souris, il serait necessaire de "prendre le point de vue de la chauve-souris", et que les visions matérialistes réductionnistes de la conscience semblent oublier cette nécessité. Chalmers propose de nommer le problème de l'experience subjective du sujet et des *qualia* le "Problème difficile de la conscience". Pour tenter de structurer ces débats sur la conscience, N.Block introduit une séparation entre la conscience *phénoménale* et la conscience *d'accès*. Pour Block, beaucoup des débats et des incompréhensions dans les recherches sur la conscience viennent d'une confusion entre ces deux *types* de conscience, comme il l'exprime dans (Block, 1995). La conscience d'accès correspond au type de conscience étudiée par les neurosciences et la psychologie et aussi celle sur laquelle se concentre Dennett et les réductionnistes, et peut être définie par ce qui est "directement disponible pour un contrôle global". La conscience phéno-

ménale, elle, ne peut pas être définie selon Block, on peut simplement indiquer qu'il s'agit de "l'effet que cela fait" et de l'expérience des *qualia* comme expliqué plus haut. Tous ces débats sur la nature de la conscience et surtout la nature très différente des points de vue et des théories qui sont explorées nous apprennent surtout que l'approche de la conscience par les sciences et la philosophie de l'esprit moderne ne nous apprends pour l'instant rien d'absolument certain, et que ce champ de recherche sera certainement amené à vivre de nombreux rebondissements.

Un des problèmes fondamentaux dans la construction d'une machine avec une conscience phénoménale est de déterminer la réussite ou non d'une telle entreprise. En effet, le flux de conscience d'une entité est quelque chose de fondamentalement subjectif et *je* ne peut pas *ressentir* le flux de conscience d'un autre, c'est-à-dire en faire l'expérience subjective. Ainsi, et de la même manière que je ne peux être *absolument certain* que de ma propre conscience et que je ne peux que très fortement supposé l'existence d'une expérience subjective chez les autres êtres humains, nous ne pourrons pas être absolument certain qu'une machine ait une expérience subjective et des *Qualia*. Il sera donc nécessaire, comme nous le faisons avec les autres êtres humains, de se convaincre par son comportement, son utilisation du langage (ou plus généralement sa manière de communiquer) qu'une machine est consciente. Se pose alors la question d'une *machine-zombie* (équivalent du *p-zombie* de Chalmers), c'est-à-dire d'une machine qui exhiberait des comportements qui sembleraient caractéristiques d'un être conscient, mais sans être réellement conscient, de la même manière que ELIZA semble comprendre le langage humain alors qu'elle ne fait que suivre un petit ensemble de règles, sans jamais former de concepts, de pensées ou d'images mentales ... Il serait alors en théorie impossible de distinguer une machine consciente d'une *machine-zombie*. Cependant, en pratique, il est peu probable qu'une *machine-zombie* parfaite existe, mais des machines non-conscientes qui présentent des comportements que l'on pourrait interpréter comme conscients vont surement voir le jour ou ont déjà vu le jour, d'où l'importance d'établir des critères de conscience et des tests de conscience pour les machines, de la même manière que les tests d'intelligence présentés dans la partie 2. Dans les années 1970, le psychologue G.Gallup propose le test de miroir afin d'estimer la conscience de soi chez l'animal. Le principe de ce test est le suivant : on place une marque sur le corps d'un individu et on place cet individu devant un miroir. Si l'individu tente d'enlever ou de toucher la trace, ou simplement semble avoir noté le changement, alors on en conclut que l'animal à conscience de son corps. Il semblerait que les machines soient aujourd'hui capables de passer ce test sans pour autant qu'on ne leur attribue une conscience. La formulation d'un ou plusieurs autres tests du même type que le test du miroir mais au jour des nouvelles approches de la conscience et ciblant précisément les éléments constitutifs de la conscience semblent nécessaire afin de guider la recherche en conscience artificielle.

3. Approches du problème.

Nous allons dans cette dernière partie discuter de certaines approches plus concrètes de la question de l'esprit et de la conscience dans les machines. Nous parlerons tout

d'abord d'informatique évolutionnaire, puis nous aborderons une tentative de définir la conscience comme un ensemble de mécanismes de traitement de l'information, enfin nous discuterons de l'approche consistant à modéliser le système nerveux.

Nous avons abordé les questions de l'essence de l'esprit et de la conscience et nous avons donné quelques éléments de classification et de définition de la conscience, mais nous ne nous sommes pas encore posé la question de la raison de l'existence de la conscience [4], c'est-à-dire, *Pourquoi sommes-nous conscients ? Pourquoi suis-je le sujet d'une expérience consciente subjective ?*. Ces questions prennent tout leur sens lorsqu'on les considère avec la théorie de l'évolution de Darwin. En effet, quel est *l'intérêt* pour un individu d'être conscient ? Un individu sans conscience ni expérience subjective mais ayant (exactement) les mêmes schémas de réponses comportementaux que l'individu conscient sera placé sur un pied d'égalité en termes de survie. Pourtant, il est bien tentant, lorsqu'on essaye d'expliquer la conscience, de dire qu'elle provient de l'évolution biologique du vivant, ou au moins qu'elle s'est complexifiée au fur et à mesure de l'évolution du vivant. Pourtant, dans (Rosenthal, 2012), Rosenthal présente des arguments convaincants sur l'idée que *l'utilité biologique* de la conscience phénoménale est faible ou inexistante. C'est-à-dire, non pas que de pouvoir accéder et manipuler des informations à propos du monde n'a pas d'utilité à la survie d'un individu, mais c'est l'accompagnement de cet accès et manipulation par une expérience subjective, par un "effet que cela fait", qui lui n'a pas de justification en termes sélection naturelle. Ainsi, il semble nécessaire d'expliquer la conscience en des termes autres que ceux de l'évolution biologique. Pour nous, cela signifie qu'une approche du problème de la conscience dans un machine par l'informatique évolutionnaire semble être une impasse. En effet, il semble que, même en plaçant des individus en tout point identique à des êtres humains mais dépourvu de conscience (ce n'est pas la possibilité d'existence réelle d'une de ces entités qui nous intéressent ici, il s'agit d'une expérience de pensée) dans une simulation informatique simulant parfaitement notre réalité, et en laissant à ces individus un temps long pour évoluer, il n'y a aucune raison particulière de penser que la conscience apparaitra chez ces individus si seul les principes de l'évolution biologique (à savoir, sélection naturelle et reproduction avec dérive génétique) sont *implémentés* dans la simulation.

Une autre approche de la construction d'une machine consciente consiste à tenter d'exprimer les propriétés computationnelles qu'un système devrait posséder pour être conscient. Bien évidemment, cette approche est lourde de conséquence sur la nature de la conscience, qui est réduite, comme le dirait D.Dennett, à un "sac rempli de tours" (notons que Dennett est un partisan de cette idée réductionniste). [5] Cette approche est en quelque sorte une vision cognitiviste réductionniste de la conscience, qui va chercher, comme pour n'importe quelle autre fonction mentale, à la décomposer en des processus simples de traitement de l'information. Une des tentatives les plus récentes de prendre

4. On parle ici bien sûr de l'expérience subjective, de la dîtes conscience *phénoménale*

5. Dans (D. Dennett, 2003), Dennett fait l'analogie entre expliquer la conscience et expliquer un spectacle de magie. La magie, tout comme la conscience, peut nous paraitre mystérieuse, inexplicable ... jusqu'à ce que les différents tours qui composent le spectacle nous soient expliqué. La conscience, pour Dennett, ne serait qu'un "sac rempli de tours" (*bag of tricks*), qu'il conviendrait d'expliquer (car ils sont explicables !) un par un.

cette approche pour déterminer les processus computationnels essentiel à la création d'une machine consciente est décrite dans (Dehaene, Lau, & Kouider, 2017), que nous allons résumer ici. Dans cet article, Dehaene Lau et Kouider émettent l'idée comme quoi les mécanismes de computation implémentés dans les intelligences artificielles actuelles correspondent aux processus inconscients chez l'humain, et proposent qu'une machine ne pourra être consciente que si (il s'agit d'une condition nécessaire) elle est en capacité de réaliser deux types de traitement de l'information :

> la sélection d'informations pour leur diffusion globale, les rendant ainsi accessibles de manière flexible pour le traitement et le rapport (C1, [...]), et l'auto-surveillance [ou l'auto-gestion] de ces traitements de l'information, produisant une sensation subjective de certitude ou d'erreur (C2, [...])

. Ainsi C1, c'est-à-dire l'accessibilité globale des informations sur l'environnement ayant de l'importance dans telle ou telle situation, permet à l'individu d'établir des stratégies efficaces et cohérentes avec son environnement pour atteindre ses buts. L'architecture neuronale implémentant C1 permettrait ainsi de récupérer une information *intéressante*, de la maintenir disponible et de la partager avec les modules du cerveau pertinents pour traiter cette information. Pour Dehaene Lau et Kouider, cette architecture neuronale serait un réseau de neurone globalement reparti à travers le cortex, avec une concentration plus grande sur les zones d'intégration associative de haut niveau, et chaque concept perçu consciemment est encodé par l'activation d'une sous-partie de ce réseau. C2, au contraire de C1, est la capacité de se représenter soi-même et sa propre cognition (ce que l'on appelle parfois métacognition). C2 est ainsi notamment la capacité d'émettre un jugement sur ses propres actions, de déterminer le niveau de confiance que l'on accorde en un choix, d'avoir conscience et d'apprendre de ses de ses erreurs afin de les limiter. Dans l'état actuel des choses, les machines ne possèdent aucune de ces deux capacités. En effet, l'intelligence artificielle se concentre sur des agents spécialisés (des *systèmes experts*) incapables d'effectuer plusieurs actions, et les machines comme les voitures ou les téléphones sont composés d'une collections de modules spécialisés qui ne partagent que très peu d'information avec les autres modules de la même machine. Introduire C1 dans une machine reviendrait alors à permettre à ses différents modules de rendre accessibles des actions aux autres modules, ce qui permettrait une coordination et une collaboration des différents module. Afin de pouvoir prendre une décision globale, la machine devra également posséder une représentation de ses propres capacités et de ses limites, ainsi qu'un moyen d'avoir connaissance des chances de succès de telle ou telle stratégie pour effectuer une tache, c'est-à-dire posséder C2. Cette approche de la conscience dans les machines est-elle suffisante pour instancier une conscience *phénoménale* ou une expérience subjective dans une machine ? Cette question est une question ouverte qui revient en fait à se poser la question de l'existence du "Problème difficile" de la conscience.

Enfin, une dernière idée pour construire une machine consciente se concentre sur l'idée que notre système nerveux est un système physique donnant naissance à la conscience, et de plus, il s'agit du seul système physique dont on ne peut pas douter qu'il le fait. Cette idée est également sous-jacente dans l'approche de Dehaene Lau et Kouider, mais ici, il ne s'agira pas de tenter d'expliquer pourquoi la conscience émerge de ce sys-

tème physique en termes de processus de traitement de l'information (ou en d'autres termes d'ailleurs). Plutôt, l'idée est de reproduire le fonctionnement du système nerveux non pas au niveau des processus de traitement de l'information, mais à un niveau plus fondamental, celui des composantes de base du système nerveux : les cellules (les neurones principalement). Cette approche de modélisation du cerveau pour faire émerger la conscience considère qu'il n'est pas nécessaire de *comprendre* la conscience pour construire une machine consciente. La Nature nous aurait en quelque sorte déjà livré un plan de construction adéquat pour la construction d'un tel type de machine : l'humain. Bien sûr pour l'instant, il existe des limitations techniques qui empêchent ce genre de machine d'exister, mais il est a noté que des simulations courtes de modèles possédant la taille d'un cerveau humain ont déjà été effectuée (voir par exemple (Izhikevich & Edelman, 2008)). Bien sûr implémenter seulement un modèle du cerveau serait insuffisant, et l'interaction de la machine avec son environnement à l'aide de capteurs, d'effecteurs et le rapport de la machine à cet environnement et à son propre *corps* est également très important. En effet, c'est par ses capteurs que la machine peut acquérir des informations sur le monde, et c'est par ses effecteurs que la machine peut émettre des comportements que nous pourrons interpréter comme signe d'une conscience. Mais surtout, en accord avec l'approche de la *cognition située*, le rapport de la machine à son corps et ses interactions avec le monde sont certainement indispensables pour qu'elle développe une conscience. Et si une machine possédait un modèle du cerveau reproduisant parfaitement les propriétés physiques du cerveau, et était capable d'interagir avec le monde de la même manière que nous, mais n'affichait pas de comportement conscient, c'est qu'il faudra chercher la conscience autre part.

Références

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, *18*(2), 227–247.

Bourgine, P., & Varela, F. J. (1992). Towards a practice of autonomous systems. *Varela and Bourgine*, *2332*.

Broad, C. D. (2014). *The mind and its place in nature*. Routledge.

Chalmers, D. (2015). Panpsychism and panprotopsychism. *Consciousness in the physical world : Perspectives on Russellian monism*, *246*.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it ? *Science*, *358*(6362), 486–492.

Dennett, D. (2003). Explaining the" magic" of consciousness. *Journal of Cultural and Evolutionary Psychology*, *1*(1), 7–19.

Dennett, D. C. (2017). *Consciousness explained*. Little, Brown.

Descartes, R., & Gilson, E. (1987). *Discours de la méthode*. Vrin.

Dowe, D. L., & Hajek, A. R. (1997). A computational extension to the turing test. In *Proceedings of the 4th conference of the australasian cognitive science society, university of newcastle, nsw, australia* (Vol. 1).

DUMOUCHEL, P., Bourgine, P., & VARELA, F. J. (1989). *Autonomie et connaissance*. Seuil.

Izhikevich, E. M., & Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the national academy of sciences*, *105*(9), 3593–3598.

Johnston, J. (2008). *The allure of machinic life : cybernetics, artificial life, and the new ai*. MIT Press.

Kari, J. (2005). Theory of cellular automata : A survey. *Theoretical computer science*, *334*(1-3), 3–33.

Langton, C. G. (1984). Self-reproduction in cellular automata. *Physica D : Nonlinear Phenomena*, *10*(1-2), 135–144.

Legg, S., & Hutter, M. (2007). Universal intelligence : A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444.

Lewes, G. H. (1877). *Problems of life and mind*. Trübner & Company.

Nagel, T. (1974). What is it like to be a bat ? *The philosophical review*, *83*(4), 435–450.

Putnam, H. (1980). The nature of mental states. *Readings in philosophy of psychology*, *1*, 223–231.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv :1511.06434*.

Ray, T. (1993). How i created life in a virtual universe. *Online at< http ://www. his. atr. jp/˜ ray/pubs/nathist/>. Accessed*, *8*, 2007.

Rosenthal, D. (2012). *Does consciousness have any utility ?* Consulté sur `https://www.youtube.com/watch?v=J8ctBHyJ3Gg` (Turing Consciousness 2012 Summer School)

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence : a modern approach*. Malaysia ; Pearson Education Limited,.

Salzberg, C., Antony, A., & Sayama, H. (2004). Complex genetic evolution of self-replicating loops. In *Ninth international conference on artificial life* (pp. 262–267).

Sayama, H. (1998). Constructing evolutionary systems on a simple deterministic cellular automata space. *Phd, University of Tokyo, Department of Information Science*.

Sayama, H. (2004). Self-protection and diversity in self-replicating cellular automata. *Artificial Life*, *10*(1), 83–98.

Searle, J. R. (2008). *Mind, language and society : Philosophy in the real world*. Basic books.

Varela, F. G., Maturana, H. R., & Uribe, R. (1974). Autopoiesis : the organization of living systems, its characterization and a model. *Biosystems*, *5*(4), 187–196.

Von Neumann, J., & Burks, A. W. (1996). *Theory of self-reproducing automata*. University of Illinois Press Urbana.