Gérard Sabah

Conscience et Intelligence artificielle(s) vues par Jacques Pitrat

Volume 3, nº 1-2 (2022), p. 179-192.

<http://roia.centre-mersenne.org/item?id=ROIA_2022__3_1-2_179_0>

Revue Ouverte d'Intelligence Artificielle
Volume 3, nº 1-2, 2022, 179-192

Conscience et Intelligence artificielle(s) vues par Jacques Pitrat

Gérard Sabah[a]

[a] 6 allée des Ormes 78470 Saint-Rémy-lès-chevreuse, France
E-mail : gsabah@free.fr.

Résumé. — Pendant de longues années, les chercheurs en intelligence artificielle et en sciences cognitives se sont gardés d'aborder le thème de la conscience, vue comme une notion trop vague pour permettre une étude scientifique. Ainsi, Bertrand Russell [14] prétendait les résultats de l'introspection scientifiquement inutilisables car n'obéissant pas aux lois physiques. De même, le behaviourisme, voulant fonder la psychologie comme science exacte, exclut toute notion d'état mental, et rejette ce qui concerne la conscience comme fondamentalement hors de son domaine.

Un renouveau de cette question est dû à la théorie darwinienne de l'évolution et au matérialisme orthodoxe. Par ailleurs, Searle [21] estime scandaleux qu'une science qui se veut étudier l'esprit ignore les aspects liés à la conscience. Le problème essentiel est alors d'expliquer les sentiments, la conscience et le libre arbitre en ne se fondant que sur les lois physiques.

Après avoir précisé quelques idées générales sur l'intelligence et la conscience, j'expliciterai les raisons qui m'ont amené à m'intéresser à la conscience à partir du traitement automatique des langues et je décrirai les aspects de la conscience que Jacques Pitrat a mis en œuvre dans CAIA (Chercheur Artificiel en Intelligence Artificielle).

Mots-clés. — Conscience artificielle, traitement automatique des langues, résolution de problèmes.

La seule façon d'exister pour la ***conscience***
c'est d'avoir ***conscience*** *qu'elle existe.*
Jean-Paul Sartre

1. Généralités

1.1. Conscience, de quoi s'agit-il ?

Un premier point à souligner est la multiplicité de concepts que recouvre la notion de conscience (comme d'ailleurs les notions liées d'intention, d'attention, de mémoire...) ; on peut donc difficilement prétendre en donner une définition unique et indiscutable. Différents sens de « conscience » apparaissent selon la langue considérée, aussi bien

à partir d'exemples issus de la Bible que de la langue anglaise qui propose déjà quatre mots pour exprimer des états très différents. Quand on parle de la *conscience* pour qualifier l'état d'éveil par rapport au *sommeil ou au coma*, l'anglais emploiera le mot *awakeness*, de la conscience de *telle ou telle information particulière*, on emploiera le mot *awareness*, de la *conscience d'être soi-même*, d'avoir la connaissance de ses propres connaissances et la perception de sa propre activité psychique, le mot *consciousness*, enfin, de la *conscience morale* pour la faculté de porter des jugements de valeur sur ses actes, le mot *conscience*, (sans parler de *conscience professionnelle*, ni de *liberté de conscience* ou autre...).

Nous nous concentrerons ici essentiellement sur le 3e point (l'introspection, c'est-à-dire la faculté de représenter ses propres opérations mentales, le monde extérieur, et de raisonner sur ces représentations [*consciousness*]) et le 4e (faculté de juger le bien et le mal [*conscience*]).

Bien que la recherche sur la conscience soit souvent bloquée par les idées métaphysiques de beaucoup des chercheurs sur le sujet, à cause de « la spécificité humaine » qu'ils revendiquent, quelques informaticiens français ont voulu approfondir le sujet (par exemple : Anceau [1], Cardon [6], Pitrat [12], Sabah [16]), supposant que les concepts évoqués ci-dessus constituent le cœur du système qui construit notre comportement et que ce peut être valable aussi pour les machines. Il semble qu'aujourd'hui le défi de fabriquer un objet doté d'une conscience réflexive ayant certaines caractéristiques de celle de l'homme soit à la portée des recherches actuelles et à venir. Jugeant de l'importance des conséquences potentielles et des précautions à prendre, l'Académie des technologies a soutenu, sur ce thème, un groupe de travail que j'ai co-dirigé avec Philippe Coiffet. Cela a amené un rapport où nous avons détaillé divers modèles de la notion de conscience et proposé une synthèse (toute théorique) des travaux français évoqués ci-dessus [18]. Nous en reprenons certains éléments dans ce qui suit.

1.2. Les limites de l'intelligence humaine

À propos du cerveau, rappelons tout d'abord quelques chiffres bien connus : il y a de l'ordre de 10^{10} neurones dans le cortex, ces neurones entretiennent 10^{15} connexions en tout, ce qui en d'autres termes correspond à 10^9 connexions par mm^3. Tout cela pour remarquer qu'il y a quelque $10^{\text{plusieurs millions}}$ combinaisons possibles d'états du cerveau... ce qui nous donne de bonnes capacités intellectuelles, malgré la lenteur des neurones (qui restent en nombre limité, même si ce nombre est très grand). Même avec les puissances actuelles, les machines sont loin de pouvoir représenter un tel ensemble d'états.

Jacques Pitrat souligne également des limites humaines : l'une d'elles vient de notre impossibilité à modifier la structure de notre cerveau ; nous sommes, par exemple, incapables d'ajouter ou de modifier des modules de traitement.

De plus, notre mémoire à court terme est relativement limitée (certains auteurs parlent de sept registres (Miller [10]) ; pour réutiliser les structures mentales qui n'y

sont plus disponibles, il faut qu'elles aient été sauvegardées dans la mémoire à long terme puis réactivées.

Enfin, même si nous pouvons expliquer ce que font nos processus conscients, outre le fait que nous sommes incapables d'accéder à nos processus inconscients, nous ne pouvons pas « méta-expliquer » (par exemple expliciter pourquoi on a les intuitions qui sont au départ de nos raisonnements).

Contrairement à ce qu'affirment certains (par exemple Eccles [7] qui estime que grâce à la conscience, l'humanité a atteint le summum de l'intelligence), Jacques Pitrat pense qu'il est très improbable que l'être humain soit au maximum de l'intelligence. Il affirme que la possibilité d'atteindre la singularité[(1)] n'existe que si notre intelligence est suffisante pour développer un « bootstrap » (amorçage) de l'IA. Pour lui, notre intelligence seule est insuffisante et la seule possibilité raisonnable est de se faire aider par l'IA elle-même, qui, elle, est capable de modifier automatiquement ses propres modules, pouvant rendre ainsi ses programmes plus efficaces.

Chez l'homme, on constate également que le niveau méta est très difficile d'accès (en particulier sans confondre les niveaux). Analysons de plus près pourquoi il est donc utile, en intelligence artificielle, de s'intéresser à la conscience, en particulier dans le domaine qui fut le mien : celui du traitement automatique des langues.

2. Traitement des langues et conscience

Jacques Pitrat s'est également intéressé au domaine du traitement automatique des langues (Pitrat [11])[(2)], développant les approches de Schank [19, 20] et mettant l'accent sur les relations entre compréhension et actions. Toutefois, bien qu'il parle déjà de méta-connaissances, il n'aborde pas directement la question de la conscience dans ce cadre. Avant d'approfondir son approche, je voudrais évoquer quelques points qui m'ont amené à m'intéresser à cette notion dans le cadre du traitement automatique des langues afin de souligner les convergences entre le traitement automatique des langues et la résolution de problèmes sur ce concept.

2.1. Où est le sens ?

Il n'est pas prouvé que les langues soient formalisables et il est même probable que l'homme ne les traite pas de façon formelle. L'adéquation du formel aux traitements automatiques doit alors être examinée soigneusement.

Une première idée est de considérer que le sens est dans les mots, dans la langue :

> *Quand ils entendent des mots, la plupart des gens pensent qu'il doit y avoir un message quelque part. (Gœthe, Faust)*

(1) L'hypothèse selon laquelle une progression forte de l'intelligence artificielle lui permettrait de dépasser l'intelligence humaine et déclencherait une croissance exponentielle de la technologie provoquant une perte de contrôle de l'humanité sur son développement.

(2) Sans parler, bien entendu, de toutes les thèses sur le sujet que Jacques Pitrat a encadrées et qui sont évoquées dans les autres articles du présent volume.

L'énonciateur a dans l'esprit un message à transmettre et des règles d'encodage. L'auditeur utilise un processus de décodage qui lui permet d'identifier les sons produits, les structures syntaxiques utilisées, les relations sémantiques correspondantes et de combiner tous ces éléments pour reconstruire le sens du message compris. Mais cette approche pose des problèmes pour modéliser les processus de compréhension, avec les ambiguïtés, les références, les métaphores, l'ironie, les performatifs... Il faut être capable, à partir du supposé sens littéral, de calculer le sens réellement transmis dans le contexte (l'intention communicative, éventuellement différente du sens littéral[(3)]).

Ainsi, on est amené à considérer que le sens littéral n'est pas pertinent, et qu'il faut donner primauté à l'interprétation contextuelle. Alors, le sens est dans la tête 1) de celui qui parle, et 2) de celui qui comprend ! Comme le dit Barthes [3] dans son langage imagé :

> *l'homme parlant parle l'écoute qu'il imagine à sa propre parole.*

L'exemple suivant indique que le locuteur n'avait (probablement) qu'une seule intention communicative dans la tête, alors que le lecteur peut en percevoir deux, ce qui, vu le contexte, déclenche le rire :

> *« Si vous voulez des enfants, adressez-vous à Monsieur le Curé »*
> *(avis aux dames catéchistes, lu dans une église)*

Prenons l'exemple de la phrase toute simple montrant l'importance du contexte : *je reviendrai*. À la fin d'une visite amicale, c'est une *promesse* ; dans la bouche d'un client, ce sera un *compliment* ; ce sera l'*avertissement* d'un policier s'adressant à un automobiliste mal garé ; une *menace* d'un propriétaire qui n'a pas été payé ; une *prédiction* consolatrice d'un soldat qui part au front...

Autre exemple montrant qu'il faut tenir compte des connaissances de l'interlocuteur :

> *Faraday présente ses travaux à Gladstone (ministre britannique des finances) qui lui demande « À quoi ça sert ? »*
> *Réponse : « Un jour vous nous ferez payer des impôts dessus »*

Ces quelques exemples montrent qu'une communication efficace nécessite d'avoir un modèle de soi-même et de son interlocuteur. C'est-à-dire, *avoir une conscience de soi et de l'autre*.

[(3)] Et c'est précisément la distinction essentielle entre les langages artificiels et les langues humaines : les premiers transmettent exactement ce qu'on dit, alors que les langues transmettent ce qu'on imagine que l'interlocuteur va interpréter. La communication est réussie lorsque c'est effectivement ce que l'interlocuteur interprète.

2.2. Des traitements imprévisibles

Les premières idées de mises en œuvre ont impliqué des utilisations séquentielles des diverses connaissances nécessaires à la compréhension (on commence par chercher la forme canonique des mots de la phrase, puis grâce à la grammaire, on construit la [les] structure[s] syntaxique[s] de la phrase, on en cherche les interprétations sémantiques pertinentes, etc.). Ces différents changements de représentation sont effectués par des modules particuliers dont chacun utilise un type de connaissances particulier. Si, à un moment donné du traitement, ses connaissances accessibles ne permettent pas au module en cours de prendre la bonne décision, on se trouve devant un *point d'embarras* : une situation où plusieurs solutions semblent pertinentes. Nous parlerons alors d'ambiguïtés « artificielles » pour les points d'embarras qui **ne** sont **pas** dus à la langue elle-même, mais au programme[(4)].

Plusieurs causes entraînent ces ambiguïtés artificielles : premièrement, on ne peut déterminer l'ordre idéal des modules en toute circonstance ; il faut le déterminer dynamiquement.

L'exemple suivant ne peut se résoudre qu'en faisant intervenir *simultanément* la résolution d'ambiguïté (père et fils dans les domaines familiaux ou religieux) et la résolution d'anaphore :

> *[Siméon, le fils de Joanis, est curé]*
> *« — Mon père, dit Joanis à son fils, je suis en grand souci*
> *— À propos de quoi, mon fils ? fit Siméon à son père. » (Jean Anglade, Les Bons Dieux)*

Deuxièmement, autre cause, on ne sait pas *a priori* quels seront les processus nécessaires :

> *Donnez le si, il pousse un if*
> *Faites le tri, il naît un arbre*
> *Jouez au bridge, et le pont s'ouvre... (Boris Vian)*

Là, il faut des connaissances phonétiques, et à propos d'autres langues (était-ce prévisible ?)

Et ci-dessous connaître des lois de l'arithmétique pour détecter l'absurde :

> *En mettant les bouchées doubles, on fait en 6 mois ce*
> *qui devait l'être en 18 (dit par un ministre à la radio)*

Enfin, il est des textes qui décrivent eux-mêmes ce qu'il faut faire pour les comprendre. Dans les deux exemples suivants, la partie soulignée décrit comment le comprendre.

---

(4) Rappelons l'exemple de Petrick : en se limitant aux seules connaissances syntaxiques, il arrive à produire 572 analyses différentes pour une phrase de 17 mots !

> *Jean écrit en remplaçant les « o » par des « x ». Il écrit à Paul : « nxus viendrxns demain » (Edgar Poe, ixage d'un paragrab)*

Ici, il faut modifier la procédure classique d'analyse morphologique : en simplifiant, il faut, quand on a détecté que c'est Jean qui écrit, vérifier si le mot tel quel existe, sinon remplacer d'abord les x par des o et voir si le nouveau mot existe.

Autre exemple :

> *« Et il m'a dit, ajouta-t-il, en jouant de petits accords aux endroits où je mettrai des points, que Chécoavins avait laissé. Trois enfants. Sans mère. Et que la profession de Chécoavins. Étant impopulaire. La génération montante des Chécoavins. Était dans une situation très difficile » (Dickens, La maison d'Apre-Vent)*

Là, il faut modifier l'analyse syntaxique en ne considérant pas les points, et modifier la procédure d'interprétation de la situation en mémorisant que des accords ont été joués là où figuraient ces points.

S'il est clair que ces exemples (loin d'être aussi isolés qu'il le semble) restent actuellement hors du champ des systèmes de traitement automatique des langues, il est tout aussi évident qu'un système de compréhension 1) ne peut prévoir toutes les situations qu'il rencontrera et 2) doit être capable de se *reconfigurer dynamiquement* en modifiant ses procédures de traitement. Seuls des systèmes *réflexifs* et *conscients* pourront les traiter efficacement. Une véritable intelligence artificielle doit donc être capable d'évaluer et de modifier ses propres programmes et donc d'avoir une conscience réflexive.

2.3. Processus automatiques vs. processus rationnels

Lorsque nous pensons, soit nous pouvons dire quelque chose de la façon dont certaines opérations mentales ont été effectuées ainsi que de leurs interactions (nous appellerons de tels processus « conscients »), soit on ne se rend compte que de leur résultat (processus « inconscients », ils agissent alors uniquement comme machines automatiques, par une activité qui relève du réflexe).

Suivant l'affirmation d'Edelman [8] :

> *Les fonctionnalités nécessaires à une véritable intelligence sont celles qui, fondées sur l'inconscient, permettent l'émergence de la conscience chez l'homme.*

nous nous sommes posé la question : *pourquoi pas chez les machines ?*

Pour répondre positivement à cette interrogation, nous avons proposé le modèle CARAMEL (*Conscience, Automatismes, Réflexivité et Apprentissage pour un Modèle de l'Esprit et du Langage*), développé et partiellement implémenté dans les années 90

[15, 17, 16]. Un premier niveau traite les perceptions par des processus non contrôlés : une extension des tableaux noirs (le *carnet d'esquisses*) permet à de tels processus d'interagir « *inconsciemment* » en tenant compte de diverses rétroactions. Le deuxième niveau se fonde sur l'idée que la réflexivité et l'IA distribuée permettent le développement de programmes capables de représenter leurs propres actions et de raisonner sur ces représentations pour adapter dynamiquement leur comportement. Un modèle simpliste de conscience établit alors un lien entre ces deux niveaux de traitement par l'intermédiaire d'un modèle de mémoire comportant une mémoire à long terme (les connaissances du système), une mémoire de travail (où sont élaborés les résultats des processus conscients et inconscients) et une mémoire à court terme (où émergent les éléments qui deviennent conscients). Un avantage de ce modèle est que le niveau conscient peut se concentrer sur les tâches les plus adaptées à un traitement rationnel, les autres problèmes étant filtrés au niveau subliminaire. Pour établir un lien entre les niveaux de traitement contrôlés et non contrôlés, la conscience joue un rôle fondamental. Elle peut être vue comme un pont entre les processus automatiques et les processus contrôlés, débouchant éventuellement sur de possibles apprentissages.

Ce programme a montré que :

(a) la modularité est une nécessité pratique,
(b) un contrôle indépendant est utile pour choisir dynamiquement l'agent à déclencher dans un contexte donné et
(c) un contrôle distribué permet aux agents de se représenter eux-mêmes, ainsi que ce qu'ils sont en train de faire (*réflexivité*), adaptant leur comportement le mieux possible à la situation.

Ainsi, cette caractéristique d'auto-représentation, d'auto-référence et d'auto-jugement semble une qualité déterminante de l'intelligence et en particulier de la compréhension du langage.

Voyons maintenant comment Jacques Pitrat a abordé cette question.

3. La résolution de problèmes

3.1. Quelques idées de base de Jacques Pitrat

En intelligence artificielle (en particulier si l'on songe à l'intelligence artificielle forte), on veut réaliser des systèmes qui ont des performances au moins analogues aux nôtres. Ils doivent montrer leur capacité dans tous les domaines où il est utile d'être intelligent. Et, bien qu'on ne cherche pas à copier l'intelligence humaine, il est souvent bon de s'en inspirer.

Les chercheurs en intelligence artificielle ont deux défauts : ils sont trop intelligents et pas assez paresseux ; du coup, ils font une trop grande partie du travail que devrait faire le système d'intelligence artificielle qu'ils développent.

D'un autre côté, l'intelligence artificielle est le problème le plus difficile auquel l'homme s'est attaqué, mais celui-ci n'est peut-être pas assez intelligent pour le résoudre.

Ces différentes idées mènent Jacques Pitrat à la conclusion qu'il faut s'aider des systèmes d'intelligence artificielle eux-mêmes pour les mettre en œuvre de la façon la plus efficace possible et donc utiliser des processus d'amorçage[(5)].

La conscience est utile pour comprendre le fonctionnement de l'être humain et elle est utile pour l'être humain ; elle a donc également toutes les chances d'être utile pour l'intelligence artificielle. Dans ce domaine, on peut s'inspirer du fonctionnement de l'homme mais ne pas se contraindre à l'imiter ; on pourra éventuellement aboutir à de nouvelles méthodes de mise en œuvre de conscience ayant peut-être des possibilités supérieures aux nôtres. Une cognition artificielle serait donc différente de la cognition humaine car certaines capacités cognitives artificielles sont inaccessibles aux humains (et vice versa). (Les ordinateurs ont un gros avantage en mode sériel mais beaucoup moins en mode parallèle qui est la base du fonctionnement de nos processus inconscients).

La conscience réflexive (capacité de s'observer et de s'analyser en train de fonctionner) permet d'analyser les différentes étapes d'un raisonnement pendant que celui-ci a lieu et donc, grâce à un méta-raisonnement d'évaluer constamment la direction à choisir ; elle donne ainsi des informations sur les actions effectuées et sur les raisons pour lesquelles elles ont été accomplies, permettant, d'une part, de justifier les plans choisis, de s'adapter à de nouvelles situations en se modifiant dynamiquement et, d'autre part, d'acquérir de nouvelles connaissances en analysant les raisons d'un succès ou d'un échec. Dans CAIA (sigle pour *Chercheur Artificiel en Intelligence Artificielle*) par exemple, Jacques Pitrat analyse les mécanismes informatiques utilisés dans les cas de succès et constate que ceux qui relèvent d'une analogie avec la conscience, lorsqu'ils sont présents, sont nécessaires à la résolution du problème.

La conscience morale (au sens anglais de *conscience*) est nécessaire pour être autonome : elle donne des informations sur le caractère bon ou mauvais, adéquat ou inadapté des décisions prises (comme le choix des buts, des actions, et les conséquences de celles-ci). Comme elle se fonde sur des principes généraux, elle est nécessaire pour s'adapter efficacement à des situations nouvelles en donnant des heuristiques pour choisir les buts à atteindre, les actions à effectuer et les conséquences de celles-ci. Bien entendu, une telle conscience n'exerce son discernement que par rapport à des règles de conduites qui peuvent être appliquées ou non et non par rapport à des lois auxquelles on ne peut se soustraire. Par ailleurs, les machines peuvent faire fi de certaines de nos limitations : leur conscience réflexive peut être ajustable (elles peuvent observer tout de leur fonctionnement), elles peuvent être clonées à faible coût (ce qui est intéressant pour tester divers modes d'apprentissage, en fonction de différences clairement identifiables) ; enfin, en ce qui concerne la conscience morale, les machines

[(5)] Utilisation d'une version n d'un logiciel pour engendrer une version n+1 ; celle-ci va servir à engendrer une version n+2, qui à son tour... (traduction de l'anglicisme « *bootstrapper* »).

peuvent être considérées comme esclaves et, comme nous l'avons déjà souligné, la survie n'étant pas leur valeur essentielle, elles peuvent prendre de grands risques.

Certains de ces aspects sont présents dans CAIA, un système général de résolution de problèmes donnés sous forme de contraintes ; ce système, réalisé par Jacques Pitrat, est décrit dans [13]. Détaillons comment ces éléments sont mis en œuvre dans la cognition artificielle de CAIA.

3.2. CAIA (Chercheur Artificiel en Intelligence Artificielle)

Quatre aspects de la cognition artificielle donnent un avantage énorme aux systèmes artificiels : leur capacité à observer leurs connaissances, leur possibilité de s'observer en cours de fonctionnement, leur aptitude à faire une méta-combinatoire bien plus considérable que celle qu'un humain peut faire, la facilité avec laquelle on peut les dupliquer.

Tous ces points se fondent sur la déclarativité des connaissances, explicites ou procédurales, et à différents niveaux méta. Ainsi, puisque chez Pitrat tout est représenté sous forme déclarative, la notion de conscience réflexive développée par Pitrat se ramène-t-elle à une gestion efficace en mémoire des diverses représentations de soi, des autres et du monde. Un autre point fondamental qu'aborde Pitrat, c'est l'évolution de ces connaissances. Les mécanismes d'acquisition de ces connaissances, leurs transformations en processus compilés et l'évaluation de leur intérêt donne également une place cruciale à la notion de conscience morale.

Nous verrons également le rôle fondamental des différents niveaux méta dans la gestion de ces connaissances, ce qui rappelle la phrase de Kant [9] : « *La conscience est une représentation qu'une autre représentation est en moi* » qui souligne le rôle clé de ces niveaux méta dans la connaissance de soi, des autres et du monde.

3.2.1. *La conscience réflexive*

Analyser ce qu'on sait. — Les humains ne savent que très peu de choses sur ce qu'ils savent (cela explique d'ailleurs les difficultés qu'on a rencontrées à réaliser des systèmes experts, dont les niveaux méta sont particulièrement faibles).

Au contraire, CAIA a une conscience effective de toutes ses connaissances car elles sont exprimées sous une forme aussi déclarative que possible. La forme déclarative des connaissances les rend plus faciles à analyser, à créer ou à modifier (par exemple par un simple changement de la valeur d'une variable). Mais, elles sont sous une forme figée, passive. Pour que le système puisse agir, il faut les transformer en connaissances procédurales, la forme active correspondante.

Les traitements qui indiquent comment traiter des connaissances déclaratives (méta-connaissances) et qui permettent cette traduction sont explicités également sous forme déclarative puis transformés en connaissances procédurales, grâce à ces mêmes méta-connaissances.

Ainsi, CAIA transforme toutes ses connaissances déclaratives en programmes C. Jacques Pitrat n'a pas écrit une seule ligne des 450.000 lignes de C qui le constituent actuellement. Afin de ne pas avoir à traiter une complexité inaccessible, ces programmes se sont développés à partir d'eux-mêmes depuis 25 ans (tout cela a été « *bootstrappé* »). Cette conscience totale de son savoir lui permet par exemple :

- d'écrire des programmes combinatoires efficaces quand il ne sert à rien d'être intelligent.
- de chercher des symétries dans l'énoncé d'un problème.
- d'optimiser les programmes qu'il crée en méta-évaluant leur exécution.
- de décider a priori comment il va utiliser une méthode de résolution particulière.

Mais il ne suffit pas d'observer, il faut comprendre ce que l'on observe pour pouvoir l'utiliser.

S'OBSERVER EN TRAIN DE FONCTIONNER. — La plus grande partie des mécanismes intellectuels humains sont inconscients (on sait, par exemple, quels raisonnements on a suivis mais on ne sait pas pourquoi on les a considérés, de même pour les analogies qui émergent « comme ça »). En outre, nous n'avons aucun moyen de rendre tout ça conscient.

Au contraire, CAIA gère une simple pile des appels de fonctions ; il peut ainsi rendre consciente toute étape de n'importe lequel de ses modules : chacun d'eux sait ce qu'il fait, peut savoir dans quel état il est quand il s'arrête et peut repartir sans que cet arrêt l'ait perturbé (au contraire des humains qui sont dérangés par une interruption et ne peuvent reprendre l'exécution telle quelle). Il a ainsi la possibilité de détecter des erreurs ou des anomalies, de contrôler les bouclages ou les explosions combinatoires... Il peut décider de se modifier dynamiquement et utiliser immédiatement les modifications qu'il a éventuellement décidé d'effectuer.

Tout cela peut se faire aussi bien en cours de traitement qu'à la fin d'un traitement (ce qui est très difficile – pour ne pas dire impossible – pour les humains parce que notre mémoire est insuffisante et que nous ne pouvons reconstituer exactement la situation passée).

Grâce à ces représentations complètes de ses exécutions, CAIA peut créer une trace de ce qu'il fait et pourquoi il le fait. Il engendre ces représentations systématiquement, ce qui lui permet d'expliciter une explication et une méta-explication. L'explication justifie la solution tandis que la méta-explication indique comment on est arrivé à penser à faire les étapes utiles aussi bien qu'à éviter les étapes inutiles : elle contient la séquence des méta-connaissances qui ont servi à sélectionner les connaissances qui ont été utilisées au cours d'une exécution.

MÉTA-COMBINATOIRE. — Pour chaque méthode, CAIA dispose de déclencheurs qui disent si elle peut être pertinente, des conditions qui vont l'interdire et des priorités qui vont décider de l'urgence de son utilisation.

Au lieu de faire uniquement de la combinatoire sur les valeurs possibles des variables, CAIA fait aussi de la combinatoire sur les méthodes possibles pouvant faire progresser la résolution.

Cette méta-combinatoire a une grande supériorité sur la combinatoire : pour justifier une solution, il n'est pas nécessaire de donner toute l'arborescence engendrée, mais seulement les méthodes qui ont servi à trouver cette solution. CAIA fait ce choix automatiquement pour créer une explication.

Donnons un exemple pour comparer la combinatoire et la méta-combinatoire :

Problème : trouver tous les nombres m et n positifs et inférieurs à 10^{18} tels que :

$$4 * m + 3 * n^2 = 817.401.078.957.542.034$$

Avec la combinatoire, on peut considérer toutes les valeurs de n permises et vérifier si on trouve un m solution (*10^{18} étapes*).

Avec la méta-combinatoire, il suffit de considérer 10^9 étapes (car il faut que $n^2 < 10^{18}$.

Mais CAIA fait encore mieux grâce au raisonnement suivant : $3*n^2$ est pair donc n^2 est pair donc n est pair donc n^2 est multiple de 4 ; le résultat de $4*m + 3*n2$ doit être multiple de 4, ce qui n'est pas le cas de 817.401.078.957.542.034, il n'y a donc pas de solution. De plus, cela s'applique aussi au cas où m et n sont des entiers compris entre $-\infty$ et $+\infty$ alors que la combinatoire ne peut être utilisée.

Bien sûr, il est des cas où la méta-combinatoire ne fait pas gagner de temps (et parfois même en perdre), mais elle garde une très grande supériorité sur la combinatoire : il n'est pas nécessaire de donner toute l'arborescence engendrée pour justifier une solution, il suffit de préciser les méthodes qui ont servi à la trouver. Ce que fait effectivement CAIA pour créer une explication.

3.2.2. *La conscience morale*

L'autonomie est donc une caractéristique essentielle des systèmes intelligents et pour qu'ils puissent s'enrichir de leur propre expérience, nous avons vu qu'une conscience réflexive était nécessaire. Mais, pour un tel programme qui fonctionne de façon de plus en plus autonome, il est impossible de prévoir tous ses comportements ; il faut donc des mécanismes de contrôle extrêmement élaborés. Pour éviter les dérives non souhaitables et faire respecter les valeurs et les principes qui fondent l'ensemble du système, il faut donc qu'il y ait un module robuste qui surveille le tout, ce que l'on pourrait assimiler à une sorte de conscience morale. Ce module devrait suivre d'une part des valeurs intangibles et d'autre part des principes généraux qui retardent ou interdisent des opérations trop dangereuses (comme la suppression définitive d'éléments

dont la création et l'adaptation ont demandé beaucoup de travail et qui se sont révélées pendant longtemps comme très efficaces)[6].

Une des questions difficiles qui n'est pas résolue est que ces valeurs et principes sont évolutifs chez l'homme, dépendent de la culture et peuvent être considérés à trois niveaux : pour l'individu, pour le groupe ou pour l'humanité entière. À la différence, ces valeurs sont fixes chez les machines (et il est d'ailleurs fortement souhaitable qu'elles le restent et que ces systèmes autonomes n'aient pas accès à ces valeurs) et peuvent être différentes des valeurs de l'homme.

3.3. Immortalité

Pour l'informatique, il est facile de reproduire un système. D'une part, il peut être très commode d'avoir plusieurs exemplaires d'un système efficace pour les tester dans des situations très diverses. Cela permet de disposer de plus de temps pour progresser, mais aussi de prendre des risques, débouchant éventuellement sur un apprentissage plus efficace.

D'autre part, cela a d'importantes conséquences sur la cognition artificielle : on peut aussi créer des copies légèrement différentes, ce qui permet de comparer les importances relatives de ces variations. On peut ainsi adapter la personnalité d'un système au type de tâche dont il est chargé : on conserve la copie la plus efficace dans un contexte donné.

Il en résulte qu'un système artificiel est quasiment immortel.

4. Conclusion

Grâce aux avantages de la cognition artificielle, que nous avons détaillés ci-dessus, l'intelligence artificielle a un potentiel de développement énorme. Mais, deux raisons font qu'il n'est pas sûr qu'il pourra être atteint :

(1) Le travail à faire est extrêmement difficile et il n'est pas certain que l'homme soit assez intelligent pour y arriver car c'est probablement le problème le plus difficile auquel il s'est attaqué.
(2) La structure de la recherche ne privilégie pas les recherches qu'il faudrait faire, c'est-à-dire expérimenter des projets ambitieux. Une raison est que l'on exige des chercheurs de beaucoup publier. Or il n'est pas facile de faire beaucoup de publications quand on réalise un système nécessitant beaucoup de travail ingrat qui ne mérite pas d'être publié. L'intelligence artificielle est une science expérimentale qui demande beaucoup d'implémentations complexes et nécessite de nombreux tests.

[6] Isaac Asimov a abordé cette question dans les années 1950 dans *Les Robots* [2]. Il a proposé ses lois de la robotique afin de gouverner des systèmes artificiellement intelligents. Une grande partie de son travail a ensuite été consacrée à tester les limites de ses lois. Comme il arrive à construire systématiquement des situations où les lois sont mises en défaut, son travail suggère qu'aucun ensemble de lois fixes ne peut anticiper suffisamment toutes les circonstances possibles.

Par ailleurs, quand on réalise un très gros système, il faut dominer ce qu'il contient. Cela exige de pouvoir y consacrer au moins 50 % de son temps. Les universitaires devraient développer ce genre de recherches dont les résultats sont pour le très long terme. Combien d'entre eux ont la possibilité matérielle de pouvoir consacrer 50 % de leur temps à leur recherche, alors qu'ils sont débordés par leurs enseignements ou leurs tâches administratives (quand ce n'est pas les deux) ?

C'est pourquoi Jacques Pitrat espère que l'informatique et les autres disciplines des sciences cognitives apporteront une aide inattendue. Il croit que l'amorçage de l'IA est la seule solution possible : les réalisations passées aideront à mettre en œuvre les réalisations futures. Et on a vu l'importance des fonctionnalités de la conscience pour cet amorçage.

Mais, il estime que ces espoirs ne se réaliseront qu'à très long terme (*une perspective à 100 ans*, comme il l'exprime dans le n° 100 du bulletin de l'AFIA [5] !).

Bibliographie

[1] F. Anceau, *Vers une étude objective de la conscience*, Hermès Science Publications, 1999.

[2] I. Asimov, *Les Robots* (titre original : *I, Robot)*, Gnome Press, 1950 (traduction française 1967).

[3] R. Barthes, *Introduction à l'analyse structurale des récits*, Communications, Paris, 1966.

[4] A. Bassi, « Un modèle dynamique de la compréhension de texte intégrant l'acquisition des connaissances », Thèse, Paris 11, 1995.

[5] Bulletin de l'AFIA, n° 100, `https://afia.asso.fr/wp-content/uploads/2018/11/100_avr18.pdf`.

[6] A. Cardon, *Modéliser et concevoir une machine pensante. Approche constructible de la conscience artificielle*, Automates intelligents, Paris, 2003.

[7] J. Eccles, *Évolution du cerveau et création de la conscience*, Fayard, Paris, 1992 (traduction française de *Evolution of the brain: Creation of the self*, Routledge, New York, 1989).

[8] G. Edelman, *Biologie de la conscience*, Éditions Odile Jacob, Paris, 2008 (traduction française de *Bright Air, Brilliant Fire: On the Matter of the Mind*, New York : Basic books, 1992).

[9] E. Kant, *Critique de la raison pure*, 8 éd., Bibliothèque de Philosophie contemporaine, PUF, 1975 (traduction française de *Kritik der reinen Vernunft,* Riga, verlegts Johann Friedrich Hartknoch, 1781).

[10] G. A. Miller, « The magical number seven, plus or minus two: Some limits on our capacity for processing information », *Psychological Review* **63** (1956), n° 2, p. 81-97.

[11] J. Pitrat, *Textes, ordinateurs et compréhension*, Eyrolles, 1985, traduit en anglais : *An artificial approach to understanding natural language.* North Oxford Academic (Grande-Bretagne) et GP Publishing (USA) 1988.

[12] ———, *Métaconnaissance, Futur de l'intelligence artificielle*, Hermès, Paris, 1990.

[13] ———, *Artificial Beings: The conscience of a conscious machine*, Wiley-ISTE, London, 2009.

[14] B. Russell, « On propositions: What they are and how they mean », *Aristotelian Society Supplementary Volume* **2** (1919), p. 1-43.

[15] G. Sabah, « CARAMEL: A computational model of natural language understanding using a parallel implementation », in *Proceedings of 9th European Conference on Artificial Intelligence (ECAI-90)*, Pitman, London/Boston, 1990, p. 563-565.

[16] ———, « The respective roles of conscious and subconscious processes for interpreting language and music », in *Proceedings of 8th International Workshop on the Cognitive Science of Natural Language Processing (CSNLP-8)* (Amsterdam), Advances in Consciousness Research, John Benjamins, 1993, p. 241-253.

[17] G. Sabah & X. Briffault, « CARAMEL: a Step towards reflexion in natural language understanding systems », in *Proceedings of IEEE 5th International Conference on Tools with Artificial Intelligence (ICTAI '93)*, IEEE Computer Society Press, Washington, 1993, p. 258-265.

[18] G. Sabah, P. Coiffet et al., *Vers une technologie de la conscience ?*, Communication de l'Académie des technologies, EDP sciences, 2013.
[19] R. Schank, « The structure of episodes in memory », in *Representation and understanding: Studies in cognitive Science*, Academic Press, New York, 1975, p. 237-272.
[20] R. Schank & R. Abelson, *Scripts, plans, goals and understanding*, N.J. Lawrence Erlbaum, Hillsdale, 1977.
[21] J. Searle, *The rediscovery of mind*, Cambridge University Press, 1992.

Abstract. — For many years, researchers in artificial intelligence and cognitive sciences have refrained from approaching the theme of consciousness, a notion considered as too imprecise to allow scientific study. Thus, Bertrand Russell [14] claimed the results of introspection scientifically unusable because they do not obey physical laws. Likewise, behaviourism, wanting to found psychology as an exact science, excludes any notion of mental state, and rejects what concerns consciousness as fundamentally outside its domain. A revival of this question is due to Darwinian theory of evolution and orthodox materialism. On the other hand, Searle [21] finds it scandalous that a science studying the mind ignores the specificities of consciousness. The main problem then is to explain feelings, consciousness and free will based only on physical laws. After having clarified some general ideas on intelligence and consciousness, I will explain why, working on natural language processing, this lead me to be interested in consciousness. Then, I will describe the aspects of consciousness that Jacques Pitrat has implemented in CAIA (Artificial Researcher in Artificial Intelligence).

Keywords. — Artificial consciousness, natural language processing, problem solving.

Manuscrit reçu le 21 février 2021, révisé le 17 octobre 2021, accepté le 30 octobre 2021.