**Philosophiques**

# Projet et signification dans des réseaux d'automates : le rôle de la sophistication

Henri Atlan

Volume 20, Number 2, Fall 1993

Perspectives sur la phénoménologie et l'intentionnalité

URI: https://id.erudit.org/iderudit/027235ar
DOI: https://doi.org/10.7202/027235ar

See table of contents

Publisher(s)

Société de philosophie du Québec

ISSN

0316-2923 (print)
1492-1391 (digital)

Explore this journal

Cite this article

Atlan, H. (1993). Projet et signification dans des réseaux d'automates : le rôle de la sophistication. *Philosophiques*, *20*(2), 443–472. https://doi.org/10.7202/027235ar

PHILOSOPHIQUES, VOL. XX, NUMÉRO 2, AUTOMNE 1993, P. 443-472

# LA PHÉNOMÉNOLOGIE ET LA NATURE DE L'INTENTIONNALITÉ

# PROJET ET SIGNIFICATION* DANS DES RÉSEAUX D'AUTOMATES : LE RÔLE DE LA SOPHISTICATION

par

**Henri Atlan**

## I. Introduction : pour une théorie physique de l'intentionnalité

On a l'habitude de distinguer l'intentionnalité au sens usuel de recherche d'un but et l'intentionnalité au sens technique de source des significations de la pensée en psychologie, reprise ensuite par la phénoménologie comme caractéristique des activités de conscience donatrice de sens, puis par l'analyse psycholinguistique, comme source des significations du langage.

Mais il existe pourtant une relation étroite entre ces deux sens, usuel et technique, et cette relation fonctionne dans les deux sens.

Il est classique d'observer que l'intentionnalité comme recherche d'un but est un cas particulier de l'intentionnalité en général, comme activité orientée de la conscience, conscience de quelque chose, origine des significations.

---

* Ce texte a fait l'objet d'une communication dans le cadre du colloque « L'intentionnalité en question » qui a lieu à Nice en juin 1992.

Mais l'expérience des machines programmées montre que cette relation fonctionne aussi en direction opposée : c'est la définition du but qui détermine les significations.

Les parties d'une machine n'ont de signification que par la fonction qu'elles assurent dans le fonctionnement global de la machine. Et dans le cas des machines artificielles ce fonctionnement global est finalisé par une certaine tâche à accomplir, un certain problème à résoudre.

Dans un programme d'ordinateur, les instructions n'ont de signification que par rapport à leur rôle et à leur efficacité dans la solution du problème que le programme doit trouver, ou en général dans la tâche qu'il doit accomplir. C'est donc le but – qui dans ce cas est fixé par le programmeur ou le concepteur de la machine – qui détermine les significations des éléments du programme et de la machine.

Mais est-il possible de concevoir une machine dont le fonctionnement finalisé, la tâche à accomplir, ne serait pas imposé de l'extérieur par un programmeur, et serait donc produit par la machine elle-même ?

C'est une façon opérationnelle, probablement réductrice mais nous allons revenir sur ce point, de poser la question : une machine peut-elle être intentionnelle ?

Comme nous le savons, il existe une réponse évidente mais triviale à cette question : oui, une machine peut être intentionnelle puisque nous-mêmes avec notre cerveau, sommes de telles machines et que nous faisons l'expérience de notre intentionnalité comme sources de significations et causes d'action.

Mais cette réponse est insuffisante car nous ne faisons pas d'expérience d'intentionnalité dans n'importe quelle machine, certainement pas dans la plupart, sinon la totalité, des machines fabriquées par des hommes; et quant aux machines naturelles il serait difficile, en dehors d'une attitude proprement animiste, d'attribuer une intentionnalité à n'importe quelle sorte de mécanisme, par exemple à une étoile ou à un tourbillon liquide, ou même à une cellule vivante.

Alors la question en se précisant, devient : quelles sortes de machines sont capables d'intentionnalité ? Jusque là, il semblerait que la seule réponse soit circulaire, à savoir : des machines dont la structure biochimique serait telle qu'elles auraient pu évoluer jusqu'à produire des cerveaux comme les nôtres. Et bien sûr, ceci ne nous dit pas grand-chose tant que nous ne savons pas

*comment* nos cerveaux produisent, si l'on peut dire, une intentionnalité, ou en sont le siège.

Pourtant, on peut faire un pas en avant en observant qu'une propriété essentielle de l'organisation biologique est sa nature dite hiérarchique, c'est-à-dire en niveaux d'intégration différents. Et beaucoup de difficultés de la biologie, mais aussi la richesse de comportements qui semblent propres aux systèmes vivants[1], concernent la nature des articulations entre niveaux.

Alors notre question peut être reformulée maintenant d'une autre facon, à partir des propriétés dynamiques de réseaux d'automates en interaction. En effet de tels réseaux sont des simulations simples mais assez bonne d'organisations en niveaux où les articulations entre propriétés locales et globales peuvent être étudiées. La question devient alors :

Est-ce qu'une intentionnalité même limitée peut émerger dans un processus auto-organisateur comme propriété globale d'un réseau d'automates ?

Les automates dont il s'agit ici ne sont pas des machines compliquées douées de facultés d'autonomie et de comportements simulant ceux d'organismes vivants – ce ne sont pas des robots intelligents peu importe ce qu'on entend par là – mais des unités matérielles obéissant à des lois relativement simples. Leur association en réseaux présente des propriétés intéressantes qu'on peut étudier dans des systèmes physiques tels que, par exemple, des verres de spins, et qu'on peut généraliser à ce qu'on appelle des réseaux de neurones formels où chaque automate est une unité simulant de façon très simplifiée certaines des propriétés à la fois physiques et logiques des neurones.

Aux yeux de philosophes, au moins de certains d'entre eux, phénoménologues en particulier, il y a peut-être quelque outrecuidance à vouloir faire une théorie physique, forcément réductrice de phénomènes considérés traditionnellement comme le propre de la conscience, tels que la signification et l'intentionnalité. Mais n'oublions pas qu'il semble que même pour Husserl l'emprise du vitalisme était telle qu'il semblait qu'une biologie physique serait impossible alors que c'est pourtant une telle biologie que nous connaissons aujourd'hui.

---

1. H. Atlan, *À tort et à raison. Intercritique de la science et du mythe*, Paris, Seuil, 1986, chap. II.

## II. Réductionnisme physico-chimique et réduction phénoménologique

C'est donc intentionnellement, si je puis dire, dans le contexte matérialiste d'une biologie physico-chimique que je veux me placer. Et ce contexte est forcément réductionniste, au moins dans sa méthode, en ce sens qu'on y met entre parenthèses les phénomènes de conscience tels que nous en faisons par ailleurs l'expérience immédiate, ou encore l'expérience médiatisée par la méditation philosophique. En conséquence, contrairement à l'analyse idéaliste phénoménologique, il n'est pas question de poser l'intentionnalité au départ comme fait de conscience (ou de langage) fondateur, mais au contraire de se situer dans ce que Husserl appelle l'attitude naturaliste, en assumant d'ailleurs cette attitude et en revendiquant pour cette attitude qui est celle des sciences de la nature, une incomplétude certes mais qui n'est pas plus grande, ou plus profonde ou plus radicale que celle de l'attitude phénoménologique. En cela évidemment nous nous séparons de la philosophie phénoménologique qui croyait pouvoir fournir des fondements absolus aux sciences de la nature à partir de l'expérience fondatrice du sujet transcendantal[2].

En d'autres termes, d'un point de vue méthodologique, le réductionnisme des sciences de la nature, forcément matérialiste et physico-chimique, est symétrique du réductionnisme de la conscience que revendique la phénoménologie comme premier moment de sa démarche, de la fameuse réduction phénoménologique (l'*époché*) par laquelle le sujet méditant et appliquant le doute méthodique à la Descartes commence par *mettre entre parenthèses* tout ce qui n'est pas son expérience vécue de ses propres courants de conscience. Bien sûr cette mise entre parenthèses est parfaitement justifiée par la méthode phénoménologique. Mais les sciences de la nature effectuent une réduction symétrique, par laquelle *c'est la conscience qui est mise entre parenthèses*, et cette réduction est tout aussi justifiée, d'un point de vue de méthode, en ce qu'elle promet, comme la réduction précédente, de retrouver plus tard ce qu'elle a mis entre parenthèses. Mais bien sûr ni l'une ni l'autre pour l'instant ne tient sa promesse.

Cette réduction naturaliste matérialiste est la symétrique inversée de la réduction phénoménologique idéaliste, car

2. E. Husserl, *Idées directrices pour une phénoménologie*, tr. fr., P. Ricoeur, Paris, Gallimard, 1950 ; et *Méditations cartésiennes*, tr. fr. G. Pfeiffer et É. Lévinas, Paris, Vrin, 1947.

contrairement à ce que supposait Husserl, il n'existe pas d'asymétrie fondamentale telle que l'intentionnalité de la conscience ainsi réduite serait *donatrice* de sens absolument, y compris, en fin de compte au monde des sciences de la nature, tandis que les savoirs scientifiques ne feraient que recevoir leur sens, et par là leurs fondements absolus, même sans le savoir, des activités de la conscience que révèlerait l'analyse phénoménologique.

Cette thèse que la réduction phénoménologique n'est pas plus fondatrice, absolument, que ne le serait la réduction physicaliste, nous pouvons déjà nous en persuader en observant que le programme des *Méditations cartésiennes* de Husserl n'a pas pu être tenu : l'analyse phénoménologique ne peut pas plus servir de fondement absolu aux sciences de la nature que la philosophie de Descartes, et pour les mêmes raisons. Les sciences de la nature se développent concrètement suivant des critères de vérité et d'efficacité locaux qui les dispensent de fondement absolu. Les changements de théories et de paradigmes laissent derrière eux les philosophies qui avaient cru servir de fondements *a priori* et éternels, et qui apparaissent alors comme des reconstructions *a posteriori*.

Pour trouver une philosophie capable de penser cette symétrie entre réduction physico-chimique de la conscience et réduction phénoménologique de la physico-chimie, c'est probablement vers Spinoza qu'il faudrait se tourner. La position originale sur les rapports du corps et de l'esprit, non dualiste mais cependant ni idéaliste ni matérialiste, qu'on a appelée en simplifiant le parallélisme de Spinoza pourrait être utile à condition de bien intégrer en quoi le monisme du corps et de l'esprit exclut pourtant une causalité de l'esprit sur le corps tout comme il exclut une causalité du corps sur l'esprit, ceux-ci ne pouvant être perçus et conçus que comme deux aspects d'une même substance, d'une même réalité qui elle, ne peut être connue et décrite comme enchaînement de causes que dans l'un *ou* l'autre de ces aspects.

Quoi qu'il en soit, la mise entre parenthèses de la conscience est évidemment le pari que font les sciences cognitives surtout depuis ces dernières années, depuis l'impulsion qu'elles ont reçue de la physique des systèmes complexes et désordonnés et aussi des théories connexionnistes en psychologie et en intelligence artificielle[3].

3. F. Fogelman-Soulié (dir.), *Les théories de la complexité*, Paris, Seuil, 1991.

## III. Complexité sans signification dans les sciences de l'information

Mais dans ce pari physicaliste une place de choix est attribuée à l'ordinateur, et c'est normal étant donné qu'on dispose là d'une machine dont on peut observer les comportements et étudier en quoi ils sont intentionnels au sens cette fois de comportement d'agents rationnels adaptant leurs moyens à leurs fins, celles-ci étant en général la solution de problèmes logico-mathématiques.

Ceci est à l'origine du rôle de l'ordinateur et de son programme comme paradigme dominant en psychologie cognitiviste — « le cerveau comme ordinateur » est devenu maintenant, grâce aux connexionnistes, « ordinateur parallèle » —, et aussi en biologie où « le développement embryonnaire est comparé à l'exécution d'un programme ».

Mais cette métaphore de l'ordinateur, si elle présente des avantages certains, en ce qu'elle propose un modèle de machines intentionnelles — ou plutôt intentionnalisées, c'est cela toute la question qui va nous occuper — présente un inconvénient majeur : elle présuppose la solution du problème qu'il s'agit de résoudre, c'est-à-dire l'origine des significations sans lesquelles aucune « computation » ne peut avoir de sens et ne peut donc réaliser un comportement intentionnel. Une telle computation sans signification n'est en elle-même qu'une expression syntactiquement correcte comme une formule algébrique, mais qui ne dit rien sur le monde physique. Elle ne peut donc dans le meilleur des cas que *simuler* un comportement réel et encore à la condition d'être *interprétée*, c'est-à-dire à condition qu'un agent intentionnel extérieur, un être humain « théorisateur », y projette des significations.

Supposer la solution d'un problème qu'il s'agit de résoudre n'est pas forcément un défaut ; cela peut être une méthode féconde, à condition que ce soit une étape provisoire, et qu'on n'oublie pas en cours de route qu'il ne suffit pas de poser que la solution existe pour qu'elle existe vraiment, *a fortiori* pour qu'on ait la moindre idée de ce qu'elle peut être.

Autrement dit, nous restons toujours confrontés à la question débattue entre partisans de l'Intelligence Artificelle (IA) forte et faible, et les réseaux connexionnistes n'apportent rien de fondamentalement nouveau là-dessus par rapport aux ordinateurs séquentiels classiques. Cette question peut être formulée de la façon suivante : quand un ordinateur calcule (« compute ») un comportement intelligent, est-ce qu'il comprend

ce qu'il fait ? On sait qu'à cette question, certains répondent oui, c'est l'IA forte; tandis que d'autres répondent non, c'est l'IA faible, pour qui l'ordinateur ne fait que simuler la compréhension sans comprendre lui-même, de la même façon qu'un ordinateur qui compute le vol d'un avion ne vole pas, la digestion de l'estomac, ne digère pas, la respiration des poumons, ne respire pas, etc... À cela les partisans de l'IA forte répliquent que lorsque l'ordinateur calcule il ne fait pas que *simuler* le calcul, il calcule effectivement. Et la compréhension n'étant pas autre chose qu'un calcul, une « computation », lorsqu'il « compute » il comprend.

On voit que cette controverse concerne en fait notre compréhension de ce qu'est la compréhension et qu'elle ne pourra être tranchée que si nous pouvons savoir en vérité ce qu'est pour nous comprendre.

Si nous pensons que comprendre, pour nous, c'est « computer », calculer d'une certaine façon, alors en effet l'ordinateur calcule vraiment et donc ne fait pas que simuler la compréhension mais comprend vraiment. Par contre si nous pensons que comprendre, pour nous, implique non seulement du calcul mais aussi par exemple notre digestion, notre respiration, c'est-à-dire au minimum le métabolisme de nos neurones, alors l'ordinateur ne fait que simuler certains résultats de ce métabolisme, comme il simule la respiration sans respirer.

Il faut donc essayer de s'attaquer à cette question de l'origine des significations dans le comportement de différentes sortes de machines intentionnelles, depuis les programmes d'ordinateurs les plus simples – dont on a l'habitude de dire qu'ils sont stupides mêmes s'ils sont très efficaces parce que de toute évidence la seule intentionnalité qui y soit à l'oeuvre est celle de leur programmeur – jusqu'à ceux, beaucoup plus complexes, qui simulent des comportements intelligents, à tel point que la question de leur intelligence et donc de leur intentionnalité peut au moins se poser.

Mais dans cette démarche nous sommes handicapés doublement. D'abord parce que cette question jusqu'à il y a peu de temps n'a pas beaucoup intéressé les informaticiens, spécialistes d'ordinateurs et de programmation. Et c'est bien compréhensible puisque ceux-ci ne s'occupent que de programmes écrits par des agents intentionnels et rationnels dont on suppose *a priori* qu'ils ne sont pas absurdes, ni dans le but qu'ils se fixent ni dans la façon d'y arriver. Autrement dit, la question de la signification des programmes a été laissée de côté par la théorie de la programma-

tion, comme celle de la signification des messages a été laissée de côté par la théorie des communications. Mais nous avons un deuxième handicap qui vient de la biologie, ou plus exactement de ses métaphores informatiques, et notamment celle du programme génétique qui elle aussi suppose le problème des significations résolu, sauf qu'ici il n'y a pas de programmeur.

Il faut donc que nous essayions de remonter à la source pour savoir où et comment s'effectue cet oubli de la question des significations tant en informatique que dans les transpositions informatiques à la biologie. Il faut essayer de savoir, comme dirait peut-être Daniel Dennett, comment s'effectue « l'emprunt sur le capital intentionnel » ou « sur l'intelligence » (« *intelligence loan* »). Cet emprunt, c'est par exemple le degré d'intelligence et d'intentionnalité qu'on prête à un programme de jeux d'échecs, ou à un organisme vivant, pour pouvoir décrire son comportement comme s'il était la conséquence d'une intentionnalité et d'une intelligence, car ce serait trop compliqué, ou même pratiquement impossible de l'expliquer exclusivement par ses déterminations mécaniques, physiques et fonctionnelles[4]. Mais comme il le dit si bien, une fois cet emprunt effectué il ne faut pas oublier de « payer sa dette », c'est-à-dire d'expliquer finalement de façon causale les mécanismes de ce comportement initialement supposé intentionnel et intelligent.

Pour commencer je voudrais montrer rapidement comment les métaphores informatiques en biologie, celle du programme génétique en particulier, souffrent de ce que la question de la signification de l'information génétique n'est en général pas posée. Cela n'est que la conséquence de ce qu'on a pris l'habitude de procéder ainsi – c'est-à-dire de supposer résolue l'origine des significations, « d'emprunter sur l'intentionnalité » et de ne plus s'en préoccuper – dans les sciences informatiques et de la programmation. Mais si cela peut se justifier des programmes d'ordinateurs écrits dans un but particulier, cela ne peut plus l'être de machines naturelles fabriquées sans but évident et sans projet explicite par le seul effet de la sélection naturelle.

4. D.C. Dennett, *Brainstorms*, Montgomery, Vermont, Bradford Books, 1978, p. 15-16.

## IV. Des sources de signification dans l'organisation biologique. Limites de la métaphore du programme génétique

L'idée d'un programme écrit dans les gènes sous la forme des séquences nucléotidiques des ADN provient schématiquement des observations suivantes et de leur utilisation implicite dans un raisonnement fallacieux.

1) L'ADN est une séquence quaternaire facilement réductible à une séquence binaire. 2) Tout programme d'ordinateur séquentiel déterministe est réductible à une séquence binaire. 3) *Ergo* : les déterminations génétiques produites par la structure des ADN fonctionnent à la manière d'un programme séquentiel écrit dans l'ADN des gènes. La fallace implicite dans ce raisonnement est évidemment la réciproque de 2, à savoir « toute séquence binaire est un programme ».

Cette fallace correspond à un besoin théorique dont on peut analyser l'origine dans l'histoire du concept de génétique et du contenu inattendu que lui a donné la biologie moléculaire.

Les gènes étaient autrefois définis de façon formelle sans qu'on en connaisse la nature, à partir de l'observation de transmissions héréditaires de caractères, qui constituent en effet des processus génétiques au sens traditionnel du terme, c'est-à-dire des processus de transformation et de production, de genèse des organismes à partir de ce qui les produit, comme un effet est produit par sa cause où, comme disait Aristote « le père est cause de l'enfant [...] ce qui produit le changement (est cause) de ce qui est changé »[5].

Or aujourd'hui, connaissant la structure physique des gènes, nous sommes dans une situation curieuse et apparemment paradoxale si l'on ne s'en tient qu'aux mots du langage, où nous devons admettre que *le processus génétique ne se trouve pas dans le gène.* Le paradoxe n'est qu'apparent dès qu'on réalise que le gène n'est pas un processus puisque c'est une molécule. La structure moléculaire statique du gène joue certes un rôle déterminant mais comme élément d'un processus de production qui implique par ailleurs d'autres molécules et surtout un ensemble de réactions, de transformations physiques et chimiques, entre ces molécules. Le rapport traditionnel entre structure et fonction a changé de

5. *Physique* II, 3.

nature. Une structure non vivante, celle d'une molécule, est responsable de fonctions qu'on percevait autrefois comme des fonctions vitales. On a bien du mal aujourd'hui à se débarrasser de la connotation vitaliste attachée à la notion même de fonction alors qu'on envisage pourtant le rôle de structures moléculaires. En un mot, le gène n'est pas vivant et il est encore censé expliquer la vie. À tel point que beaucoup de non-spécialistes ne peuvent pas se résoudre à admettre qu'un gène n'est qu'une molécule.

Si l'on admet, comme on le dit trop vite, que le génome, ensemble des gènes, contient le secret de la vie, et qu'en conséquence la découverte de chaque gène dévoile un peu plus de ce secret, alors il est en effet bien difficile de comprendre en même temps qu'un gène est une molécule. Car si un gène contient une partie des secrets de la vie, comment pourrait-il n'être qu'une molécule, c'est-à-dire un morceau de matière non vivante ?

Mais une fois la nature et l'origine de cette fallace reconnues, nous devons nous demander quelles sont les autres possibilités pour une séquence binaire outre celle d'être un programme.

Une première possibilité serait qu'il s'agisse d'une séquence aléatoire. Si par là on entend une séquence sans signification c'est bien difficile à accepter car on comprendrait mal alors comment de telles séquences pourraient déterminer des fonctions biologiques qui constituent, dans la métaphore informatique, la signification de l'information génétique. Si par séquence aléatoire on entend une séquence incompressible par aucun algorithme plus court qu'elle-même – suivant la théorie classique de la complexité algorithmique – rien n'empêche alors que l'on retrouve l'éventualité précédente d'un programme d'ordinateur. On sait en effet que la séquence binaire qui traduit un programme d'ordinateur peut être elle-même une séquence aléatoire dans ce sens. Il s'agit là d'ailleurs d'une des insuffisances de cette théorie de la complexité algorithmique qui présuppose l'existence de significations dans une telle séquence – celles que réalise l'exécution du programme – sans avoir à en rendre compte de façon explicite dans la structure de la séquence. Nous reviendrons sur ce problème qui vient de ce que la théorie ne concerne qu'une complexité mesurée en unités de calcul sans tenir compte de la signification éventuelle de ce que le calcul effectue.

Mais il existe une autre possibilité qui va nous retenir maintenant : que la séquence binaire ne soit ni programme, ni sans signification mais qu'elle constitue des *données*.

Pour envisager cette possibilité il faut d'abord justifier la distinction programme – données que tout l'effort de l'informatique théorique a eu pour résultat de supprimer[6]. Il faudra ensuite préciser par quelle sorte de programme ces données seraient traitées.

## V. La « sophistication » comme mesure de complexité porteuse de signification

On sait que dans le cadre des machines de Turing universelles une séquence binaire peut être indifféremment traitée comme un programme ou comme des données[7]. La machine de Turing lit la séquence et l'interprète ainsi comme une description binaire d'un objet à fabriquer ou d'une tâche à réaliser, dans laquelle les parties programme et données sont indiscernables et interchangeables. Mais cet état de choses résulte de ce que la théorie s'occupe d'artefacts – objets ou machines fonctionnelles – dont la signification n'est qu'implicite sous la forme du but ou de la tâche assigné par le programmeur et n'est jamais prise en compte explicitement par la théorie. C'est précisément pour cela que, de façon apparemment paradoxale, la complexité algorithmique maximum est réalisée par une suite aléatoire. Ce qui semble être un défaut de la théorie n'en est pas un quand il s'agit d'artefacts, c'est-à-dire d'algorithmes dont on sait *par ailleurs* qu'ils ont une signification, celle que le programmeur leur a assignée sous la forme d'une tâche à accomplir.

Au contraire, se contenter d'une mesure de complexité sans signification est une insuffisance de la théorie quand il s'agit de description d'objets naturels que nous observons sans savoir dans quel but ils sont formés. Là, il faut tenir compte du contenu sémantique plus ou moins important de ces objets, à définir de façon telle qu'une suite qui ne serait qu'aléatoire devrait avoir une complexité porteuse de signification nulle. Et pour formaliser une telle complexité porteuse de signification il nous faut retenir et préciser la distinction entre les parties programme et données d'une description. *C'est la partie programme qui explicite un but, une finalité source de significations.* C'est elle qui définit une *classe d'objets*

6. A.V. Aho, J.E. Hopcroft, J.D. Ullman, *The Design and Analysis of Computer Algorithms*, Reading Mass., Addison Wesley, 1974.

7. *Ibid.*; cf. aussi H. Atlan, *L'organisation biologique et la théorie de l'information*, Paris, Hermann, 1972, 2e éd., 1992.

*partageant une même structure.* Au contraire, *les données spécifient un objet particulier dans cette classe.*

Un exemple très simple permet de fixer les idées. Supposons un objet décrit par la séquence suivante : 0011110000001100 11. Cette séquence est produite en dédoublant chaque digit de la séquence : 0110001101. On y distingue donc facilement une partie programme qui consiste à dédoubler chaque digit dans n'importe quelle suite, et une partie données qui est la séquence sur laquelle le programme est appliqué. Le programme définit une classe infinie d'objets qui partagent tous la structure en digits dédoublés. Les données spécifient dans cette classe un objet particulier.

La définition classique de la complexité algorithmique d'un objet peut ainsi être modifiée de façon à tenir compte d'une *mesure de complexité avec signification.* Rappelons que la complexité algorithmique classique d'un objet décrit par une séquence S est la longueur H(S) d'une description minimale, faite d'un programme et de données, telle que cette description entrée sous forme de données dans une machine de Turing suffit pour générer S.

H(S) = min { |P| + |D| | (P, D) génèrent S en étant entrées dans une machine de Turing}.

La description minimale (P,D) contient une partie programme P et une partie données D, de longueurs respectives |P| et |D| mais l'important est la longueur minimale totale sans qu'il soit nécessaire de distinguer de façon absolue et invariante suivant la machine de Turing considérée ce qui est programme et ce qui est données.

Avec M. Koppel[8], nous avons défini une mesure de complexité avec signification, que nous appelons « sophistication » comme la seule longueur |P| de la partie programme de la description minimale. Une des conséquences de cette définition corrige le défaut de la théorie en ce qu'une longue suite aléatoire a classiquement une grande complexité mais une sophistication quasi nulle. En effet, pour la reproduire telle quelle, sa description minimale contient une partie programme qui se réduit à l'instruction « PRINT » et une partie données qui n'est pas autre chose que la suite elle-même.

---

8. M. Koppel et H. Atlan, « Les gènes : programme ou données ? Le rôle de la signification dans les mesures de complexité » dans F. Fogelman-Soulié (dir.), *op. cit.*, p. 188-204.

Ce n'est pas ici le lieu d'entrer dans les détails techniques exposés ailleurs[9], permettant une généralisation de cette distinction telle que la séparation entre programme et données (avec la mesure de sophistication qui en suit) soit aussi invariante par rapport à la machine de Turing considérée que l'est la mesure classique de complexité, c'est-à-dire invariante à une constante additive près[10].

Autrement dit on se donne une constante c ajoutée à la longueur de la vraie description minimale d'un objet particulier, pour éviter que cette description ne soit minimale que pour cet objet particulier et non la classe des objets partageant la même structure. On cherche ainsi à éviter que la description minimale ne soit la plus courte qu'à cause de propriétés particulières de l'objet envisagé, par exemple dans le cas où la séquence S est elle-même courte. Il peut être alors plus économique de la décrire en ignorant sa vraie structure, par un programme *ad hoc* qui contient une part de données. À la limite il peut être plus court de la reproduire telle quelle par « PRINT » comme s'il s'agissait d'une suite aléatoire.

Par exemple supposons que S soit la séquence ABCABCABCABC. Comme ABC n'est répété que quatre fois, il est plus économique pour produire S d'utiliser la description « PRINT ABCABCABCABC » comme si la séquence S était aléatoire et n'avait pas de structure. Dans ce cas, la sophistication serait la longueur de PRINT qui n'exprime évidemment pas la structure générale de la classe des objets suggérés par ABC répété n fois.

Par contre, si S est constituée par ABC répété un grand nombre de fois, alors il deviendra plus économique de la décrire par un programme « PRINT AND REPEAT N TIMES » avec « ABC » comme données. Cette dernière description est plus longue que la précédente pour un objet S où n est petit. Ainsi, quand S est courte la vraie description minimale peut être « PRINT S ». En ajoutant une constante c à la description minimale on se donne une marge qui

---

9. M. Koppel, « Structure », dans R. Herken (dir.), *The Universal Turing Machine. A Half-Century Survey*, Oxford University Press, Londres, 1988, p. 435-452; et M. Koppel et H. Atlan, « An Almost Machine-Independant Theory of Program Length Complexity. Sophistication and Induction », *Information Sciences*, 56, 1991, p. 23-33.

10. Pour cela on définit une description dite c-minimale (P,D) de S telle que :

$$|P| + |D| < H(S) + c$$

et la c-sophistication de S :

$$SOPH_c(S) = \min \{ |P| \, D, (P,D) \text{ est une description c-minimale de S} \}.$$

permet d'utiliser comme description minimale une description, certes plus longue dans ce cas-là, mais qui devient plus courte dès que n devient grand, c'est-à-dire dès que l'objet particulier à réaliser dans la classe devient plus long. Ainsi, c mesure le caractère non *ad hoc* de la description minimale retenue, en sorte que sa partie programme exprime véritablement la structure de la classe et non un cas particulier dans cette classe. Un bon « vrai » programme sera toujours la partie programme d'une description c-minimale même pour de grandes valeurs de c (c'est-à-dire pour de longs objets), contrairement à « PRINT » qui n'est la partie programme d'une description minimale que pour des objets courts (à moins évidemment qu'il ne s'agisse d'objets sans structure, de séquences aléatoires).

On peut montrer que dans le cas d'une suite S infinie, c se réduit à zéro.

## VI. ADN : programme ou données

Il nous suffira de noter que cette distinction permet de poser la question du rôle des séquences nucléotidiques des ADN dans les déterminations génétiques sous la forme d'une alternative entre un rôle de programme et un rôle de données. Cette alternative permet alors de poser la question du rôle de la machinerie cellulaire toujours associée aux ADN dans la production de ces déterminations. Si les ADN sont un programme, cette machinerie joue le rôle d'un interprêteur de programme. Si les ADN sont des données, la machinerie cellulaire joue le rôle de programme traitant ces données. Il s'agit là bien entendu de deux métaphores complémentaires dont nous avons discuté par ailleurs[11] les mérites respectifs quant à leur pouvoir explicateur de certains mécanismes connus de différenciation cellulaire et de l'évolution.

Outre ses mérites propres, la deuxième métaphore a l'avantage de provoquer la discussion et de mettre en question la première métaphore. Car il est clair que la notion de programme, inscrit dans la séquence nucléotidique des gènes à la façon d'un programme d'ordinateur, a eu en son temps une valeur heuristique et opérationnelle indéniable; ne serait-ce que pour se représenter de façon globale un mode d'action des déterminations

11. H. Atlan et M. Koppel, « The Cellular Computer DNA : Programm or Data ? *Bull. Mathem. Biol.*, 52, 3 1990 p. 335; voir aussi *art. cit.* note 7.

génétiques sur le développement des organismes apparemment dirigé vers leur réalisation future. Mais comme il arrive souvent, cette notion utilisée de façon non critique, en oubliant qu'il s'agit d'une métaphore et non d'une connaissance explicite de mécanismes bien identifiés, peut devenir un écran et empêcher de progresser dans la recherche de ces mécanismes.

C'est en cela qu'une métaphore alternative peut être utile. Dans celle que nous proposons, les déterminations génétiques résultant de la structure séquentielle des ADN fonctionnent non pas comme un programme mais comme des données mémorisées, traitées et utilisées dans un processus dynamique qui, lui, joue le rôle d'un programme. Ce processus est produit par l'ensemble de nombreuses réactions biochimiques couplées du métabolisme cellulaire. Un tel processus dynamique est comparable à celui d'un réseau d'automates, d'une machine d'états dont les travaux actuels en Intelligence Artificielle montrent qu'elle est capable d'adaptation, d'apprentissage non programmé et, de façon générale, d'auto-organisation structurale et fonctionnelle.

L'avantage de cette nouvelle métaphore est d'indiquer un déplacement de centre d'intérêt. D'une attention portée presque exclusivement au « tout génétique », on passe à la recherche de processus épigénétiques et à l'analyse des mécanismes régulateurs de l'expression génique.

C'est ainsi que Walter Gehring[12], après avoir découvert la fonction régulatrice de certaines séquences du génome de drosophiles (dites *homeobox*) posait la question : « Comment les régulateurs sont-ils régulés ? ». Et il suggérait de chercher une réponse dans le cytoplasme et son information de position. À propos de cette nouvelle métaphore, on évitera bien sûr autant que possible de tomber dans le même piège qui consisterait à la prendre à son tour trop au sérieux.

Car il est probable que la réalité doit se trouver quelque part entre les deux métaphores. Entre la vision d'un programme d'ordinateur inscrit dans les séquences nucléotidiques des gènes, et celle de données mémorisées, traitées par un réseau de réactions métaboliques, comme dans un programme distribué, la réalité doit se trouver quelque part entre les deux puisqu'on ne peut pas

12. W. G. Gehring, « The Molecular Basis of Development », *Scientific American*, Oct. 1985, p. 153-162.

nier que la structure des gènes détermine en retour, bien qu'à une échelle de temps plus longue, la structure du réseau métabolique.

C'est ainsi que derrière ces deux métaphores on peut concevoir l'image d'un réseau évolutif (nouvelle métaphore ?) où deux dynamiques seraient superposées, à des échelles temporelles différentes. Une dynamique du premier ordre dépendrait de la structure du réseau métabolique et des données qu'il reçoit sous la forme de gènes actifs exprimant des protéines qui servent de substrats ou de catalyseurs aux réactions du métabolisme. Mais par une dynamique du deuxième ordre, *plus lente*, des états stables du réseau modifieraient l'activité de certains gènes, en sorte que certaines réactions du métabolisme dépendant de cette activité s'arrêteraient et d'autres seraient déclenchées, produisant ainsi une modification de la structure du réseau de réactions. Celui-ci serait ainsi conduit par une nouvelle dynamique du premier ordre vers un nouvel état stable, et ainsi de suite ...

Mais cette question des ADN comme programme ou données n'a été discutée ici que pour introduire et illustrer la notion de sophistication, à propos d'un exemple d'organisation naturelle.

Cette notion est voisine de celle de « *logical depth* » proposée par Bennett[13] comme mesure du temps qu'il aurait fallu à l'évolution pour produire des algorithmes de plus en plus complexes et performants du point de vue de leur contenu sémantique.

Nous en aurons besoin plus loin, en considérant la possibilité d'objets à sophistication infinie, quand nous essaierons de caractériser formellement les propriétés d'une auto-organisation intentionnelle.

## VII. Typologie des auto-organisations

Mais auparavant, pour situer le problème, et revenir à la question plus générale de l'intentionnalité des machines, je voudrais d'abord proposer une distinction entre différentes sortes d'auto-organisation qu'on rencontre dans la littérature.

Nous distinguerons d'abord entre deux sortes d'auto-organisation, l'une dite « au sens faible », et l'autre « au sens fort ». Ensuite, nous distinguerons, à l'intérieur de la deuxième sorte, entre systèmes auto-organisateurs non intentionnels et

13. C.H. Bennett, « Logical Depth and Physical Complexity », dans R. Herken, *op. cit.*, p. 227-258.

intentionnels, ces derniers n'étant pleinement réalisés que dans l'espèce humaine, même si l'on peut en trouver des formes simplifiées dans des espèces animales voisines.

## A) Auto-organisation au sens faible

Des exemples d'auto-organisation *au sens faible* sont fournis par la plupart des applications du calcul par réseaux neuronaux aux techniques d'intelligence artificielle destinées à fabriquer des machines à apprendre et à mémoire distributive[14]. Ce calcul est effectué en parallèle par un grand nombre d'unités, dites neurones formels, qui effectuent chacune des opérations élémentaires simulant de façon très simplifiée le fonctionnement électrique de neurones, d'où leur nom. Mais ces unités en grand nombre – au moins plusieurs milliers – sont interconnectées dans des « architectures » particulières de réseaux qui acquièrent de ce fait des propriétés d'apprentissage non programmé et de mémoire associative. Ils sont capables, par exemple, de compléter et de corriger des formes imparfaites au moment de leur rappel en mémoire, reproduisant ainsi certains aspects de nos activités d'apprentissage et de mémorisation. En cela, ces réseaux sont beaucoup plus performants que les ordinateurs déterministes programmables classiques, et c'est pourquoi ils constituent une nouvelle génération dans les techniques d'intelligence artificielle. On parle, à propos de ces réseaux, d'auto-organisation, ou d'émergence, parce que leurs performances sont réalisées grâce à des modifications de connexions non programmées explicitement. Un exemple typique de telles performances est la reconnaissance d'un nombre infini de variations d'une forme (comme une lettre alphabétique manuscrite) après exposition à un nombre limité, relativement restreint, d'échantillons de telles variations. Contrairement aux programmes de reconnaissance de formes classiques, la classe de formes à reconnaître (la lettre avec toutes ses variations possibles) n'est pas définie par des caractères spécifiques, en sorte que des instructions de programme puissent permettre de rechercher et d'identifier ces caractères. Il existe seulement des règles d'apprentissage, qui sont des règles très générales, valables pour n'importe quelle forme à reconnaître. Suivant ces règles, l'exposition aux échantillons pendant la

14. F. Fogelman-Soulié, « De la complexité dynamique et de son utilisation dans les réseaux connexionnistes », dans F. Fogelman-Soulié (dir.), *op. cit.*, p. 48-57.

période d'apprentissage, produit des modifications dans les connexions du réseau. En conséquence, l'exposition ultérieure du réseau à d'autres formes appartenant à la même classe produira leur reconnaissance avec un taux de succès raisonnablement élevé. C'est en ce sens que le réseau s'organise lui-même pendant la phase d'apprentissage, car les règles d'apprentissage, qui, elles, sont programmées, sont générales et non spécifiques pour une forme (une lettre) particulière. Les règles sont les mêmes pour toutes les classes possibles de formes qui seront définies et spécifiées par l'expérience que constitue l'exposition elle-même, pendant la phase d'apprentissage. Alors, des connexions spécifiques sont établies automatiquement, de façon non programmée explicitement, par application des règles non spécifiques sous exposition du réseau aux stimuli externes, c'est-à-dire aux échantillons différents de la forme qui, implicitement, définissent la classe. L'auto-organisation de la structure spécifique des connexions est donc réelle en ce sens qu'elle n'est pas programmée explicitement. Cependant, l'on ne peut parler que d'auto-organisation au sens faible car la tâche à accomplir par le réseau est définie à l'avance, bien qu'implicitement, de l'extérieur par son constructeur : le but de cet apprentissage est de reconnaître une forme. Plus généralement, le but de ces machines, comme dans le cas des programmes, est de résoudre un problème bien défini (la reconnaissance de certaines formes en est un) posé *a priori* par le concepteur du réseau. Les règles d'apprentissage sont jugées efficaces ou non suivant un critère établi à l'avance, c'est-à-dire dans la mesure où elles permettent la solution de problèmes posés et la réalisation de tâches définies, autrement dit dans la mesure où elles atteignent le but poursuivi par le concepteur et l'utilisateur du réseau. En d'autres termes, la signification du fonctionnement du réseau est encore définie *a priori* par son concepteur humain sur la base d'un critère d'efficacité ou d'utilité qui est donc fourni de l'extérieur au réseau.

## B) Auto-organisation au sens fort

Au contraire, une auto-organisation *au sens fort* implique que même la tâche à accomplir, le but à atteindre, c'est-à-dire ce qui définit la signification de la structure et du fonctionnement de la machine, soit une propriété émergente de l'évolution de la machine elle-même.

C'est ce qui se produit dans des systèmes naturels non programmés où l'on observe l'émergence de structures *et de*

*fonctions* à un niveau macroscopique, à partir de contraintes physico-chimiques peu spécifiques au niveau microscopique.

Un tel comportement peut être simulé par des programmes d'ordinateur particuliers qui ne sont en fait programmés explicitement pour rien de spécifique et qui pourtant exécutent quelque chose qui a un sens. Un programme de ce type, appelé « *Soar* » par ses auteurs A. Newell, J. Laird et P. Rosenbloom[15], se présente comme un système expert programmé pour résoudre des problèmes. Mais à la différence d'un système expert classique auquel serait proposé un problème pour lequel la base de connaissances serait insuffisante, Soar ne s'arrête jamais. Il propose toujours une solution, même inadéquate et, à la limite, en partie aléatoire, à tout problème qui lui est proposé, en recherchant dans son « espace de problèmes ». Il applique ainsi des règles très générales qui ne sont pas destinées de façon spécifique à résoudre tel ou tel problème mais à se déplacer dans son espace de problèmes. Il mémorise ensuite le problème et sa « solution » enrichissant ainsi ses connaissances à partir d'« expériences » non programmées. Comme le dit Newell de façon imagée : « Soar n'a pas à être programmé pour faire quelque chose » (« *Soar does not have to be programmed to do something* »); ou encore : « Soar est finalisé (« *goal oriented* »), mais pas seulement parce qu'il a appris des buts dans sa mémoire. Il est finalisé, orienté vers des buts parce que ses buts émergent de ses interactions avec l'environnement. Il construit ses propres buts chaque fois qu'il ne peut pas simplement continuer[16]. »

Ainsi, ce qui caractérise une auto-organisation au sens fort est l'absence de but défini à l'avance et l'émergence de ce qui apparaît après coup comme un comportement fonctionnel c'est-à-dire ayant un sens.

Nos propres travaux sur l'émergence de procédures de classification dans des réseaux booléens[17], fournissent d'autres exemples de simulation d'auto-organisation au sens fort.

---

15. A. Newell, *Unified Theories of Cognition*, Cambridge Mass., Harvard University Press, 1990.
16. *Ibid.*, p. 228-229.
17. H. Atlan, E. Ben Ezra, F. Fogelman-Soulié, D. Pellegrin et G. Weisbuch, « Emergence of Classification Procedures in Automata Networks as a Model for Functionnal Self-Organization », *J. Theoret. Biol.* 120, 1986, p. 371-380. Voir aussi H. Atlan, « Self Creation of Meaning », *Physica Scripta* 36, 1987, p. 563-576.

On observe dans des réseaux d'automates en partie aléatoires, des propriétés de classification et de reconnaissance de formes sur la base de critères auto-engendrés, non programmés. Il s'agit là de simulations d'auto-organisation fonctionnelle où ce qui « émerge » est non seulement une structure macroscopique à partir de contraintes microscopiques peu spécifiques mais une fonction au sens biologique. Celle-ci, observée au niveau global, est le produit de contraintes locales peu spécifiques, partiellement aléatoires, en ce sens qu'elles ne sont pas programmées en vue de la fonction émergente qui en résulte.

Typiquement, on construit un réseau de quelques centaines d'éléments où chacun reçoit deux entrées binaires de deux de ses voisins et envoie une sortie binaire dont la valeur est celle de son propre état à deux autres de ses voisins.

Les éléments du réseau sont donc des automates qui calculent à chaque unité de temps une fonction booléenne de deux variables. La structure du réseau, c'est-à-dire la fonction assignée à chaque automate, est produite en distribuant au hasard sur les automates les 14 fonctions booléennes non constantes. Un état initial du réseau est ainsi déterminé en tirant au sort l'état initial de chaque automate. Le réseau change alors d'état de façon discrète, car chaque automate calcule son nouvel état à partir des états précédents de ses voisins qu'il reçoit comme entrées.

En général, le réseau se stabilise dans l'un de ses attracteurs après quelques dizaines d'unités de temps de calcul. L'attracteur présente typiquement une structure spatiotemporelle caractérisée par l'existence de sous-réseaux constitués par des éléments stabilisés dans un de leurs deux états possibles, séparés par d'autres sous-réseaux dont les éléments ont un comportement oscillant périodique, passant de facon répétitive par une séquence d'états relativement courte (de quelques unités à quelques dizaines). Il s'agit là d'un exemple d'auto-organisation structurale[18], où des contraintes microscopiques, partiellement aléatoires et relativement simples, créent une dynamique conduisant à l'émergence de structures macroscopiques, et constituent ainsi un modèle de

18. F. Fogelman-Soulié, « Réseaux d'automates et morphogénèse », dans P. Dumouchel et J. P. Dupuy (dir.), *L'auto-organisation, de la physique au politique*, Colloque de Cerisy, Paris, Seuil, 1983, p. 101-114. Voir également H. Atlan, « L'émergence du sens et du nouveau », *ibid.*, p. 115-138.

morphogénèse. Depuis les travaux de A. Turing en 1952[19] qui décrivaient pour la première fois de tels modèles de morphogénèse à partir de couplages de réactions chimiques et de diffusion, on connaît maintenant de nombreux exemples de dynamiques, discrètes ou continues, produisant ainsi un transfert de structure du local au global avec émergences de formes macroscopiques stables à partir d'états homogènes et de structures microscopiques[20].

Mais de plus, nous avons montré[21] que de telles structures peuvent présenter des propriétés fonctionnelles qui n'avaient pas été programmées, et qui apparaissent elles aussi, *a posteriori*, comme propriétés émergentes de la dynamique. Une telle auto-organisation fonctionnelle est observée sur nos réseaux booléens quand on perturbe un élément d'un sous-réseau préalablement *stabilisé* dans un attracteur, en lui *imposant* une séquence temporelle de signaux binaires. On recherche alors les modifications produites par cette séquence perturbatrice sur les autres éléments du réseau. Dans ces conditions, on observe que ces réseaux fonctionnent comme des systèmes capables de reconnaître des formes qui leur sont présentées de l'extérieur. Les formes à reconnaître sont des séquences temporelles binaires de 0 et 1 arrangées d'une certaine façon et imposées sur l'élément du réseau qui sert ainsi d'entrée dans le système de reconnaissance. Ces séquences ont un rôle perturbateur qui déstabilise certains éléments stables. Mais curieusement, certaines de ces séquences ont aussi pour effet de *stabiliser certains éléments qui étaient oscillants*. Ce phénomène est une sorte de résonance entre la structure temporelle de la séquence perturbatrice et la séquence d'états périodiquement répétitive du ou des éléments oscillants. Toutefois, s'il ne s'agissait que d'une résonance simple, une seule séquence perturbatrice aurait cette propriété grâce à une structure périodique unique qui compenserait exactement les oscillations de ces éléments. En fait, la structure temporelle de ces séquences stabilisatrices (c'est-à-dire « reconnues ») n'est que partiellement

19. A.M. Turing, « The Chemical Basis of Morphogenesis », *Phil. Trans. Roy. Soc.*, Londres, B 237, 1952, p. 37-72.

20. I. Prigorine, « Structure, Dissipation and Life », dans M. Marois (dir.), *Theoretical Physics and Biology*, Amsterdam, North Holland Publ. Co., 1969, p. 23-52 ; A. Katchalsky, « Biological flow-structures and their relation to chemicodiffusional coupling », *Neurosciences Res. Progr. Bull.*, vol. 9 n° 3, 1971, p. 397-413; S.A. Kaufman, « Emergent Properties in Random Complex Automata », *Physica*, 10 D, 1984, p. 145-156.

21. *Cf.* note 16.

périodique, en ce qu'elle est constituée par une séquence répétée présentant des variations aléatoires. Typiquement, une telle suite est représentée par des 0, des 1, et des astérisques (par exemple 00* 1* 01* 1*0); lorsque cette suite est répétée pour constituer une séquence périodique binaire, les astérisques sont remplacés par 0 ou 1 indifféremment, de façon aléatoire. Il en résulte que cette structure pseudo-périodique (ou partiellement aléatoire) définit non pas une seule séquence qui serait reconnue par stabilisation d'un élément oscillant, mais *une classe* de telles séquences réalisées par toutes les variations possibles dans le remplacement des astérisques par un signal binaire. La reconnaissance consiste à différencier une séquence appartenant à cette classe, de toute autre séquence perturbatrice. Et le critère de reconnaissance n'est donc pas autre chose qu'une structure pseudo-périodique donnée, dont la propriété stabilisatrice n'est qu'une conséquence de l'état final d'organisation macroscopique du réseau, lui-même émergent à partir des contraintes microscopiques imposées par les fonctions booléennes et les connexions. C'est en ce sens que l'on peut parler ici d'auto-organisation au sens fort : les contraintes microscopiques et les conditions initiales n'étant pas programmées (puisqu'elles sont établies au hasard), les fonctions qu'on observe (reconnaissance de séquences temporelles) au niveau des structures macroscopiques qui en résultent ne sont elles-mêmes programmées en aucune façon. Elles ne sont le résultat d'aucune finalité intentionnelle, d'aucune recherche intentionnelle de but. Bien au contraire, leur émergence semble être une création de ce qui nous apparaît comme source de signification, sous la forme de critères de classification qui n'ont pas été programmés comme tels.

## C) Auto-organisation intentionnelle

### a) Le rôle de l'interprétation

Mais nous devons faire maintenant une autre distinction, car pour décrire ces propriétés émergentes, à la fois structurales et fonctionnelles, on ne peut pas éviter de tenir compte de l'existence et du point de vue de l'observateur (non pas avec le sens d'une subjectivité, mais, de façon habituelle en physique, avec le sens de conditions objectives d'observation et de mesure). Considérer un réseau comme ayant acquis — par auto-organisation — la signification d'une machine à reconnaître des formes est une *interprétation* par l'observateur. Plus précisément, c'est une projection de nos propres expériences cognitives de reconnais-

sance de formes – soit directement par le moyen de nos machines faites pour cela – sur le comportement observé dans le réseau. Par conséquent, même dans le cas de systèmes auto-organisateurs au sens fort, censés simuler des systèmes naturels mécaniques non produits par l'homme – tels que par exemple des plantes ou des animaux –, nous restons avec une dernière question sur la signification de *l'observation* et de *l'interprétation* qui produisent la signification.

De ce point de vue, les systèmes humains, tant individuels que sociaux, apparaissent dans une situation intermédiaire curieuse entre des systèmes naturels dont l'origine des significations est interne, et ne peut être observée que de l'extérieur par projection interprétative, et des machines artificielles dont la finalité, et l'origine des significations, est connue, en ce qu'elle est planifiée et observée directement par les hommes eux-mêmes. Les systèmes humains sont donc à la fois des machines et des concepteurs/observateurs de ces machines.

Ainsi, nous sommes amenés à considérer les systèmes humains comme occupant une place particulière en ce que, du fait de leurs conditions particulières d'observateurs-observés, on doit nécessairement y prendre au sérieux la question de l'intentionnalité, alors qu'on pourrait la considérer comme illusoire ou épiphénoménale dans le cas d'autres systèmes mécaniques.

Observons d'abord que ceci ne présume en rien que nous devions accepter la *réalité du libre arbitre* et ne pas supposer l'existence de déterminations causales pour les intentions elles-mêmes. Seulement, l'intentionnalité créatrice de projet est reconnue comme une causalité efficiente particulière et, comme telle, un objet spécifique des sciences de l'homme. La question du libre arbitre est mise de côté jusqu'à ce que nous connaissions en détail comment les intentions spécifiques sont déterminées causalement – si cela est possible un jour, et si la sous-détermination des théories par les faits[22] ne constitue pas une limitation irréductible de cette connaissance.

### b) Transformation d'une séquence causale en procédure

Dans cette recherche de mécanismes physiques d'intentionnalité, nous devons donc tenter d'aller plus loin. À partir de nos modèles d'auto-organisation au sens fort, rien n'empêche de

22. H. Atlan, *Tout, non, peut-être. Éducation et vérité*, Paris, Seuil, 1991, chap. 3 et 4.

concevoir que la capacité de faire des projets, puisse être comprise elle aussi dans son principe général, et modélisée un jour comme résultat d'un mécanisme d'auto-organisation – au sens fort et intentionnel donc, c'est-à-dire humain – dans le fonctionnement de réseaux de neurones. Cette conception est tout à fait cohérente avec le fait que nous pouvons observer certains types de comportements apparamment intentionnels, même limités, produits par des réseaux de neurones différents des nôtres, en l'occurrence des cerveaux d'autres animaux pourvu qu'ils soient suffisamment semblables aux nôtres, tels ceux de mammifères évolués. En effet, il semble bien que certains singes ayant effectué une première fois des gestes sans signification fonctionnelle apparente, découvrent les effets de ces gestes et en reproduisent la séquence de façon apparemment finalisée avec l'intention de reproduire ces effets. C'est ainsi que la fabrication d'outils primitifs orientée vers leurs utlilisations futures ne peut être comprise, comme chez les hommes, que grâce à la capacité de faire des projets capables de déterminer des comportements que nous avons toutes les raisons de qualifier d'intentionnels.

Une machine relativement simple capable d'apprentissage inventif de ce type, reproduisant donc un comportement intentionnel, pourrait fonctionner de la façon suivante.

Comme nous l'avons déjà observé, il peut arriver qu'une structure temporelle, sous la forme d'une succession d'états d'un réseau émergeant d'une dynamique purement causale d'interactions entre ses éléments, acquière une signification fonctionnelle aux yeux d'un observateur extérieur. C'est ce qui se produit lorsque cette succession d'états aboutit à un état final où le réseau effectue quelque chose qui apparaît alors comme une fonction que la série des états précédents a déterminée. Cette série apparaît alors *a posteriori* – toujours aux yeux d'un observateur extérieur – comme celle des moyens utilisés pour arriver à cette fin.

Supposons maintenant que cette séquence d'états soit mémorisée comme telle. Tout se passe alors comme si c'était une procédure qui est ainsi mémorisée puisque l'état final y est retenu – avec son effet – en tant qu'état final de cette séquence. Supposons, de plus, qu'un réseau neuronal puisse fonctionner comme un observateur de lui-même – chose que nous ne savons pas encore faire dans des réseaux artificiels, mais dont semble être capable, apparemment, la substance réticulée de nos cerveaux. Il peut alors arriver que l'état final de la séquence précédente d'états soit rappelée en mémoire, de façon associative, par une séquence

d'états différente de la précédente, déclenchée elle-même par la perception ultérieure de l'effet de cet état final – c'est-à-dire par ce qui est maintenant interprété comme la fonction de cet état par le réseau du fait de son observation de lui-même. Ce qui sera rappelé en mémoire sera alors non seulement cet état final mais toute la procédure, puisqu'elle a été mémorisée en même temps que lui.

Par exemple, imaginons que l'état final de la séquence consiste en une modification de la forme d'un os qui lui permet d'être utilisé comme un outil par un grand singe en faisant, disons, un trou dans la terre. Supposons maintenant que le réseau puisse – ce que nous ne savons pas encore modéliser – fonctionner comme observateur de lui-même. Alors, certains effets de ce même état final – un trou dans la terre – pourront produire, par mémoire associative, l'état final en question correspondant à la modification de l'os effectuée précédemment. Le trou dans la terre sera interprété comme une conséquence de la modification de l'os, alors même qu'il est la cause du rappel, par mémoire associative, de l'état du réseau correspondant à cette modification. De plus, cela n'est pas seulement cet état qui est rappelé, mais toute la séquence d'états qui l'avait produit, et c'est toute cette séquence qui est alors observée par le réseau comme une procédure dont la signification est de faire un outil en vue de creuser un trou dans la terre. Chaque fois que cette séquence d'états sera rappelée en mémoire à l'occasion d'associations avec des stimuli divers, elle aura la même signification d'un comportement intentionnel, réalisation apparente d'un projet dirigé vers et par le futur. En fait, ceci n'est que le résultat de ce que toute la série des causes et des effets a été mémorisée, y compris le dernier effet. Lors de la répétition de cette procédure que déclenche son rappel par mémoire associative, ce qui est rappelé c'est sa représentation mémorisée où sa fin est rappelée en même temps que son commencement.

Ainsi, nous pouvons imaginer un modèle mécanique et purement causal de notre capacité de faire des projets, telle qu'elle semble être présente déjà chez les chimpanzés.

Un degré de plus serait atteint avec la mémorisation des activités elles-mêmes de mémorisation de procédure. C'est cela qui permettrait de modéliser notre expérience de fabrication de projets en général, indépendante de tel ou tel projet spécifique; autrement dit, qui produirait en nous la conscience

de l'intentionnalité. Comme nous l'avions suggéré autrefois[23], notre conscience volontaire serait le résultat de mémorisation de phénomènes auto-organisateurs qui produisent, de façon inconsciente, l'avenir de la nouveauté : association de vouloir inconscient à une conscience-mémoire instituant l'unité temporelle fragile d'un soi.

Dans tous les cas, le projet sur l'avenir ne serait que le résultat du retournement d'un effet en cause dans la représentation. Ce qui était la fin d'une procédure et son effet dans une action qui a précédé sa représentation est retourné en cause dans la représentation elle-même. Et ce retournement n'a rien de bien mystérieux car il est rendu possible par l'aplatissement du temps que produit toute mémorisation. Dès qu'une séquence temporelle est mémorisée, l'ordre passé-futur est transformé en un ordre symbolique où le temps disparaît, puisque toute la séquence est présente en même temps dans la mémoire; ce qui n'empêche pas chaque événement de la séquence de rester affecté d'un indice de temporalité puisque celui-ci n'est évidemment que symbolique et non réellement temporel.

### c) Conscience-mémoire et auto-organisation inconsciente

Ainsi, si nous savions comment construire une machine capable de s'observer elle-même, il semble que nous saurions alors construire un modèle purement causaliste de notre capacité de faire des projets, qui semble déjà observable chez les grands singes.

Un pas de plus pourrait être fait si le processus de mémorisation de procédures était lui-même mémorisé. Ceci produirait alors un modèle de l'expérience que nous avons d'une capacité de faire des projets en général, indépendamment de tel ou tel projet particulier. En d'autres termes cela produirait alors un modèle mécanique de notre conscience comme conscience d'intentionnalité au sens limité et réduit évidemment que nous avons envisagé.

On voit que dans un tel modèle, la conscience ne serait pas visée vers, orientée vers un inexistant futur ou simplement hors d'elle, mais au contraire elle serait principalement mémoire d'états et de processus passés. Par contre l'intentionnalité créative, comme inventrice de projet et donatrice de signification,

23. H. Atlan, *Entre le cristal et la fumée*, Paris, Seuil, 1979, ch. 5.

serait le propre de phénomènes auto-organisateurs au départ inconscients, produits par l'émergence de structures et de fonctions nouvelles, comme dans nos réseaux d'automates. Ces structures et fonctions nouvelles apparaîtraient évidemment comme telles dans une activité secondaire lorsque la conscience-mémoire observerait cette émergence et la mémoriserait à son tour.

Maintenant, c'est donc sur ce mode d'interaction entre conscience-mémoire et auto-organisation inconsciente qu'il faut réfléchir pour tenter d'entrevoir comment une conscience intentionnelle ou une intention consciente peut ainsi apparaître non pas comme phénomène premier, fondateur, mais comme phénomène secondaire, dérivé[24].

### d) Sophistication infinie

Dans la tâche qui tend à unifier le calcul par réseaux de neurones et notre expérience d'intentions et de significations, un dernier pas consisterait maintenant à trouver un nouveau principe d'architecture et d'organisation, qui serait le propre de notre intentionnalité et de nos capacités sémantiques, en ce qu'il rendrait compte de notre capacité d'interprétation apparemment infinie. Nous sommes apparamment capables de donner à n'importe quoi une signification, en projetant des significations sur tout objet perçu; aussi bien d'ailleurs des significations mécaniques, celles du monde naturel, que des significations proprement intentionnelles rapportées à la conscience, suivant les règles du jeu que nous nous donnons nous-mêmes.

Dans ce travail, qui n'existe bien sûr qu'à l'état de projet de recherche, un tel principe d'architecture et d'organisation se fonderait sur l'idée qu'un modèle d'auto-organisation intentionnelle tel que celui que nous avons discuté pourrait être généralisé encore plus.

Nous sommes partis de la capacité de créer un projet, en *transformant* une conséquence-effet d'une séquence d'évènements, *en cause* déclenchant une procédure (le trou dans le sol, effet, est transformé après mémorisation et rappel, en cause déclenchant une procédure permettant de le produire). Ensuite, nous avons suggéré qu'une capacité de mémoriser ce processus de mémorisation des procédures lui-même est un substrat possible pour notre conscience d'intentionnalité de la conscience.

---

24. *Ibid.*

Ce qui manque maintenant, c'est une capacité non seulement de mémoriser des procédures avec leur signification fonctionnelle, et ensuite mémoriser ce processus de mémorisation lui-même, mais encore *modifier les significations des procédures en n'importe quelles circonstances qui perturbent et modifient ces procédures de façon imprévue, c'est-à-dire de façon non déjà mémorisée.*

Évidemment, à l'origine de cette idée se trouve la notion qu'on a déjà rencontrée que la signification d'un objet ou d'un énoncé n'est pas une propriété intrinsèque de cet objet ou de cet énoncé; mais que cette signification est toujours générée par un acte d'interprétation, à l'interface entre l'observateur et l'observé. Pourtant, l'observation et même l'auto-observation ne sont pas suffisantes. C'est pourquoi la seule existence de la formation réticulée dans le cerveau d'espèces de vertébrés autres que les hommes – alors que chez l'homme cette formation semble indispensable à la conscience de soi – peut nous conduire à conclure que ces animaux peuvent avoir nécessairement des capacités sémantiques et intentionnelles semblables à celles dont nous faisons l'expérience.

Pour fonctionner comme source apparemment infinie d'interprétations, l'auto-observation ou la conscience de soi ne doit pas être un simple organe passif de mémorisation. Ce système doit en plus être connecté à un ou plusieurs organes auto-organisateurs, c'est-à-dire capables de produire indéfiniment de la nouveauté – sous forme d'états ou séquences d'états – qui seront alors interprétés chaque fois, indéfiniment, par le système lui-même avec une nouvelle signification.

La question la plus difficile reste évidemment : quelle serait l'architecture, la nature d'une telle connexion, entre mémoire du passé et auto-organisation productrice de nouveau d'où sortirait une capacité infinie de signifier ce nouveau en l'interprétant ?

C'est là que du point de vue de la théorie de la complexité des algorithmes une telle capacité d'interprétation pourrait être assignée à une classe particulière d'algorithmes formellement définis comme capables de générer des objets infinis avec une sophistication apparemment infinie.

C'est là donc que nous retrouvons notre sophistication comme mesure d'une complexité porteuse de signification.

Nous avons vu que la sophistication est définie comme la partie qui véhicule la signification dans la mesure classique de la complexité d'un algorithme, en établissant une distinction

formelle entre la partie programme (qui définit la structure d'une classe avec une certaine signification) et la partie données (qui ne fait que spécifier une instance particulière parmi tous les membres de la classe d'objets partageant la même structure).

Cette définition permet de distinguer la complexité classique des suites aléatoires infinies, elle-même infinie, d'avec leur sophistication qui est nulle. De façon générale, une suite infinie a en général une sophistication finie, longueur de la partie programme dans la description minimale capable de la produire. Sophistication finie qui peut être très grande bien sûr si la suite produite a une structure très complexe en ce sens qu'elle a besoin d'un programme minimum très long.

Mais il est aussi possible de concevoir une classe d'objets évolutifs, descriptibles par des suites infinies, dont la sophistication elle aussi serait infinie. Cela voudrait dire, en gros, que la signification de tels objets, l'ensemble des significations qu'on pourrait y découvrir ou projeter, se modifierait sans cesse au fur et à mesure qu'on découvrirait une nouvelle partie de l'objet.

De tels objets évolutifs ont la propriété curieuse d'être non récursifs en ce qu'on ne peut pas les produire de façon mécanique par un algorithme, et pourtant non aléatoires puisque leur sophistication est grande, alors que classiquement un objet ne peut être que récursif, c'est-à-dire produit par un programme, ou aléatoire.

Cette propriété particulière de n'être ni récursif ni aléatoire est ce que nous attendons de choses porteuses de significations auxquelles de nouvelles significations sont attribuées sans cesse, à l'occasion de réorganisations incessantes et inattendues.

Ainsi, si l'on admet que la sophistication des séquences d'états dans nos cerveaux est assez grande pour être pratiquement assimilée à l'infini, – notamment du fait des communications entre cerveaux différents dans l'espace et le temps à l'aide de mémoires accessoires et de systèmes de calcul – ceci pourrait expliquer la capacité apparente de créer indéfiniment de nouvelles significations par interprétation.

Alors, si des machines produisant des comportements de sophistication apparemment infinie pouvaient être construites par des techniques d'I.A., notamment sous la forme d'architectures de réseaux d'automates partiellement aléatoires et en plusieurs niveaux, alors, peut-être la question pourrait se poser de savoir si de telles machines feraient la même expérience

d'intentionnalité et de don de signification que nous faisons nous-mêmes.

## Conclusion

Ainsi, l'analyse de mécanismes d'auto-organisation tels que nous pouvons les observer dans des systèmes physico-chimiques ou dans des modèles de ces systèmes, peut nous servir de fil conducteur qui nous permettra peut-être, de façon progressive, de concevoir la notion de signification à partir d'un concept physique, pas moins physique même s'il est aussi abstrait, que celui d'énergie, ou d'entropie, ou de fonction d'onde, ou d'état cérébral.

*Hôtel-Dieu de Paris*