Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

# THESIS / THÈSE

**MASTER EN SCIENCES INFORMATIQUES**

**Formalisations informatiques du comportement éthique**
**Justification et explication dans les systèmes multi-agents**

Elskens, Thomas

*Award date:*
2021

*Awarding institution:*
Universite de Namur

Link to publication

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 07. Jul. 2025

UNIVERSITÉ DE NAMUR

Faculté d'informatique

Année académique 2020-2021

# Formalisations informatiques du comportement éthique

# Justification et explication dans les systèmes multi-agents

Thomas Elskens

Promoteur : Nathalie Grandjean

(Signature pour approbation du dépôt - REE art. 40)

Mémoire présenté en vue d'obtenir le grade de
Master en Sciences Informatiques

# Résumé

Au premier chapitre, l'éthique des machines, discipline nouvelle, qui a pour objet le calcul du comportement éthique, est présentée. Ses principales problématiques et réalisations sont exposées, avec une attention particulière pour la justification de l'action. L'intérêt de la justification est double : d'une part, sur le plan conceptuel, elle joue un rôle dans la motivation et l'interprétation de l'action ; d'autre part, sur le plan technique, la justification requiert à la fois une approche sous-symbolique et, dans sa qualité de discours intelligible, une approche logique.

Le deuxième chapitre présente en profondeur le paradigme multi-agents : il scrute ses différentes manifestations technologiques où, malgré leur apparente diversité, le mémoire s'attache à retrouver les caractéristiques qui en font l'unité. La diversité éclate toutefois dans le traitement de l'environnement : alors que celui-ci disparaît dans une technologie comme FIPA ACL, qui repose sur la communication de messages entre agents, il prend une place prépondérante dans la simulation à base d'agents, autre manifestation du paradigme. Le cas de la simulation est particulièrement intéressant, dans la mesure où la technique se double d'une ambition explicite de production de connaissance. Nous nous penchons sur la nature et la fonction de cette connaissance : la mesure dans laquelle elle peut servir dans une justification, la menace qui plane sur elle d'être prise pour un simulacre, dépourvu d'exigence propre.

Le troisième chapitre s'intéresse de près au potentiel des systèmes multi-agents dans le calcul des aspects éthiques du comportement, ceci à 3 niveaux – intrapersonnel, interpersonnel, impersonnel – et d'un point de vue tant téléologique que déontologique. Y sont abordées en détail les questions de la motivation, du jugement éthique en situation, de la réputation et de la confiance, de la négociation des fins, de l'internalisation de la norme, les différentes façons dont la norme peut constituer une épreuve à laquelle notre comportement doit s'affronter, mais aussi les modalités de création, de diffusion et d'apprentissage de la norme. Ce chapitre se clôt sur l'analyse de trois cas : d'abord une étude de deux œuvres de science-fiction, œuvres dont les auteurs sont aux prises avec différentes déclinaisons d'intelligence distribuée et centralisée ; ensuite une étude des outils de modélisation à l'aide de la prise de décision en développement durable ; enfin un accident de la route à issue mortelle impliquant une voiture autonome. Ce dernier cas permet d'aborder le rôle de la simulation à base d'agents dans une perspective d'innovation responsable. Les trois cas s'intéressent tous au temps : sous la figure de l'anticipation d'un côté, la prévision de l'autre, le traitement du temps en SBA peut prétendre à un rôle dans une optique d'éthique orientée vers le futur.

# Abstract

In the first chapter, the new discipline machine ethics, whose object is the computation of ethical behavior, is presented. Its main problems and realizations are discussed, with a particular focus on the justification of actions. The significance of justification is twofold: on the one hand, on the conceptual level, justification plays an important role in motivating and interpreting actions; on the other, on the technical level, justification requires a subsymbolic approach and, while it is meant to be a human-readable discourse, a logical approach, too.

The second chapter yields a detailed presentation of the multiagent paradigm: the chapter unravels its manifold technological manifestations, in which, despite their apparent diversity, this memoire is committed to retrieving the characteristics warranting its unity. Diversity, however, seems unavoidable when it comes to handling the environment: indeed, environment almost disappears in a technology like FIPA ACL, based on the communication of messages between agents. On the contrary, it becomes of outmost importance in agent-based simulation, another manifestation of the paradigm. Simulation is even a particularly striking example of it, in so far as technics is coupled with an explicit ambition to produce knowledge. We have a closer look on the nature and the function of this kind of knowledge: how can it serve the end of justification and the way it is threatened by the accusation of simulacrum, devoid of any authentic imperative.

The third chapter devotes close attention to the potential of multiagent systems in computing ethical aspects of behavior, at 3 levels – intrapersonal, interpersonal, impersonal – and this both from teleological and deontological viewpoints. Are discussed here questions like motivation, situated ethical judgment, reputation and confidence, negotiating means-ends, norm internalization, different means for the norm to be a test to which our behavior should be submitted, in addition to the modalities of norm creation, diffusion and acquisition. Finally, the chapter culminates with the analysis of three illustrative cases: first, the analysis of two works of science-fiction, in which the authors grapple with different variations of distributed versus centralized intelligence; then, a study of modelling tools to help with decision-making in the field of sustainable development; and lastly, a lethal traffic accident involving an autonomous car. This last case also enables us to tackle the role of agent-based simulation in the perspective of responsible innovation. All the three cases have in common their concern for time: be it in the form of anticipation or prevision, time-handling in agent-based simulation can claim a role in a future-oriented ethics.

*À Nathalie Grandjean, qui n'a pas hésité à accepter un sujet aux contours pourtant si incertains.*

*Aux amis fidèles, J.D et J.B, qui plus d'une fois, de jour et de nuit, nous ont généreusement secouru lorsque la volonté de continuer risquait de faire défaut.*

*Je ne t'apporte pas des roses,*

*Car je n'ai pas touché aux choses,*

*Elles aiment à vivre aussi.*

Charles VAN LERBERGHE

# Introduction

Ce mémoire ne manque pas, à sa manière, d'une certaine ambition, celle de conjuguer deux fils thématiques qui, dans l'opinion commune, sont réputés difficiles à nouer ensemble. En même temps en effet, nous visons, d'une part, un thème informatique, les systèmes multi-agents, et d'autre part, un thème relevant de la réflexion philosophique sur la place de l'informatique dans la cité des hommes. D'une part, une présentation rigoureuse d'une technologie particulière ; d'autre part, une réflexion sur les tenants et aboutissants de celle-ci en matière éthique.

Ce mémoire s'inscrit dans une réflexion sur la technique, dont l'importance philosophique a été rappelée, au XX^e^ siècle, par Martin Heidegger[1], selon qui la technologie est le propre de l'homme. Mais le père de la philosophie contemporaine regarde la technologie encore comme un ensemble de *moyens*, construits par l'homme, pour arriver à certaines *fins*, également posées par l'homme. Cette conception étroite va progressivement s'élargir pour trouver, dans les études des sciences et technologies, sa définition la plus vaste : la technologie n'est pas faite que de ses objets, mais de l'ensemble des pratiques, systèmes de savoirs et de faire, relations sociales rendus possibles par ceux-ci. Entre ces deux conceptions – l'une étroite, l'autre très large – une leçon fondamentale a été retenue : l'homme donne naissance à des outils, après quoi ceux-ci finissent par *refaçonner* (*shape*) leur géniteur[2].

Ce mémoire se veut également – et tout autant – une réflexion à portée éthique. C'est loin d'être un sujet commode : il suffit de remplacer le terme d'éthique par son synonyme quasi-parfait « morale » pour nous convaincre de la profonde ambivalence de notre temps à l'égard des exigences que l'éthique incarne. Perdus que nous sommes entre un désir du risque zéro promis par le contrôle absolu et les aspirations à la liberté, entre l'homogénéisation induite par la mondialisation et le souhait – souvent sincère – de vivre dans une société plurielle, l'acuité de nos contradictions fait bégayer le discours éthique. De façon plus fondamentale, un certain héritage logiciste, qui nous a appris à trancher entre *valeur* et *fait*, renvoie l'éthique à la sphère des idéalités : belles peut-être, mais superflues à celui qui se trouve aux prises avec le réel. Le discours relativiste, véritable signe du temps, prompt à intoxiquer n'importe quel débat comportant un volet éthique, ne se prive que rarement de cette arme de pointe.

Et pourtant, l'exigence éthique garde de tout temps une sorte d'évidence, car oui, plusieurs choix se présentent à nous, et oui, le « seul vrai » choix n'existe probablement pas. Mais la pluralité – réelle ! – des options ne doit pas faire oublier que tous les choix ne se valent pas, et que nous sommes tenus, dans nos choix, à toutes les rigueurs de la sincérité, de la cohérence et, plus généralement, de la

[1] Cité dans D. G. JOHNSON, *Computer Systems*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 170-171.

[2] Le propos est du philosophe canadien Herbert Marshall McLuhan, qui pense d'abord aux nouvelles technologies qui ont permis aux médias de masse de prendre leur essor, dont au premier chef la radio et la télévision. Cité par A. K. MACKWORTH, *Architectures and Ethics for Robots*, dans *op. cit.*, pp. 345-347.

Raison. Le choix éthique doit engager toutes nos facultés ; or l'engagement passionné que l'éthique réclame est aux antipodes de l'indifférence, du repli ou du chemin de la moindre résistance.

Tout au long de ce mémoire, nous parlerons donc éthique et technique, nous essayerons de dire les spécificités de l'une comme de l'autre : l'éthique – nous le savons depuis l'aube des temps – a ses exigences ; or depuis au moins Simondon, nous avons pris conscience que la technique n'est pas moins dépourvue de ses exigences propres, qui réclament tout le sérieux et toute la rigueur dont notre raison peut se montrer capable, sous peine d'inefficacité et par là, de non-sens : car quelle absurdité serait une machine qui ne « marcherait » pas, qui ne ferait pas ce pour quoi elle a été conçue !

Ce qui précède ne doit cependant pas induire le lecteur en erreur : si l'éthique et la technique importent, chacune de son droit propre, il ne faudrait pas en conclure que nous nous contenterons, dans ce mémoire, de *juxtaposer* les exigences de l'une et de l'autre. Car si la technique est le propre de l'homme, elle dit à ce titre quelque chose sur l'homme, sur ce qu'il peut et doit faire, sur les possibilités de son être-au-monde, elle ne saurait être neutre éthiquement : son message doit interpeller quiconque s'intéresse à l'éthique. Parcourant la même interdépendance dans le sens inverse, nous devons encore dire que les réalisations techniques, ou même seulement les aspirations techniques de l'homme, font partie intégrante de la question « que dois-je faire ? » : elles peuvent – et doivent – faire l'objet de la délibération éthique sur les choix qui se proposent à l'homme. Car en technique, comme dans d'autres sphères de l'agir humain, tous les choix ne se valent pas.

Le « problème moral » nous interpelle depuis toujours ; le « problème technique » est peut-être plus récent ; il n'en est pas moins important pour autant, ni moins urgent. Pour ne citer qu'un seul exemple du progrès fulgurant des techniques ces quelques dernières années : au concours DARPA en 2004[3], la meilleure voiture autonome tombait en panne après seulement 14 kilomètres. Huit années plus tard seulement, tant Google que Tesla pouvaient déjà se faire fort d'avoir conçu des voitures capables de parcourir de façon autonome des centaines de milliers de kilomètres ! De tels exploits de l'agir humain doivent, *a priori*, nous réjouir : espérons que la réflexion sur leur *bonne* insertion sociale leur emboîtera prestement le pas.

## i) Vous avez dit « éthique » ?

Le terme « éthique » s'utilise dans des acceptions diverses et variées. Afin de guider notre recherche, précisons ce qui, pour nous et en première approximation, désigne ce vocable. Au sens plein et premier, nous y voyons une *tension* entre un *contexte*, une *situation*, et finalement une *valeur*. Le contexte se comprend ici au sens large : tout ce qui constitue l'environnement dans lequel la situation a lieu. Il faut y inclure toutes les contraintes familiales, légales, tous les devoirs « moraux » qui nous

[3] Le *DARPA Grand Challenge* est une compétition organisée par la DARPA (agence états-unienne en charge de projets de recherche militaires), dont les protagonistes sont des véhicules terrestres pleinement autonomes. Le circuit du concours se trouve dans le désert des Mojaves, en Californie. Signalons par ailleurs qu'un tableau reprenant tous les acronymes se trouve à la fin du mémoire.

incombent, mais aussi nos croyances, nos idéologies et nos illusions. La situation est ce sur quoi l'éthique s'exerce, que ce soit un jugement éthique, ou une décision à prendre. Si le contexte présente quelque généralité, la situation est toujours particulière.

La valeur, finalement, est ce au nom de quoi l'acte éthique va être posé. Dans le sens riche de l'éthique auquel nous adhérons ici, cette valeur renvoie toujours, en dernière instance, à une *image de l'homme*. Par cette notion d'image de l'homme – *ethos* dans un vocabulaire hérité d'Aristote – nous retrouvons l'étymologie du mot éthique. Par les valeurs auxquelles il adhère et qui guident ses actions, l'homme se dévoile. Pour bien faire sentir l'importance de l'ethos, disons encore que nous croyons – c'est une valeur à laquelle nous tenons – que l'homme est, au moins en partie, un être *rationnel*. En tant qu'être rationnel, les connaissances qu'il acquiert au sujet du monde et sur soi affectent à leur tour cette image qu'a l'homme de lui-même. Il en va de même de l'œuvre de ses mains, des productions diverses que l'homme donne au monde : par l'exploration du génie technique qui lui est propre, l'homme ajuste pour ainsi dire constamment son ethos.

Nous avons fait état d'une *tension*, c'est-à-dire qu'il ne suffit pas d'appliquer une règle (générale, relevant du contexte) à une situation (particulière) de façon automatique. Pour tout dire, ce n'est qu'au moment où il y a *conflit* entre normes, entre valeurs, qu'il est possible de parler d'un moment éthique. Un tel moment requiert un choix entre plusieurs possibilités, qu'il n'est possible de trancher qu'en invoquant une valeur. Toute possibilité de choix n'est donc pas nécessairement éthique, si le choix n'implique pas, n'engage pas, l'homme.

Un tel moment éthique est exemplifié par le choix que doit faire Antigone quand elle s'oppose à Créon : faut-il, ou non, accomplir les rites mortuaires dus à son frère, Polynice ? Dans cette situation, il y a conflit entre la loi de la Cité, défendue par Créon, et les lois religieuses qui exigent des rites. Antigone pourrait céder à la peur, *esquiver* le moment éthique, mais si elle décide finalement de braver la loi des mortels pour accomplir la volonté des dieux, c'est en raison de l'image qu'elle a d'elle-même en tant que sœur : l'homme n'est lui-même que s'il s'inscrit dans une chaine d'êtres qui, résistant à l'épreuve du temps, accède sinon à l'immortalité, sinon à la nécessité, au moins au sens. Finalement, Antigone ne prend pas parti *contre* la loi de la Cité – elle ne veut pas être une mauvaise citoyenne – mais elle se décide *pour* son devoir familial. Au moment où la valeur est choisie, la situation s'éclaire par un jour nouveau, agir autrement ne serait plus pour Antigone que lâcheté : la valeur donne couleur et relief, un sens et un horizon, au moment éthique. Notre propre personne est engagée, et cet engagement que tout moment éthique implique nous apprend autant – voire davantage – sur nous-mêmes que sur la chose à juger ou la décision à prendre, comme si les différents traits de la situation finissaient par donner forme à notre propre portrait.

La valeur est informante. Si elle n'est peut-être pas ce qui donne sa structure à l'expérience, du moins elle possède indéniablement une vertu herméneutique, elle fournit une clef d'interprétation. En cela, elle s'oppose à toute approche exclusive, tel un certain logicisme pour qui l'offense majeure sera toujours le manquement à la vérité, même en dehors du débat académique. Nous pouvons avoir plusieurs valeurs, certes, mais tout au plus une viendra éclairer le moment éthique. L'accent mis sur la valeur – nous voulons connaître l'homme par la valeur qu'il se donne – a quelque chose de lévinassien, si nous pouvons interpréter ainsi son propos de voir l'éthique comme la discipline la plus fondamentale de la philosophie.

Être éthique, c'est d'abord une question de discernement : il faut voir les moments éthiques, voir la multitude de choix possibles, et puis faire un choix en fonction d'une valeur plutôt que par calcul… à moins, bien sûr, d'avoir l'intime conviction que l'homme, *in fine*, est un être calculateur. Mentionnons à ce propos un sens dérivé de l'éthique, là où l'éthique devient principalement affaire de précaution et de prudence :

> *[…] l'échec de l'éthique réside dans le présupposé même devant lui assurer une valeur opératoire et une reconnaissance institutionnelle, la théorie de la prudence et du choix rationnel, l'évaluation du risque et une analyse coût/bénéfice. […] l'approche éthique n'est qu'une extension du processus de rationalisation propre aux sociétés industrielles.*[4]

L'éthique, dès lors, est une technique sociale pour accompagner l'innovation ; une technique née dans une société qui a appris que l'homme est capable de tout, et du pire en premier. Être « éthique », au sens premier du terme, implique dès lors bien souvent de prêter attentivement l'oreille à ce que nous dit « l'éthique », dans ce sens plus restreint.

Quelques remarques sur les contours des « êtres » impliqués dans cette conception de l'éthique s'imposent. Traditionnellement – première remarque – l'éthique est conçue de façon individualiste : c'est l'individu qui est placé devant des choix, qu'il opère en faisant référence à des valeurs plus ou moins partagées, mais dont l'horizon est finalement toujours assez court. Il ne doit pas en être ainsi : le contexte peut faire référence aux savoirs et réalités sociales, économiques, politiques et écologiques. La situation peut être un cas de notre vie quotidienne, mais elle peut aussi se présenter à une collectivité, à l'occasion d'une grande décision à prendre qui va influencer la vie de ceux qui viendront « après nous ».

Deuxième remarque, cette conception de l'éthique soulève une question importante : n'est-elle pas fondamentalement anthropocentrique ? L'homme est-il seul à faire l'expérience de moments éthiques ? Et ses valeurs doivent-elles forcément renvoyer à une image de lui en tant qu'il est homme ? Même si les vues exposées ici n'excluent pas les collectivités, voire des entités non-humaines, il est vrai que celles-ci seront toujours vues comme agissant en quelque sorte « par procuration », ou comme ayant de la valeur par rapport au sens que l'homme leur donne. Ainsi, si l'homme protège la nature, c'est parce qu'il ne veut pas avoir de lui l'image d'un parasite. Il est possible de considérer qu'il s'agit là d'une faiblesse de la conception, mais en dernière analyse, nous pensons que l'homme ne peut créer du sens que par rapport à soi.

## ii) Quelle éthique pour une machine ?

Le premier chapitre examinera à la loupe la discipline nouvelle qui a pour nom « éthique des machines » (*machine ethics*). Caractérisons-la – brièvement – comme l'effort de calculer les aspects éthiques du comportement, que celui-ci soit humain ou robotique. En effet, comme nous le verrons,

[4] A.-M. Rieu, *Face aux nanotechnologies*, dans D. Parrochia et V. Tirloni, *Formes, systèmes et milieux techniques après Simondon*, p. 223.

même si la visée première de la discipline est pratique – offrir les meilleures garanties possibles quant aux comportements de robots agissant à proximité d'êtres humains – une visée secondaire est d'étendre le champ de la calculabilité au comportement humain. Là, une distinction supplémentaire pourra encore être faite entre le comportement humain réel – visée descriptive – et un comportement humain idéalisé – visée prescriptive.

Nous commencerons le chapitre par plusieurs mises au point notionnelles et définitions, notamment en situant l'éthique des machines par rapport à une discipline bien mieux connue, nous voulons dire l'éthique de l'informatique (*computer ethics*). Nous donnerons quelques exemples de ses premières réalisations, avant de nous engager plus en avant dans une discussion sur son objet : qu'est-ce qui, au juste, doit – ou peut – être calculé ?

L'éthique des machines – pour des raisons que nous verrons – se cantonne délibérément au comportement ; étant une discipline issue d'une approche d'ingénieur, son souci premier est l'opérationnalisation des concepts éthiques afin de guider les processus d'automatisation. Aussi tout un ensemble d'aspects traditionnellement rapportés à l'éthique sont-ils d'emblée exclus de l'analyse. Nous en analyserons les conséquences pour la conception éthique qui peut s'en dégager, car, toute opératoire qu'elle soit, à elle s'imposent les mêmes exigences que nous sommes en droit d'attendre de toute conception éthique quelle qu'elle soit. Anticipant un peu les résultats de cette analyse, nous pouvons déjà annoncer que le calcul du seul comportement observable ne suffit pas : les raisons qui guident la machine à agir comme elle le fait doivent faire partie des livrables de l'éthique des machines. Ces raisons – qui s'inspirent d'arguments éthiques classiques, le plus souvent de nature déontologique ou utilitariste – doivent faire partie intégrante du calcul.

La calculabilité dont il est question ici présente ainsi un double visage : elle préside à l'action ; elle la justifie, aussi, dans la mesure où le calcul doit pouvoir rendre compte du comportement de la machine. En tant que moteur de l'action, elle ressemble à l'équivalent machinal d'une motivation. En tant que justification, elle constitue un certain type de discours, dont nous aurons à analyser les modalités. Fidèle à la conception de l'éthique que nous venons d'esquisser, nous nous pencherons plus particulièrement sur le rôle de la valeur dans la justification.

Signalons, pour terminer l'annonce du premier chapitre, que chaque section s'y termine par un ensemble de questions : cette façon de procéder s'est imposée à nous assez naturellement, dans la mesure où la matière – tout aussi vaste qu'elle est nouvelle – se laisse ainsi concevoir comme une matière à exploitations diverses. Quelques-unes de ces questions recevront un début de réponse dans les chapitres ultérieurs ; d'autres, en revanche, resteront essentiellement ouvertes, faisant signe par là même à de riches gisements, encore à prospecter.

## iii) Les systèmes multi-agents

Ce mémoire n'a pas pour ambition de dire quelque chose de la technologie en tant qu'ensemble sinon homogène, du moins suffisamment cohérent pour être synthétisable. Nous nous limiterons à une technologie particulière, celle dite des systèmes multi-agents : elle fera l'objet de notre deuxième chapitre. La notion de système multi-agents est assez difficile, car chacun de ses

termes – *système*, *multi*, *agent* – résonne de diverses manières dans le concert des savoirs et des pratiques. À cela s'ajoute que les pratiques informatiques qui se disent multi-agents sont elles-mêmes diverses. À première vue, leur seul point commun est négatif, celui de l'abandon d'une intelligence centrale unique. Chaque « agent » a quelque latitude pour prendre ses propres décisions, pour gérer ses propres perceptions du monde. À partir de là, les applications de l'idée multi-agents sont vastes ! Dans le monde physique d'abord, la notion est appliquée dans la robotique : citons l'exemple de Stanley, un système de contrôle de voiture autonome qui a remporté la compétition DARPA en 2005[5]. Le système multi-agents se fait conducteur de voiture, agit directement sur le monde réel à partir de ses percepts « personnels ». Dans le monde logiciel ensuite, des agents se chargent de gérer le trafic aérien. Ici, à la différence du premier cas, les agents informatiques ne pilotent pas (encore) eux-mêmes les avions, mais se contentent d'émettre des plans de vol continûment ajustés en fonction des vicissitudes climatologiques et plus largement environnementales[6]. Dans le monde virtuel, enfin, nous trouvons les simulations « à base d'agents » (ou SBA). Elles sont de plus en plus populaires en sciences, y compris dans celles qu'on désigne parfois par « humaines » ou « sociales ». C'est sur cette dernière forme des systèmes multi-agents que notre attention se portera tout particulièrement.

Un exemple éclairant d'une simulation à base d'agents est donné par Virginie Mathivet[7] : elle propose de montrer comment un phénomène collectif tel un banc de poissons peut être formé sur la base de quatre règles simples, règles exécutées par chaque agent individuellement : première règle – l'océan étant modélisé par une grille – il faut éviter de dépasser les limites du monde virtuel ; deuxième règle, s'il y a un autre poisson dans la zone très proche, il faut s'en éloigner ; de même en présence d'un obstacle (troisième règle). La quatrième règle, enfin, édicte que s'il y a un poisson à distance moyenne, l'agent poisson s'aligne sur lui. Ces règles, pour rudimentaires qu'elles soient, suffisent à voir apparaître un banc de poissons aux mouvements harmonieux… si harmonieux que nous pourrions croire que les individus qui composent le banc agissent de concert, comme s'ils étaient animés d'une seule et unique intention de faire route ensemble. Les moyens mis en œuvre par l'exemple du banc de poissons sont très simples. Ainsi, le flux de contrôle décentralisé est ici seulement mimé[8]. Pourtant, une telle illustration souligne déjà à sa manière les points forts de la simulation : étude des comportements individuels inscrits dans le temps et l'espace *en les reproduisant* dans un environnement virtuel. Cet environnement virtuel peut faire l'objet d'une visualisation à l'écran et déjà ainsi révéler des informations à celui qui le scrute.

La SBA innove donc par rapport à mainte autre forme de simulation. Outre la gestion explicite du temps et de l'espace que nous venons de voir, ajoutons encore que les comportements individuels incorporent volontiers une part d'aléatoire. Ainsi, au fil des exécutions d'une même simulation, les résultats obtenus peuvent varier stochastiquement : l'explication des régularités s'affranchit ainsi des rigidités déterministes, selon des modalités qu'il conviendra d'examiner en temps voulu. Prises

[5] M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 99-101.
[6] G. WEISS, *Multiagent Systems*, pp. 463-467.
[7] V. MATHIVET, *L'Intelligence Artificielle pour les développeurs*, pp. 365-383.
[8] L'exemple étant conçu à des fins didactiques, le recours à des raccourcis d'implémentation abonde. Ainsi, chaque poisson a accès à l'ensemble des positions de ses congénères, alors qu'il n'est censé « voir » que les poissons les plus proches (le code se trouve *ibid*., p. 377).

individuellement, aucune de ses caractéristiques, gestion décentralisée, gestion de l'espace et du temps, place faite à l'aléatoire, ne distinguent pourtant uniquement la SBA ; c'est leur co-occurrence à l'intérieur d'une même méthode qui constitue son originalité. Elle fait le pari que certains phénomènes sont à saisir dans un contexte ou milieu singulier, où la rencontre plus ou moins fréquente, plus ou moins fortuite, entre certains types d'éléments, de particules ou d'agents, donne lieu à des phénomènes qui dépassent les individus considérés. Ainsi la SBA nous renseigne sur le devenir d'un phénomène ; elle enquête sur ses origines et les raisons de sa stabilité ; plutôt qu'à s'en tenir à des moyennes de moyennes – apanage des méthodes statistiques – elle cherche à nous renseigner sur des configurations spatiales ou, à tout le moins, interprétables par application d'une métaphore spatiale. Les questions que les scientifiques posent à ces simulations sont toujours singulières, fluctuent au gré des domaines d'études et des problèmes à explorer.

Quel intérêt, finalement, des systèmes multi-agents dans un mémoire portant sur des questions éthiques ? Sans trop vouloir anticiper sur nos développements, prenons un autre exemple qui a déjà fait couler beaucoup d'encre, celui des insectes sociaux et des colonies de fourmis en particulier. La vie d'une fourmilière a été magistralement décrite par Maurice Maeterlinck[9]. Tout au long de son ouvrage, l'auteur décrit en détail l'étonnante organisation d'une fourmilière et l'intelligence collective de ses membres. Une question obsède l'auteur, qui n'a de cesse de s'interroger sur l'organisation de la fourmilière : *qui* décide ? *qui* détermine combien de reines il faut, combien d'ouvrières, combien de mâles, et comment répartir les tâches entre ouvrières ? comment, en cas de danger, l'alarme est-elle donnée de façon quasi instantanée, comme une fulgurance ? Ainsi, au fil des pages, l'auteur s'interroge sur l'intention planificatrice qui semble prévaloir dans la fourmilière, sur ce qui fait son *unité*, sur sa *finalité* :

> *Nous retrouvons ici [dans la fourmilière] le grand problème de la ruche et de la termitière. Qui règne et qui gouverne dans la cité ? Où se cache la tête ou l'esprit, d'où émanent des ordres – qui ne sont jamais discutés ? […] la fourmilière devrait être considérée comme un individu dont les cellules, au rebours de celles de notre corps qui en compte environ soixante trillions, ne seraient plus agglomérées mais dissociées, disséminées, extériorisées, tout en restant soumises, malgré leur apparente indépendance, à la même loi centrale.*[10]

Et l'auteur pousse même la réflexion plus loin, allant jusqu'à s'interroger sur le destin des colonies de fourmis :

> *Jusqu'où iront-elles [les fourmis] ? Sont-elles à leur apogée ou déjà sur le déclin, comme pourraient le faire craindre les ferments étrangers et morbides que sèment les parasites dans leurs meilleures républiques ? Ont-elles un autre avenir devant elles ? Qu'attendent-elles ? Voilà des millions d'années qui n'ont pas compté, par conséquent des milliards de milliards de vies et de morts qui n'ont pas compté davantage.*[11]

[9] M. MAETERLINCK, *La vie des fourmis.*
[10] *Ibid*., pp. 22-23.
[11] *Ibid*., pp. 179-180.

L'interrogation sur la finalité et le destin aboutira finalement sur la question de la valeur : si destin il y a, *vaut-il* d'être poursuivi ?

> *Qu'est-ce qui compte enfin ? Ont-elles [les fourmis] atteint leur but et quel est donc ce but ? Si la terre, la nature, l'univers n'en ont pas que nous puissions entrevoir, pourquoi en auraient-elles, pourquoi en aurions-nous un ? Naître, vivre, mourir et recommencer jusqu'à ce que tout disparaisse, n'est-ce pas suffisant ? Quelqu'un ouvre un œil dans la nuit, voit un coin de terre ou de mer, quelques étoiles, une face humaine, puis le referme pour toujours. De quoi se plaindrait-il ? N'est-ce pas ce qui nous arrive ? Tout, ne fût-ce qu'une seconde, ne vaut-il pas mieux que de n'avoir pas été ?*[12]

Ainsi, en s'interrogeant sur la fourmi, son organisation, sa vie en société, son devenir au fil du temps, Maeterlinck s'interroge aussi, par procuration, sur l'homme. Puissions-nous, à notre manière, en faire autant, sonder les profondeurs de l'homme en nous penchant sur ses sociétés, ses œuvres, ses ambitions et ses désirs. Cependant, notre interrogation se réclamera d'une méthode inaccessible à un homme du début du XX^e^ siècle : faire voir les possibilités de la simulation à base d'agents pour dépasser la simple observation, pour restituer une certaine épaisseur du réel à l'écran, contrôlant ce même réel pour mieux l'expliquer, tel sera un objectif de ce mémoire.

A l'intérieur du paradigme multi-agents, nous nous intéressons principalement à deux technologies phares : les moteurs d'inférence décisionnels « BDI » (pour *Beliefs, Desires, Intentions*), d'une part, les simulations à base d'agents, d'autre part. Un tel choix peut surprendre, car à première vue tout sépare les agents très raisonneurs du BDI des réflexes comportementaux souvent très simples de la SBA : alors que les premiers sont utilisés pour venir à bout d'une complexité comportementale parfois vertigineuse, les deuxièmes servent souvent à mettre en lumière des phénomènes complexes émergeant d'interactions rudimentaires. La SBA, en d'autres termes, est peut-être l'émanation du paradigme où le « multi » se révèle décisif. Nous consacrerons une partie conséquente du deuxième chapitre à mettre en lumière l'unité du paradigme, sous la diversité protéiforme de ses manifestations. Nous nous interrogerons, chemin faisant, sur la notion même d'un agent : qu'est-ce que cela veut dire d'être source ou origine d'actions ?

Le BDI trouve son champ d'application principal en robotique ; la SBA, en sciences sociales. Cette dernière utilisation invite pour ainsi dire naturellement à une réflexion sur la portée épistémologique du paradigme multi-agents, étant donné son ambition explicite en la matière : qu'est-ce que la SBA permet de connaître ? Anticipant quelque peu les développements du mémoire, la nature exacte des connaissances nous importe moins que leur potentiel à entrer dans une justification. Ce potentiel n'a rien d'évident : il doit être démontré. Notamment, il convient d'écarter le soupçon qui ne voit, en toute simulation, que simulacre, artifice qui court-circuite les exigences difficiles d'une science rigoureuse. C'est ici qu'une interrogation sur l'aspect « système » d'un système multi-agent peut se révéler utile, car qu'est-ce qu'un système sinon un espace ou un processus continu d'échanges d'informations constitutives ? Que font les fourmis, sinon s'échanger des signaux de toutes sortes en vue de coordonner leurs actions ?

[12] *Ibid*., p. 180.

Nous aurons ainsi à cœur de vérifier si la pensée systémique peut être de quelque secours à lever le soupçon du simulacre : cette vérification nous conduira à un effort de caractérisation rigoureuse de quelques notions servies à beaucoup de sauces, dont au premier chef celles de réduction et d'émergence, mais aussi la notion au fondement même de l'éthique des machines, la calculabilité formalisante, les interfaces entre celle-ci et les problèmes mondains qu'elle cherche à décrire.

Parmi les techniques de simulation, la SBA se distingue par une excellente prise en compte du temps et de l'espace : nous verrons sur quoi repose cette excellence. Même si nous devons remettre à un cas pratique du troisième chapitre une comparaison entre différentes techniques – et encore, celle-ci restera sommaire – nous nous intéresserons dès le deuxième chapitre à cette caractéristique de la SBA.

## iv) L'éthique des systèmes

Qu'il nous soit permis d'illustrer le paradigme multi-agents encore d'une autre façon, peu orthodoxe peut-être, mais très certainement éclairante. Nous voulons parler ici d'un cas classique d'algorithmie, le tri des éléments selon une relation d'ordre donnée. La façon traditionnelle d'aborder ce problème est un algorithme par insertion. Les éléments à trier sont alors stockés dans une collection – tableau ou liste – en suivant une règle très simple : l'élément est inséré dans la collection de façon à ce que celle-ci reste triée. Ainsi, lorsque le tableau comporte la séquence {1,5,10}, l'ajout de l'élément 8 donnera le tableau {1,5,8,10}. Dans cet exemple, l'intelligence du tri est *externe* aux éléments considérés. Du moment que les éléments se laissent docilement ramener, par la relation d'ordre, à un chiffre, leur individualité cesse de compter.

Le paradigme multi-agents, si jamais il était appliqué au problème du tri, nous inviterait à voir les choses autrement, un peu à la manière d'une cour de récréation pleine d'élèves, qu'un maître d'école voudrait ordonner selon leur taille. Notre invariant – une collection déjà triée – s'évanouit. Nous sommes ici en présence d'un ensemble bigarré et chaotique d'individus, qui peu ou prou ne voient pas plus loin que le bout de leur nez. En effet, ils n'embrassent pas la totalité de la cour du regard, mais n'en perçoivent que leurs voisins proches : si leur voisin de gauche est plus grand, ils changent de place avec lui ; si leur voisin de droite est plus petit en revanche, c'est avec lui qu'ils se livrent à la ronde. Il faut ajouter à ceci qu'*a priori*, les élèves sont libres de considérer d'abord leur voisin de gauche ou d'abord celui de droite ; ils peuvent même changer de préférence au cours de l'exercice ; le déterminisme ne se drape ici d'aucune vertu particulière.

La *perception limitée de l'environnement* n'est pas la seule caractéristique du paradigme. Ainsi, si les élèves ont l'habitude de cet exercice, ils *apprennent*. Notamment, avec l'expérience ils finissent par savoir s'ils se trouvent plutôt au début, au milieu, ou à la fin de la file. C'est dire qu'ils peuvent rendre leur agir plus efficace en adoptant un autre *comportement* : celui de gagner d'abord leur place approximative dans la file, avant de se lancer dans la comparaison minutieuse avec les camarades. Ils pourront encore *transférer* leur apprentissage sur d'autres domaines : ainsi lorsqu'il s'agira de trier un ensemble de pastèques selon leur taille, chaque élève pourra en prendre une et, pour la suite de

l'exercice, ne plus considérer sa propre taille mais au contraire celle de la pastèque, ou de tout autre objet, en fonction de *l'objectif* qu'il s'est donné.

Cet exemple, pour trivial qu'il soit, est cependant riche de renseignements, que notre troisième chapitre se propose d'explorer. Là où l'algorithme de tri se fonde sur une collection d'éléments quelconques, interchangeables, le paradigme multi-agents est utilisé dans des cas où le luxe de telles abstractions n'est plus tenable ; les agents dont il s'agit sont dotés de leurs facultés spécifiques, s'engagent dans une dynamique de groupe, sont régis par un cadre institutionnel précis, hors duquel leur comportement n'aurait guère de sens.

Nous venons de décrire un comportement organisé en vue d'une *fin*, en l'occurrence de produire une rangée d'élèves ordonnés selon leur taille. L'éthique, en effet, en tant que philosophie pratique, ne peut faire l'abstraction de cette caractéristique fondamentale de l'action, qui est de toujours se proposer un but à atteindre. Nous commencerons le troisième chapitre par une mise au point de cet aspect téléologique de la réflexion éthique, avant de plonger dans les efforts calculatoires des systèmes multi-agents en cette matière. Seront abordées ici, évidemment, la question de la motivation de l'agir, mais aussi des questions téléologiques qui engagent d'emblée les agents vivant en société : le jugement éthique, l'importance, pour l'agir, de la réputation et de la confiance, la négociation des buts à atteindre. Nous conclurons cette section sur les fins dont les organisations et les collectivités peuvent se révéler les dépositaires, plus ou moins indépendamment des individus qui les composent.

Une réflexion éthique ne peut cependant pas se cantonner à l'examen des fins dernières de l'agir : elle s'intéresse également de près au *comment*, aux façons de procéder que l'homme met en œuvre. Car si les fins que les hommes et les sociétés se proposent nous racontent beaucoup de choses sur eux, les moyens mis en œuvre pour y arriver nous en apprennent peut-être encore davantage. Nous continuerons donc notre enquête en examinant les travaux multi-agents consacrés aux aspects déontologiques de l'éthique : nous ferons état, dans cette section, de l'internalisation de la norme par les agents, des façons diverses et variées dont la norme peut peser sur leur comportement, des manières enfin dont les normes peuvent voir le jour et se répandre ; nous verrons également l'importance des collectivités dans cette vie des normes.

L'exemple de la cour de récréation met par ailleurs aussi le doigt sur l'importance de la mise en situation : en dehors de l'institution scolaire, il n'y a ni maître d'école, ni élèves ; c'est dire l'importance des institutions humaines. Il y a cependant plus : l'injonction de se ranger en file ne prend sens que dans certains lieux et moments spécifiques : à la cour de récré au moment où les cours reprennent, à l'arrêt de bus lorsqu'une classe part en excursion, etc. Le moment opportun est décisif : à 23h, le maître d'école ne pourrait donner aucune injonction de se mettre en file ordonnée… faute d'élèves ! Bref, une mise en situation ne se réduit pas aux agents en interaction – dans le cas de l'exemple, les élèves et leur maître d'école – elle est bien plus large, comprend l'espace et le temps, les institutions et les rôles dévolus à chacun en leur sein. Le troisième chapitre scrutera à la loupe ce contexte élargi, en interrogera la pertinence éthique dans le cadre des systèmes multi-agents et ce, sur les deux plans que nous venons de voir : téléologique d'une part,

déontologique de l'autre. Nous serons attentif aussi aux articulations entre les deux plans, articulations pour lesquelles nous nous baserons largement sur les analyses de Paul Ricœur.

Un mémoire en éthique ne saurait être complet sans la discussion de quelques cas illustratifs : c'est sur eux que se terminera le troisième chapitre. Dans le premier cas, nous nous arrêterons sur le temps de la technique. En effet, la réflexion sur la technique se veut orientée vers l'avenir ; cependant, cette ambition s'achoppe à l'impossibilité de ne rien dire de certain du monde à venir. Nous explorerons, dans ce cadre, une voie inattendue, qui est celle de la science-fiction. Nous commencerons par exposer les attentes légitimes qu'il est possible de nourrir vis-à-vis du genre, les perspectives qu'elle peut ouvrir et leurs implications éthiques. Puis nous examinerons deux œuvres qui, chacune à sa manière, nous disent quelque chose sur l'idée à la base du paradigme multi-agents, à savoir l'intelligence distribuée.

Si l'éthique est éminemment une philosophie pratique, éclairant l'homme sur le difficile problème de son action, la matière de notre deuxième cas s'attaque à la prise de décision, notamment en matière de développement durable. Le cas est intéressant à plus d'un titre, car loin de se cantonner aux décisions privilégiées par les moralistes traditionnels, il s'attaque de front aux problématiques ouvertes par le développement de la technique et de la pensée au XX$^{e}$ siècle : que faire, en tant que collectivité, pour assurer une vie, digne de l'homme, à notre postérité ? Nous aurons l'occasion de voir différentes méthodes de modélisation à l'œuvre – dont la simulation à base d'agents – et de comparer leurs mérites respectifs.

Un dernier cas prend pour objet un accident de la route impliquant une voiture autonome qui a coûté la vie à une piétonne. Nous examinerons le cas sous toutes les coutures – juridique, technique, informatique – avant de méditer ses implications pour une innovation responsable. Nous y verrons notamment l'importance capitale de la décision initiale d'automatiser (ou non) une activité donnée. Nous consacrerons également une section à nous interroger de quelle lumière le paradigme multi-agent peut éclairer ce genre de cas.

Voilà le programme de ce mémoire ; si le tracé en paraît de prime abord sinueux comme une trajectoire de fourmi, laborieux comme l'effort assidu de toute une colonie, nous espérons que le lecteur y fera, sous le fourmillement des idées et des observations, la même expérience de découverte que lorsque son regard émerveillé prend connaissance de la vie et de l'organisation d'une république myrmécéenne.

# Chapitre Ier. L'éthique des machines

*La machine est un geste humain déposé, fixé,*

*devenu stéréotypie et pouvoir de recommencement.*

Gilbert SIMONDON

## 1.1. Éthique des machines et éthique de l'informatique

Si l'éthique des machines est une discipline récente, l'éthique de l'informatique (*computer ethics*) est, elle, beaucoup plus connue. Il est dès lors intéressant de comparer les deux disciplines et de montrer en quoi elles sont différentes. L'éthique de l'informatique a pour objet d'étude *l'usage* des technologies informatiques et ce, sur deux plans. Le premier plan est celui de l'éthique appliquée : l'éthique de l'informatique se pose alors des questions telles que la protection de la vie privée, la fracture numérique, ou encore des questions de propriété ou de justesse des données[1].

Au deuxième plan, nous trouvons des préoccupations « macro-éthiques » : l'éthique de l'informatique s'intéresse alors aux entités et aux valeurs qui méritent un « statut » éthique. Contrairement aux familles macro-éthiques classiques que sont le contractualisme, le conséquentialisme, le déontologisme et l'éthique de la vertu, l'éthique de l'informatique cherche à déplacer le centre de gravité du regard éthique : plutôt que de se concentrer sur le sujet humain et l'inspiration ou les conséquences de ses actions, le regard de l'éthique de l'informatique se tourne vers l'objet technique en tant que tel, que celui-ci soit matériel comme une machine ou « immatériel »[2] comme une donnée ou un logiciel.

L'éthique des machines, en revanche, prend l'éthique elle-même pour objet d'étude, et se pose la question de savoir dans quelle mesure celle-ci peut être formalisée. Les visées sont tout d'abord pratiques : comment empêcher que des machines, de plus en plus indépendantes de l'homme dans leurs opérations, ne s'écartent des chemins balisés par un comportement éthique ? Elle a également un volet méta-éthique, dans la mesure où elle réfléchit sur la signification de ses formalisations pour

[1] Toute notre présentation de l'éthique de l'informatique est inspirée de L. FLORIDI, *Information ethics*.

[2] Clarifions tout de suite le sens de ce terme : aucune réalisation informatique n'existe sans support matériel, cela va sans dire. Cependant, le support matériel est insuffisant pour rendre compte des effets produits et opérations mises en œuvre par la machine informatique. Le *sens* de l'informatique se situe sur le plan des idéalités, non sur celui du support matériel (J.-M. SALANSKIS, *Le monde du computationnel*, pp. 142-147).

l'éthique. Même si l'éthique des machines est un champ relativement récent, les sujets de recherche se sont déjà considérablement diversifiés, comme nous allons le voir[3].

Ainsi, l'éthique des machines peut s'intéresser aux origines évolutionnaires du comportement éthique et, plus généralement, altruiste. Elle peut se pencher sur les manières d'éviter des usages abusifs ou des comportement immoraux de la machine. Elle peut implémenter – ou simuler – des théories de raisonnement éthiques, à des fins purement théoriques ou pour créer des conseillers éthiques. Enfin, les théories du raisonnement peuvent être traduites en programmes – ou modules – éthiques, servant à faire produire à une machine ses propres justifications éthiques, justifications qui à leur tour peuvent guider la machine dans ses actions à prendre.

Nous aurons l'occasion de revenir à ces différents champs de recherche dans les paragraphes qui suivent. Contentons-nous pour l'instant d'observer qu'en tant que production informatique, les implémentations de l'éthique des machines sont un sujet d'étude comme un autre pour l'éthique de l'informatique : autant dire que les deux disciplines sont étroitement liées, et qu'il est préférable de considérer l'éthique des machines comme une sous-discipline de celle, plus globale, de l'éthique de l'informatique. Ceci est d'autant plus vrai que selon ses partisans, l'éthique des machines devrait être déterminante dans le choix des systèmes autonomes autorisés à interagir avec l'homme[4].

| **Question** |
| --- |
| **Les SMA peuvent-ils modéliser (autre chose que) des objets techniques ?** |

## 1.2. Une approche fonctionnelle

Même si l'intérêt pour la formalisation informatique du comportement éthique se manifeste dès les années '80 du siècle dernier, il a fallu attendre le 21e siècle pour que cet intérêt débouche sur une démarche consciente d'elle-même. Le premier article à parler « d'éthique des machines » date de 2001, le terme sera ensuite repris par les époux Anderson en 2004. L'éthique des machines s'intéresse aux comportements des machines et des hommes, cette partie de la vie éthique qui se *donne à voir* à l'observateur externe. En cela, elle s'inscrit dans le programme éminemment pragmatique qui est celui de l'intelligence artificielle. La question fondamentale, qui structure son programme de recherches, est celle de savoir si la prise de décision éthique est *calculable*.

De prime abord, la question peut paraître excessivement ambitieuse. En effet, prendre en compte l'intégralité de la chaîne éthique a été qualifié de problème « IA-complet » (*AI-complete*), c'est-à-dire

[3] Pour cet aperçu, nous sommes redevable à Dr. McDermott, *What Matters to a Machine?*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, pp. 88-90.
[4] La revendication est formulée par M. Anderson et S. L. Anderson, *Case-Supported Principle-Based Behavior Paradigm*, dans R. Trappl, *A Construction Manual for Robots' Ethical Systems*, p. 167.

que sa résolution présuppose réglée l'entièreté des problèmes auxquels l'IA peut vouloir s'attaquer[5] : la collecte et la sélection d'informations pertinentes, la compréhension du langage naturel, la planification, etc. Il n'en demeure pas moins que plusieurs théories macro-éthiques (le déontologisme à la Kant et l'utilitarisme) répondraient en principe favorablement à la calculabilité du raisonnement éthique.

La démarche est dite « fonctionnelle », selon le terme de W. Wallach et C. Allen[6] : elle met entre parenthèses les questions métaphysiques sur les qualités qui font une personne, une entité morale… pour se concentrer – nous l'avons dit – sur le comportement. Prenant l'exemple du vol, ces auteurs soutiennent que tant les oiseaux que les avions volent, au même titre, même si *l'implémentation*, dans les deux *systèmes*, diffère du tout au tout. Les deux types de vol, cependant, sont fonctionnellement *équivalents* :

> *Flight is a functional property – it does not matter how you do it, so long as you get airborne and stay airborne for a decent amount of time. Because it is a functional property, flight can be manifested by a wide range of different systems made out of lots of different materials*[7].

David Gunkel a qualifié la démarche de kantienne[8] : il faut suspendre notre jugement pour laisser parler les réalisations. Voyons ce qu'être éthique « veut dire » pour les ingénieurs qui ont implémenté ces dernières. À ceci s'ajoute que la priorité donnée au comportement sur les facultés constitutives n'est pas simplement une affaire d'opportunisme. En effet, il y a également d'excellentes raisons de croire que nos facultés naissent de nos comportements et de nos interactions. L'idée a tout d'abord des antécédents philosophiques : il suffit de se rappeler comment l'abbé de Condillac explique l'émergence de la conscience de soi chez le petit enfant sous l'effet d'une double sensation[9] : par le tact, en touchant des objets, des corps, il fera dans un premier temps l'expérience de l'étendue. Puis, quand il se touchera lui-même, il ne recevra pas seulement une réponse tactile de sa main, mais également de la zone de son corps qu'il a touchée. Ainsi progressivement il fera la différence entre *son* corps et le monde extérieur ; sa subjectivité et son expérience de l'étendue extérieure vont de pair. La conscience de soi naît donc d'un comportement particulier dans un contexte bien particulier.

Ensuite, le primat du comportement a également frayé son chemin dans le domaine de l'intelligence artificielle, chez les tenants de l'intelligence artificielle dite « incarnée » (*embodied*)[10] : celle-ci se démarque tant de l'intelligence artificielle symbolique que du connexionnisme en adoptant une vision élargie de la cognition. Plutôt que de se cantonner aux traitements informationnels de haut niveau, ce courant émanant de la robotique préconise de s'intéresser en priorité au contexte

---

[5] Dr. McDermott, *What Matters to a Machine?*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, p. 93.
[6] W. Wallach et C. Allen, *Moral Machines*, pp. 55-71.
[7] *Ibid.*, p. 67.
[8] D. J. Gunkel, *The Machine Question*, p. 75.
[9] D'après X. Papaïs, *Condillac. Traité des sensations*, dans L. Jaffro et M. Labrune, *Gradus philosophique*. Plus récemment, dans une veine toute condillacienne, H. A. Simon a soutenu la thèse qu'un enfant doit apprendre à associer les informations sensorielles qu'il reçoit en provenance de son environnement pour déterminer les effets de ses initiatives motrices et, ultimement, à apprendre à se proposer ses propres buts (*Les sciences de l'artificiel*, pp. 210-220).
[10] Notre présentation de l'IA incarnée se base sur celle qu'en donne C. Misselhorn, *Grundfragen der Maschinenethik*, pp. 27-30.

géographique et physique dans lequel le robot va devoir évoluer. Plutôt que d'assumer des mondes clos ou d'autres modèles simplifiés, le but de l'intelligence artificielle devrait être de permettre au robot de percevoir le monde réel, s'y mouvoir, y manipuler des objets, etc., afin de maximiser son interactivité avec celui-ci.

À cet endroit, il vaut la peine d'ouvrir une parenthèse sur les buts et visées de l'intelligence artificielle. La formulation traditionnelle de la discipline – faire adopter un comportement à une machine qui serait tenu pour intelligent s'il était le fait d'un être humain – est très prudente. Elle fait toutefois la part belle au « comme si », ce qui entraîne deux inconvénients : premièrement, cette appellation invite à voir l'artifice, *le faux*, derrière chaque réussite de la discipline, aussi belle qu'elle soit ; deuxièmement, cette chasse aux chimères est renforcée du fait que, dans nombre d'acceptions du terme, « l'intelligence » ne recouvre *pas* une propriété fonctionnelle, dont au moins certains traits essentiels se laisseraient capturer du dehors, dans le comportement auquel ils donnent lieu. Dans cette pensée commune, il ne suffit pas d'obtenir un résultat d'un certain niveau intellectuel pour que nous puissions être taxé d'intelligent : ainsi, produire un calcul rapidement en se servant d'une calculette, n'est pas, sous ce rapport, qualifiable d'intelligence. De même, trouver la réponse à une question sans raisonnement, mais simplement en puisant dans notre savoir encyclopédique sur le monde, ne sera pas considéré intelligent non plus. En définitive, ce qui compte dans l'intelligence, c'est la manière dont nous arrivons à un résultat, le *comment* prime en quelque sorte sur le *quoi*. Ajoutons à cela, pour couronner le tout, que le concept d'intelligence est finalement assez creux : nous avons tellement pris l'habitude de ranger tellement de choses sous cette bannière, nous ne savons même pas si c'est une faculté unifiée ou un amalgame de différentes facultés cognitives à la mode. D'où une profusion de discussions stériles sur la *vraie* intelligence, tout comme, d'ailleurs, sur la *vraie* intentionnalité, la *vraie* conscience, etc.

Si nous voulons faire droit aux avancées à maints égards vertigineux de la discipline, un autre regard s'impose, une vision de l'IA qui ne serait pas de simuler quelque chose qu'elle n'est pas. La proposition de Yuk Hui[11] est de voir l'IA comme la discipline qui crée un *environnement* dans lequel les êtres humains et les machines peuvent *interagir* au travers de *relations matérialisées*. Ces relations sont interobjectives, c'est-à-dire que la matérialisation de la relation crée un nouveau système – ou « milieu », pour parler en termes simondoniens. Cette conceptualisation permet d'accommoder l'aspiration de l'éthique de l'informatique de se déprendre d'une vue trop exclusivement centrée sur l'homme et ses actions. De par les questions abordées, de par le choix de la méthode aussi, l'éthique des machines touche donc de près au cœur même des préoccupations de l'intelligence artificielle.

Cela dit, le primat accordé au comportement ne trouve pas sa motivation principale dans des considérations théoriques, mais dans les préoccupations éminemment pratiques de ce courant de recherche, en provenance du monde des ingénieurs. À ce propos, il est intéressant de se rappeler les enjeux d'une démarche d'ingénieur. L'ingénieur n'est ni un philosophe – c'est un tropisme que de l'affirmer – ni non plus, un scientifique. Un point important qui distingue l'ingénieur du scientifique est son rapport au réel. Le scientifique construit un modèle du monde où les phénomènes obéissent

[11] Y. Hui, *On the Existence of Digital Objects*, pp. 151-161.

à des lois. Le scientifique ne s'intéresse qu'aux phénomènes où la relation de cause à effet peut se dissoudre dans un ensemble d'équations productrices d'équivalence. Sa démarche d'abstraction l'autorise ensuite à écarter les autres phénomènes comme contingents ou impurs.

L'ingénieur, en revanche, même s'il est grand consommateur des modèles fournis par le scientifique, est toujours aux prises avec le monde concret. Tout ce que le scientifique peut taire, l'ingénieur en fait le centre de ses préoccupations et de sa pratique[12]. Alors même que le scientifique, dans cette optique, ne s'intéresse qu'aux lois qui agissent sur des corps simples, éventuellement idéalisés, l'ingénieur adopte un mode d'explication qui n'est pas celui des lois, mais celui de l'agencement des différents composants au sein d'un mécanisme ou d'un *système*[13]. Nous aurons amplement à revenir sur cette préférence pour le système propre à l'ingénieur.

Ce paragraphe sur le fonctionnalisme ne saurait cependant être complet sans un survol des principales critiques qui lui ont été faites[14]. Tout d'abord, le primat du comportement amène la question de savoir si « ça marche ? »[15] : quels sont les critères en fonction desquels nous pouvons dire que le comportement est moralement réussi ? Un « test de Turing »[16] moral existe, dont il est cependant clair qu'il est trop discursif. Ensuite, le focus de l'éthique des machines porte sur les effets plutôt que sur les causes de l'action. Fort bien en soi, si ce n'est que l'action morale est presque toujours mesurée à l'aune de *ses effets sur l'homme*. L'approche fonctionnaliste, dès lors, est très anthropocentrique, le plus souvent de façon implicite. Enfin, une question pourtant centrale de la réflexion éthique a une grande incidence sur le comportement, sans paraît-il pouvoir être expliquée par lui ; c'est la question de la valeur : qu'est-ce qui compte ?

| Questions |
| --- |
| **Quels critères de succès sont utilisés par les SMA ? Comment savons-nous que leur modélisation « a marché » ?** |
| **Les SMA peuvent-ils nous apprendre quelque chose respectivement sur les causes et les effets de l'action morale ?** |
| **Les SMA peuvent-ils modéliser un système ?** |

[12] Sur cette qualification de l'ingénieur, voir I. STENGERS, *Cosmopolitiques I*, pp. 129-133.
[13] Cf. D. ANDLER, A. FAGOT-LARGEAULT et B. SAINT-SERNIN, *Philosophie des sciences II*, pp. 760-763.
[14] Les points faibles ont été récapitulés par D. J. GUNKEL, *The Machine Question*, pp. 83-87.
[15] Dans son ouvrage *Les Sciences de l'artificiel*, SIMON traite longuement de la question de savoir si « ça marche ? », question qui est au cœur de la pensée systémique. Simon y distingue les problèmes d'optimisation des problèmes auxquels font face les systèmes complexes : ceux-ci ne peuvent faire autrement que d'y apporter des réponses adaptatives, contextuelles, que Simon désigne par le néologisme *satisficing*.
[16] Un test de Turing fait référence à un procédé imaginé par le père de l'informatique, le mathématicien britannique Alan Turing : il s'agissait de déterminer l'intelligence d'un ordinateur en lui faisant tenir une conversation avec un être humain par écran interposé. Si l'interlocuteur ne parvient pas à décider s'il a affaire à un ordinateur ou non, l'ordinateur est considéré avoir réussi le test. Dans une variante du test, l'interlocuteur fait la conversation sur plusieurs écrans : un seul parmi ceux-ci est nourri par un ordinateur, les autres affichent les réponses d'interlocuteurs humains tout comme lui. Signalons que le test d'origine, lui aussi, a été critiqué pour son logocentrisme.

## 1.3. Comportements éthiques implicites et explicites

L'éthique des machines part d'une préférence du sens commun, celle d'un comportement machinal « précablé » (*hard-wired*), où le robot n'a d'autre *choix* que celui de bien agir. Ainsi un distributeur de billets ne délivre sa manne pécuniaire qu'à une personne dûment identifiée grâce à sa carte et son code secret. En outre, il ne lui permettra pas de retirer plus qu'elle ne possède, lui évitant ainsi l'endettement. Il s'agit là d'un comportement moral *opératoire* ou *implicite* : la programmation du distributeur est *contrainte* de telle façon qu'un mauvais comportement soit d'emblée impossible. Or cette vue traditionnelle est insuffisante lorsque nous considérons le progrès fulgurant de l'apprentissage automatique. Au fur et à mesure que le champ d'action des machines s'étend, il devient de plus en plus ardu de prévoir tous les cas de figures possibles. Un module éthique *généraliste* s'impose, où la prise de décision éthique se fonde sur des représentations *explicites*.

L'éthique des machines a dès lors pour objet des représentations éthiques explicites : au-delà de la contrainte implicite, un agent technologique pourrait agir en vertu de certaines normes éthiques appliquées à une situation particulière, potentiellement inédite. La norme éthique doit être énoncée en termes suffisamment généraux, sous peine de se confondre avec des normes dites de conception : celles-ci permettant d'évaluer un agent technologique sur le résultat attendu en vertu de sa conception, nous retrouvons donc là, en quelque sorte, les contraintes implicites. Notons cependant que les contraintes implicites et les représentations explicites ne doivent pas s'exclure. Au contraire, elles peuvent très bien être combinées.

Cette distinction entre critères implicites et explicites n'est pas exempte de critiques. On peut notamment lui reprocher d'être trop simpliste. Dans cette optique, Tamas Madl et Stan Franklin[17] proposent quatre niveaux de contraintes au lieu de deux. Le premier niveau est celui du *non-cognitif*, nous pourrions également l'appeler le niveau *matériel*, car les auteurs y rangent toutes les contraintes mécaniques. Par exemple, un robot d'assistance aux personnes âgées est construit de façon à n'avoir qu'une vitesse très limitée, ou à ne pas pouvoir exercer trop de force sur les objets à sa portée. Le deuxième niveau est dit *réactif* : un distributeur automatique n'est ainsi autorisé à dispenser des billets que lorsque l'utilisateur a introduit sa carte et le code correspondant. Le troisième niveau est *délibératif* : nous y trouvons les règles éthiques explicites qui peuvent contraindre le robot dans la résolution de problèmes. Le dernier niveau est celui où nous trouverions des métarègles éthiques, telles que l'impératif catégorique de Kant, que la machine pourrait utiliser pour éprouver la validité des règles qu'elle aurait apprises[18]. Parmi ces niveaux, il est facile de voir que les deux niveaux supérieurs, les niveaux délibératifs et métacognitifs, sont explicites, et que le niveau le plus bas, le non-cognitif, est implicite. Le niveau réactif se laisse cependant moins facilement réduire à cette dichotomie. Il s'avère que le niveau est très large : ainsi un robot pourrait être contraint d'abandonner une action s'il constate qu'elle entraîne une réaction émotionnelle négative de la part du « bénéficiaire » de son action. Un autre niveau intéressant est le métacognitif : alors que les trois autres sont « explicitement » destinés à réguler le comportement, à agir sur lui, le

[17] T. MADL et S. FRANKLIN, *Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, pp. 139-140.

[18] Les auteurs s'empressent de préciser que ce niveau reste tout théorique et que, pour l'heure, aucune implémentation n'existe.

niveau métacognitif a pour objet des représentations internes à la machine. Nous voyons ainsi apparaître une autre dichotomie, dont la dichotomie de départ ne peut rendre compte.

Il n'en demeure pas moins que, en proposant ce programme de recherches, l'EM innove. Pour s'en convaincre, il suffit de se rappeler les principales lignes directrices qui guident actuellement les chercheurs et fabricants en robotique[19]. Au Royaume-Uni, l'EPSRC – acronyme pour *Engineering and Physical Sciences Research Council* – a publié une série de principes en 2011, à la suite d'un atelier où étaient présents des acteurs de l'industrie, des juristes, ainsi que des chercheurs d'horizons divers. Ces principes adoptent une vue traditionnelle : la machine y est assimilée à un outil, tel un marteau, dont chaque usage peut être imputé à un être humain. C'est-à-dire que pour chaque usage, un agent humain responsable doit pouvoir être désigné. En 2012, un projet lancé par l'Union européenne, *Making Perfect Life*, aborde le problème du point de vue des données échangées. Le rapport final distingue entre machines de surveillance, entre machines « partenaires de communication » de l'homme, et finalement, des machines comme instances de calcul généralisé. La machine y est abordée sous son aspect de producteur et relai de données. Par voie de conséquence, les recommandations se situent surtout sur le plan de la protection des données, respect de la vie privée et de la transparence.

Citons encore MEESTAR, un outil d'évaluation des implications éthiques des systèmes sociotechniques (*Model for the Ethical Evaluation of Socio-Technical Arrangements*). Même si l'outil cible explicitement des applications d'assistance aux personnes âgées, son champ d'application peut être élargi. La grille d'analyse de MEESTAR fait état de bon nombre de points d'attention qui tombent hors d'atteinte non seulement de l'agent technologique lui-même, mais également de ses créateurs, et doivent faire l'objet d'un débat public plus large. En cela, une prise de position étroitement technocrate est certes évitée. Cependant, comme guide pour ceux qui sont confrontés dans l'immédiat à des choix d'implémentation, l'outil est d'une aide assez limitée.

Ce rapide survol permet de dégager quelques éléments importants : tout d'abord, les cadres conceptuels proposés favorisent largement les agents éthiques implicites, c'est-à-dire contraints par conception à ne pas pouvoir sortir de leur champ d'application initialement prévu. En d'autres termes, les trois autres niveaux qu'il serait possible d'exploiter ne sont guère pris en compte. Ensuite, les préoccupations éthiques que la machine suscite semblent indissociables de la manière qu'on a de la voir : selon qu'on la considère comme couteau suisse ou comme usine à données… C'est souligner qu'une mise en perspective éthique de la machine ne peut se passer de développer sa propre compréhension de son objet d'étude.

À la lumière de la précédente remarque, la démarche préconisée par l'éthique des machines prend tout son sens. En effet, partant du point de vue de l'ingénieur, elle se veut *interne* à l'objet étudié, elle développe sa compréhension éthique en prenant en compte la machine de l'intérieur.

## Questions

[19] Ces informations proviennent de Br. KRENN, *Robot: Multiuse Tool and Ethical Agent*, dans R. TRAPPL, *op. cit.*, pp. 13-17.

| |
|---|
| **Quels niveaux de contrainte les SMA peuvent-ils prendre en compte : non-cognitif, réactif, délibératif, voire métacognitif ?** |
| **Les SMA présentent-ils des biais – ou des préférences – en faveur d'une ou plusieurs de ces contraintes ?** |

## 1.4. Démarches descriptives et prescriptives

La section précédente a présenté les différents types de travaux sous un jour thématique, comme autant de sujets réunis dans un grand panier de fruits dans lequel un chercheur qui s'intéresse au champ pourrait venir piocher à sa guise. Une telle présentation homogène ne doit cependant pas faire oublier un clivage capital du point de vue éthique. Nous pouvons en effet discerner deux grandes familles de recherches en éthique des machines. L'une, remontant aux travaux de Robert Axelrod, est *descriptive* : la question de recherche centrale qui l'anime est de savoir si le comportement éthique *ordinaire* est calculable. Une deuxième famille, dont les Anderson sont les représentants actuellement emblématiques, part d'une inspiration *prescriptive* : sa préoccupation est de savoir si la prise de décision éthique *idéale* est calculable. Dans le projet de conseiller éthique dominent les travaux normatifs. Les travaux d'Axelrod et ceux qui s'intéressent à des cadres « cognitivement plausibles » pour y inscrire le comportement éthique, sont quant à eux d'inspiration descriptive.

Le dialogue entre ces deux types de travaux ne coule pas de source : s'y pose de façon aiguë le problème de savoir dans quelle mesure l'être humain peut servir de modèle pour l'éthique des machines. C'est un point souligné à mainte reprise par les Anderson :

> *It can easily be argued that the « ethical » values of most human beings are insatisfactory. Ordinary humans have a tendency to rationalize selfish, irrational, and inconsistent behavior. Wouldn't we like machines to treat us better than most humans would?*[20]

De fait, les auteurs de la veine prescriptive multiplient les arguments qui problématisent la pertinence de l'éthique ordinaire comme guide de l'EM. Ainsi le comportement éthique ordinaire serait entaché[21] d'irrationalité, d'égoïsme, d'esprit grégaire… notamment à cause de nos émotions, de notre manque de bons modèles à suivre, ainsi que de notre recherche de gratification instantanée (et interne) plutôt que différée (et externe).

Face à ce portrait désolant, la machine présente plusieurs avantages[22] : elle n'a ni intérêts propres, ni émotions. En cas d'urgence, elle fait preuve d'une rapidité de calcul époustouflante ; elle est très rigoureuse dans l'application des calculs, ce qui lui donne un avantage certain pour appliquer les normes avec impartialité. Certainement dans un calcul éthique dans la veine utilitariste, la machine pourrait faire bien mieux que l'homme pour maximiser l'utilité globale escomptée. Pour couronner

[20] M. ANDERSON et S. L. ANDERSON, *Approaches to Machine Ethics*, dans IID., *Machine Ethics*, p. 238.
[21] Voir S. L. ANDERSON, *How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better*, dans *op. cit.*, pp. 524-525.
[22] EAD., *Philosophical Concerns with Machine Ethics*, dans *op. cit.*, p. 166.

le tout, le programme informatique n'admet ni le flou ni la contradiction, et veillera donc à la complétude et à la consistance des règles éthiques. Bref, l'idée est que la machine peut mieux faire, agir plus éthiquement, que l'homme.

Parmi les travaux prescriptifs, outre le projet de conseiller éthique déjà cité et sur lequel nous aurons à revenir, le but visé est de concevoir un module éthique qui pourrait guider un robot dans un grand nombre de contextes d'interaction. Les considérations éthiques doivent permettre de poursuivre deux buts complémentaires : d'une part, interdire aux robots certains comportements inadmissibles ; d'autre part, guider leurs efforts vers les meilleurs effets possibles[23]. Accessoirement, un tel module générique rendrait plus comparables des choix d'implémentation éthiques entre différentes machines.

Du côté des travaux descriptifs, là encore, nous retrouvons deux types de travaux bien distincts. Ainsi, un premier groupe d'études s'intéresse aux comportements éthiques d'un groupe, aux interactions qui s'y observent. C'est, par exemple, l'approche prise par Axelrod, puis par Danielson[24]. Un autre ensemble de travaux, d'inspiration cognitiviste, s'intéresse à l'individu, aux raisonnements qu'il tient, réduisant le contexte et l'environnement à des paramètres d'entrée-sortie de son sujet d'étude. Ce type d'études, plutôt que de mettre en exergue la faiblesse de caractère du genre humain, s'arrête bien plus volontiers à l'extraordinaire complexité de l'éthique ordinaire. Ses tenants rappellent ainsi que la prise de décision requiert différents types de raisonnement[25] : qualitatif, analogique, par principes premiers, qui à leur tour présupposent la compréhension du langage naturel, l'application de lois et de contraintes, la planification, ainsi que l'optimisation de l'utilité sociale[26].

Afin de rendre ces tendances plus parlantes, nous allons voir un exemple de chacune dans les paragraphes qui suivent, en commençant par MoralDM, exemplaire de la démarche descriptive, avant de traiter un exemple de l'inspiration prescriptive.

### 1.4.1. MoralDM

MoralDM est un système qui formalise la prise de décision éthique chez l'être humain[27]. Les auteurs insistent sur le fait que leur système permet de mieux rendre compte des résultats obtenus par certaines études psychologiques que les explications utilitaristes habituelles. Leur propos est donc

---

[23] M. ANDERSON et S. L. ANDERSON, *Case-Supported Principle-Based Behavior Paradigm*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, p. 166. Si nous nous en référons à la classification de C. MISSELHORN (*Grundfragen der Maschinenethik*, pp. 46-47), le project de conseiller éthique s'inscrit donc clairement dans une perspective d'éthique appliquée, dans la mesure où les questions particulières d'importance sociale ou professionnelle, par opposition à une éthique normative générale, où la discussion porte en toute généralité sur l'adéquation éthique des règles, comportements ou sentiments, voire des traits de caractère.

[24] Voir respectivement R. AXELROD, *The Complexity of Cooperation* et P. DANIELSON, *Artificial Morality*.

[25] M. DEHGHANI, K. FORBUS, E. TOMAI et M. KLENK, *An Integrated Reasoning Approach to Moral Decision Making*, dans M. ANDERSON et S. L. ANDERSON, *op. cit.*, p. 422.

[26] Dr. MCDERMOTT, *What Matters to a Machine?*, dans *op. cit.*, p. 90.

[27] M. DEHGHANI, K. FORBUS, E. TOMAI et M. KLENK, *An Integrated Reasoning Approach to Moral Decision Making*, dans *op. cit.*, pp. 424-433.

résolument descriptif[28] : la *bonne* façon de raisonner est comprise ici comme celle qui correspond le mieux à ce que fait l'être humain. Les études psychologiques en question traitent du phénomène des « valeurs protégées », qui ont la singulière propriété de faire échec aux raisonnements utilitaristes. Elles donnent lieu à des réactions émotionnelles fortes, voire le refus d'envisager la possibilité de comparer les conséquences. En d'autres termes, elles entrainent une diminution significative de la sensibilité quantitative : la nature de l'action prime alors sur ses conséquences.

MoralDM est un moteur de raisonnement complexe, qui commence par soumettre tout nouveau dilemme à son module de compréhension du langage naturel (EA NLU, pour *Explanation Agent Natural Language Understanding*) afin de traduire de l'anglais (simplifié) vers un calcul des prédicats d'ordre supérieur qui comprend des opérateurs modaux. Le recours au langage naturel présente plusieurs avantages : non seulement il réduit le biais qui consiste à choisir la formalisation qui donne les meilleurs résultats, mais il réduit aussi les efforts d'encodage, ainsi que le risque d'erreur. Le niveau de langage compris par le NLU est assez élémentaire sur le plan syntaxique. Les efforts ont été concentrés sur l'analyse sémantique, qui se fait à deux niveaux. Au niveau de la phrase d'abord, ou toute ambiguïté potentielle reçoit un marquage, ce qui permet de différer la désambiguïsation d'interprétations sémantiques concurrentes ; au niveau du discours ensuite, où le contexte doit permettre de résoudre les ambiguïtés trouvées. Afin d'obtenir des représentations sémantiques non ambiguës, le module a recours à une base de connaissances basée sur une ontologie formelle de plus de deux millions de faits.

La « traduction » est ensuite analysée par un module de calcul qualitatif (OMR, pour *Orders of Magnitude Reasoning*) de l'ordre des grandeurs. OMR rappelle un peu les principes de la logique floue : les utilités peuvent être proches, comparables (plus grandes que), ou distantes. Plusieurs facteurs contribuent au calcul de l'ordre des grandeurs : premièrement, la présence ou non de valeurs protégées : la présence de valeurs protégées tire l'équilibre vers l'intervalle « proche ». Deuxièmement, le calcul va déterminer si l'effet escompté de l'action présentée dans le dilemme moral entraîne des effets négatifs sur l'agent ou le patient de l'action. Plus la causation du mal est directe, plus l'agent aura tendance à l'inaction et sa sensibilité à l'égard de l'utilité sera diminuée. Troisièmement, les utilités escomptées des différentes alternatives d'action sont également mises dans la balance.

En fonction de l'ensemble de ces facteurs, MoralDM va privilégier respectivement un style de raisonnement utilitariste ou déontologique. C'est le rôle du modèle FPR (pour *First-Principles Reasoning*), soit le module de raisonnements par règles. S'il n'y a pas de valeurs protégées (ou si l'ordre de grandeur est différent), le système opte pour un calcul utilitaire. Sinon, il va préférer une méthode déontologique qui peut violer la logique utilitaire.

Parallèlement au raisonnement par règles, MoralDM va d'ailleurs toujours essayer un raisonnement casuistique. C'est le rôle du module AR (pour *Analogical Reasoning*), soit le module de raisonnements par analogie. Chaque cas de base est structuré en entités, attributs et relations. Les cas de bases sont

[28] Si nous suivons la classification de C. MISSELHORN (*op. cit.*, pp. 46-47), le propos est plus précisément méta-éthique, dans la mesure où il s'attache à décrire des principes généraux de raisonnement éthique, plutôt que la description de normes particulières dans un contexte socio-historique donné et que l'auteure désigne d'éthique descriptive empirique.

ensuite projetés sur les cas cibles et reçoivent un score d'évaluation structurelle. S'il y a des éléments dans le cas de base qui manquent dans la cible, le système peut construite une inférence sur la base de la structure et ainsi « découvrir » une connaissance nouvelle. La combinaison des modules FPR et AR s'est avérée payante, car s'ils échouent l'un ou l'autre régulièrement à rendre une décision, ils n'échouent que rarement ensemble[29] !

Même si ces premiers résultats sont prometteurs, il faut bien garder à l'esprit que MoralDM ne peut comparer que des effets d'un même type (p. ex : vies contre vies), aucun mélange n'est possible. Les auteurs misent sur la prise de décision émotionnelle afin de remédier à ce souci. C'est dire que nous sommes très loin de la sensibilité prescriptive, dont nous allons voir un exemple au paragraphe suivant.

### 1.4.2. Le conseiller éthique

Après l'exposé de MoralDM, une section sur un conseiller éthique peut paraître surprenante à première vue : tout compte fait, MoralDM ne nous conseille-t-il pas sur des questions éthiques ? Il faut répondre par la négative : MoralDM ne donne pas de conseils, il simule ce que serait une prise de décision faite par un être humain. Nous aurons à revenir sur la simulation au prochain chapitre, retenons pour l'instant simplement que « simulation », dans ce contexte, veut dire « description » : MoralDM cherche à décrire – afin d'expliquer – le raisonnement éthique humain. Ce faisant, il ne s'interroge nullement sur la valeur éthique intrinsèque des avis qu'il émet : par exemple, il prend en compte des valeurs protégées quelles qu'elles soient. Tout ce qui l'intéresse, c'est de coller au plus près à l'évaluation que ferait un être humain de la même situation, tout en montrant les différents types de considérations prises en compte par ce raisonnement.

Les travaux liés au conseiller éthique, en revanche, s'attaquent à une autre problématique : ils cherchent à concevoir un système qui, face à un dilemme éthique donné, peut éclairer notre lanterne. Pour donner ces conseils, ils peuvent à l'occasion faire fi du raisonnement éthique humain ordinaire. L'approche, encore une fois, est prescriptive : le raisonnement appliqué se veut impartial, objectif, inspiré par des théories les plus généralistes et les plus universelles possibles, guidé par la seule raison. Le résultat de l'assistant est une séquence d'actions assortie de leur justification, disant *pourquoi* telle action est préférée à telle autre[30].

Quelles sont donc les théories dont ces conseillers s'inspirent ? Comment savoir qu'on a la « bonne » ? Cette question, de nature méta-éthique, est importante, mais non point absolue. Il est, en effet, parfaitement admissible d'obtenir des réponses contradictoires en consultant des conseillers différents, tant que les justifications individuelles produites par chacune d'eux soient

[29] Notons que les auteurs ne font état que de 8 cas de tests : si chaque module échoue à rendre une décision correcte pour 3 d'entre eux, l'autre module corrige toujours le tir. Les auteurs se montrent cependant confiants que des résultats similaires seraient obtenus en augmentant le nombre et la complexité des cas à traiter.
[30] H. SEVILLE et D. G. FIELD, *What Can AI Do for Ethics?*, dans M. ANDERSON et S. L. ANDERSON, *op. cit.*, p. 508.

cohérentes[31]. Le rôle premier du conseiller est de nous rendre attentifs à nos filtres subjectifs, nos inconsistances, notre ignorance – complaisante ou non. L'accent mis ici sur les exigences de répétabilité et de consistance s'accommode même de conseils déterministes, voire y voit un avantage. Cela explique aussi que l'absence d'émotions ou un manque de vécu soient valorisés positivement[32].

Les tenants de l'approche normative tiennent le conseiller éthique pour un but intermédiaire[33]. Un projet de conseiller éthique, qui ne décide rien lui-même, présente en effet certains avantages. Des avantages d'ordre pratique d'abord. Il est ainsi facile de limiter le champ d'application du conseil : le conseiller s'en tiendra à un thème ou un sujet bien spécifique. Ensuite, comme le conseiller ne fait que répondre à des questions, il peut facilement avoir accès à des données supplémentaires en demandant des compléments d'informations à son interlocuteur humain.

D'autres avantages sont plutôt d'ordre stratégique : comme le conseiller ne prend aucune décision par lui-même, il ne prend aucune responsabilité. Celle-ci revient toujours à l'homme. La responsabilité de ce dernier s'en trouve même accrue, puisque le recours au système d'assistant technique peut attirer l'attention du décideur sur des aspects auxquels il n'aurait *a priori* pas pensé lui-même[34]. Cet aspect est particulièrement prégnant dans le cas d'une théorie conséquentialiste[35] : la prise en compte des conséquences *engendre* plus de dilemmes éthiques qu'elle n'en résout, donnant ainsi lieu à un élargissement de la conscience éthique. La garantie ainsi fournie par le conseiller éthique contre la déresponsabilisation devrait par ailleurs faciliter l'acceptation de l'idée d'une moralité artificielle.

Enfin, certains avantages sont plutôt de fond : comme le conseiller éthique n'agit pas « sur le terrain », il n'est pas impliqué dans les conseils qu'il donne ; il est, par construction, externe à la situation à juger. L'avantage est, dès lors, qu'il est possible de faire l'économie sur le statut propre de la machine : la question de sa *valeur propre* ne se pose pas.

Dans les paragraphes à venir, nous examinerons en quelque détail un exemple de conseiller éthique conçu par les époux Anderson[36]. Le système se base sur la théorie des devoirs conditionnels (*prima facie*) de W.D. Ross. L'idée de la théorie est que les devoirs auxquels nous sommes soumis (par exemple, le devoir de tenir nos promesses) ne nous contraignent que tant qu'ils ne donnent pas lieu à des conséquences néfastes. La théorie combine ainsi les inspirations déontologiste et conséquentialiste ; toutefois, elle ne fournit pas de procédure de décision en cas de devoirs conflictuels.

---

[31] B. WHITBY, *On Computable Morality*, dans M. ANDERSON et S. L. ANDERSON, *op. cit.*, pp. 148-149.
[32] Notons que le déterminisme et le manque d'implication sont des choses bien différentes de l'objectivité, comprise comme une absence de préjugés. Ceux-ci peuvent provenir des programmeurs et/ou de l'apprentissage (cf. B. WHITBY, *loc. cit.*, p. 144).
[33] S. L. ANDERSON, *Machine Metaethics*, dans ANDERSON et EAD., *Machine Ethics*, pp. 23-26.
[34] H. SEVILLE et D. G. FIELD, *What Can AI Do for Ethics?*, dans *op. cit.*, p. 507.
[35] La remarque provient de H. SEVILLE et D. G. FIELD, *What Can AI Do for Ethics?*, dans *op. cit.*, p. 502.
[36] Tiré de leur article *A Prima Facie Duty Approach to Machine Ethics*, *dans op. cit.*, pp. 476-498.

Le contexte dans lequel opère le conseiller éthique est celui des soins : faut-il accepter qu'un patient refuse un traitement, même si on est persuadé que ce traitement puisse améliorer son état de santé ? Les devoirs, en l'occurrence, sont ceux de l'éthique biomédicale de Beauchamp et Childress, à savoir le respect de l'autonomie du patient, la non-malfaisance, et la bienfaisance. Afin de mesurer le degré de satisfaction de ces devoirs, une échelle quantitative simple est utilisée, allant de -2 à 2. L'échelle peut se lire comme suit : -2 est une violation forte du devoir, -1 une violation légère. Symétriquement, 2 est une satisfaction forte du devoir, 1 une satisfaction légère. 0 signifie que le choix est neutre par rapport au devoir considéré.

Le noyau de système est un moteur de programmation logique inductive. Les auteurs ont formalisé 18 cas d'une étude en bioéthique, dont les utilités ont été évaluées par des éthiciens. Quatre cas ont été fournis au système, après quoi celui-ci déduit un principe général qui permet de statuer sur les 14 autres cas. Une décision A prévaut sur une décision B si

1. la différence d'autonomie escomptée est supérieure ou égale à 3 ;

   OU

2. la différence de non-malfaisance est supérieure ou égale à 1 ET la différence d'autonomie est plus grande ou égale à -2 ;

   OU

3. la différence de bienfaisance est supérieure ou égale à 3 ET la différence d'autonomie est plus grande ou égale à -2 ;

   OU

4. la différence de non-malfaisance est supérieure ou égale à -1 ET la différence de bienfaisance est supérieure ou égale à -3 ET la différence d'autonomie est plus grande ou égale à -1.

Ainsi la programmation logique inductive a permis d'apprendre une règle : le professionnel de la santé devrait essayer de convaincre le patient de revenir sur sa décision si celle-ci n'est pas pleinement autonome et il y a soit une violation quelconque de la non-malfaisance, soit une violation sévère de la bienfaisance. Les auteurs conviennent du fait que ce principe n'a rien de révolutionnaire et ne vient qu'expliciter ce que des éthiciens savaient déjà plus ou moins intuitivement. Cependant, le principe n'a jamais reçu une représentation formelle et constitue donc, aux dires des auteurs, une avancée dans notre compréhension des règles éthiques, fût-elle modeste. Le principe généré doit être complet et consistent. S'il y a des incohérences, cela veut dire qu'il y a des désaccords entre éthiciens. Le formalisme permet de mieux cerner le désaccord et la machine ne devrait pas prendre des décisions dans ces cas. Ce noyau a été implémenté dans Ethel, un système de rappels de médicaments à prendre. Lorsque le système le juge opportun, il avertit un surveillant.

Quoique ce système soit un succès, sa portée est très limitée : le prototype a été nourri de tous les devoirs pertinents (et rien qu'avec eux). D'où la nécessité de concevoir un système plus général.

Plutôt que de recevoir d'emblée les devoirs et les traits éthiques pertinents, le système doit les demander auprès de son interlocuteur. Par « traits éthiques pertinents », les auteurs entendent tout ce qui permet de spécifier les utilités : l'effet de bien-être escompté de la prise d'un médicament, la durée après laquelle l'omission d'un médicament se fait sentir, etc. Comme procédure de décision, ce système général s'inspire de l'équilibre réflexif de John Rawls. Il s'agit d'une procédure cyclique : d'abord, on généralise à partir d'intuitions sur des cas particuliers, puis on mesure les généralisations à l'aune de cas concrets supplémentaires. On répète ces étapes jusqu'à ce qu'un principe soit trouvé qui respecte l'intuition tout en étant applicable au plus grand nombre de cas possibles, sans conflit restant. Si un nouveau cas se présente avec un autre jugement, le système signale l'incohérence potentielle à l'éthicien. Celui-ci doit alors soit revoir le jugement, soit fournir au système un nouveau devoir, trait, voire un nouveau degré d'utilité. Le nombre et le domaine des valeurs des paramètres sont essentiellement ouverts.

Au fil des cas, le système général évolue assez vite vers la même procédure que le prototype. Le système connaît toutefois une limitation de taille : dans un dilemme éthique, les traits éthiques en jeu peuvent présenter des différences soit quantitatives, soit qualitatives : une utilité a dès lors plus d'importance qu'une autre. Il faut alors, pour parler prosaïquement, une clef de pondération des traits éthiques. Or le système actuel ne peut gérer ce type de situation.

Nous retrouvons là une limitation très similaire à celle contre laquelle MoralDM se bute. Ceci n'a en soi rien d'étonnant, car la difficulté est de taille. Nous aurons à revenir là-dessus. Retenons, à ce stade précoce de notre analyse, qu'il convient d'être attentif à cette tension, et de toujours garder à l'esprit que les deux démarches peuvent être amenées à nourrir des attentes différentes par rapport aux systèmes multi-agents.

| Questions |
| --- |
| **La SBA peut-elle être utilisée tant à des fins prescriptives que descriptives ?** |
| **Comment est-elle pour l'heure utilisée ?** |
| **La SBA peut-elle modéliser des différences qualitatives dans les traits éthiques à comparer, ou uniquement quantitatives ?** |

## 1.5. Démarches ascendantes et descendantes

En éthique[37], il est possible d'adopter deux types de point de vue : soit descendant, soit ascendant. Dans la manière ascendante de voir l'éthique, une place prépondérante revient au contexte. Réagir éthiquement dans un contexte donné requiert du savoir pratique, une faculté de discernement, un ensemble de vertus en somme, qui ne peuvent être acquis que par l'habitude ou l'exercice. Certains cadres particularistes (s'ils évitent le relativisme), casuistiques ou basés sur une éthique de la vertu rentrent dans ce paradigme.

Le même point de vue se retrouve aussi dans un programme de recherche de l'intelligence artificielle que nous avons déjà rencontré, celui de la cognition dite incarnée : le comportement complexe ne naît pas d'une représentation centrale dans le cerveau, mais émerge à partir de processus de bas niveau. Dans cette optique, l'être humain est un système complexe, lui-même composé de sous-systèmes multiples[38]. Nous y retrouvons également les travaux évolutionnaires d'Axelrod et Danielson : la théorie des jeux permet alors de comprendre l'émergence d'un système de valeurs à partir d'interactions simples[39]. Dans les deux exemples, certes, le sens exact à donner à l'émergence peut grandement varier. Cependant, pour la question qui nous occupe ici, il est clair qu'ils supportent l'idée de la prééminence du particulier vis-à-vis de la règle générale.

En revanche, une démarche descendante se base sur l'idée d'une théorie éthique généraliste universellement applicable (déontologique, utilitariste…). En d'autres termes, elle est caractérisée par l'idée d'une *procédure de décision indépendante* de celui (ou celle) qui l'applique, et qui se prêterait alors bien à la formalisation. Notons qu'indépendant ne veut pas dire insensible au contexte : la procédure peut prendre en compte un nombre important de paramètres contextuels. Il est possible, toutefois, pour deux agents donnés, ayant une connaissance équivalente du monde, de l'appliquer de la même manière. Certaines casuistiques peuvent très bien être dites descendantes d'ailleurs, pour peu qu'elles adoptent une vue large sur leur applicabilité. En effet, une casuistique peut être pensée comme un utilitarisme retourné : plutôt que de s'efforcer de calculer une utilité maximale des conséquences possibles à *l'avenir*, elle regarde derrière elle, le plus souvent au nom du principe de la *tradition* ou de la *coutume*, afin d'y trouver une *similitude* maximale par rapport aux cas *passés*.

C'est dire que la démarcation entre démarches ascendantes et descendantes est parfois délicate, dépendante comme elle l'est de ce que l'on attribue respectivement au contexte et aux agents. Par voie de conséquence, les confusions ne sont pas impossibles, d'autant plus que l'accent mis par certaines formulations particularistes sur le thème de l'apprentissage rappelle une distinction familière à l'ingénieur entre méthodes analytiques et méthodes d'apprentissage. Pour simplifier, disons que l'ingénieur dispose de deux options d'implémentation : soit l'algorithme classique, soit

[37] Ce paragraphe se fonde sur C. MISSELHORN, *Grundfragen der Maschinenethik*, pp. 96, 114-117, ainsi que sur W. WALLACH et C. ALLEN, *Moral Machines*, pp. 79-81. La distinction recoupe *grosso modo* la terminologie, plus courante en philosophie éthique, de généralisme vs particularisme.
[38] W. WALLACH et C. ALLEN, *Moral Machines*, pp. 64-65.
[39] W. WALLACH et C. ALLEN, *Moral Machines*, pp. 101-103.

une méthode d'apprentissage[40]. Dans le premier cas, l'algorithme classique, l'ingénieur spécifie, étape par étape, de façon analytique, comment la machine doit procéder. Dans les méthodes d'apprentissage en revanche, il peut n'avoir aucune idée de la meilleure façon d'obtenir le résultat escompté : dans ce cas, l'ingénieur peut simplement laisser faire la machine par essai et erreur, et la récompenser en cas de réussite. Or force est de constater qu'un raisonnement analogique sur base de cas (nous l'avons vu dans MoralDM) peut très bien s'implémenter de façon analytique. À l'inverse, l'ingénieur peut très bien s'inspirer de Kant lorsqu'il attribue des récompenses à une machine en phase d'apprentissage : la machine, dès lors, aura appris – et mettra en œuvre – une vision descendante. L'ingénieur n'aura ni spécifié ni implémenté cette vision descendante, mais rien n'empêche en principe l'intelligence artificielle de décrire formellement le résultat de son propre apprentissage[41]. Même si les méthodes d'apprentissage ont le vent en poupe, l'ingénieur n'attache en principe aucune importance particulière à l'une ou l'autre démarche, pourvu que « ça marche ». D'ailleurs, dans les systèmes les plus complexes, les deux types de technique se combinent le plus souvent[42].

Une autre source de confusion provient de la volonté de traiter, dans un même élan, prise de décision et jugement éthiques. Or nous venons de voir que la distinction prend sens à partir de la notion d'une procédure de décision, indépendante de celui qui l'applique : le général – la procédure – prime sur le particulier – les paramètres d'entrée. Or comme nous le rappellent Boltanski et Thévenot[43], un jugement va toujours du particulier (les circonstances à juger) au général, en fonction du principe d'ordre que l'émetteur du jugement estime approprié. Le mouvement est donc topique, et par là même, nécessairement ascendant en un certain sens. Cette confusion est par exemple visible lorsqu'un auteur suppose que l'utilisation d'un réseau de neurones formels pour une tâche de jugement éthique pourrait apporter quelque lumière sur le débat généraliste[44]. Or le jugement part nécessairement d'un mouvement ascendant, visant le général au travers du particulier, peu importe le statut de ce général à l'égard de ce particulier. La distinction ascendant-descendant qui nous occupe ici ne peut donc pas s'appliquer au *mouvement* ou à l'acte de juger, mais uniquement au critère de jugement : le regard qui juge est-il universellement applicable, ou dépend-il de celui qui voit ?

---

[40] Assez malencontreusement, les auteurs auxquels nous nous référons ici désignent cette distinction entre méthodes analytiques et d'apprentissage également par celle entre méthodes ascendantes et descendantes (cf. W. WALLACH et C. ALLEN, *op. cit.*, pp. 80-81). Comme – de l'aveu même des auteurs – la distinction dans les moyens ne recoupe nullement la distinction philosophique, nous avons préféré abandonner cette terminologie, source probable de confusions entre les moyens et les fins.

[41] C'est une idée défendue notamment par N. BOSTROM, *Superintelligence* (pp. 258-259), lorsqu'il parle de normativité indirecte : plutôt que de doter une « superintelligence » d'un système de valeurs *a priori*, il pourrait être plus prudent de se contenter d'une contrainte d'apprentissage assez générique et de laisser le soin de trouver comment satisfaire la norme à la machine. Il faut donc, en tout temps, bien distinguer apprentissage et exercice d'une compétence.

[42] Cf. D. J. GUNKEL, qui se fait l'écho de l'idée que la complexité du fait éthique est telle que les simplifications ne sont pas permises, et que tous les moyens à notre disposition sont à hybrider. Il convient également de noter qu'au fur et mesure que la connaissance mathématique des méthodes d'apprentissage s'affermit, les similarités profondes entre les deux types de méthodes sont de plus en plus évidentes (cf. N. BOSTROM, *Superintelligence*, p. 11).

[43] Voir L. BOLTANSKI et L. THÉVENOT, *De la justification*, pp. 16-20.

[44] M. GUARINI, *Computational Neural Modeling and the Philosophy of Ethics. Reflections on the Particularism-Generalism Debate*, dans M. ANDERSON et S. L. ANDERSON, *op. cit.*, pp. 328-332.

Une dernière source de confusion est l'assimilation d'ascendant à descriptif, et de descendant à prescriptif. De fait, si nous dressions un bilan des articles publiés dans ce champ de recherche, un tel rapprochement se verrait peut-être corroboré statistiquement – toujours avec la même réserve que le chemin ascendant n'est pas, à ce jour, fréquemment emprunté. Corrélation n'est pas raison cependant, et le rapprochement n'a rien de nécessaire.

Pour nous en convaincre, reprenons l'exemple du conseiller éthique des époux Anderson, qui a été traité dans le paragraphe relatif aux démarches descriptives et normatives. Dans les prototypes existants, la démarche est essentiellement descendante : les devoirs de base sont connus, les traits éthiques pertinents aussi, le cadre de classification aussi. Or leur dernier projet[45] est de concevoir un conseiller qui pourrait déduire lui-même les traits et devoirs pertinents, rien qu'en se basant sur ses discussions avec des éthiciens. Ainsi, les Anderson restent résolument dans une approche prescriptive ; ils ne changent fondamentalement rien non plus dans leurs moyens technologiques (puisque tous leurs prototypes se basent sur la programmation logique inductive). Cependant, le nouveau système pourrait apprendre ses propres règles, qui ne soient pas déductibles d'un principe généraliste.

Concluons ce paragraphe en rappelant que non seulement les moyens techniques peuvent être combinés, mais les points de vue éthiques aussi. Ainsi, il est tout à fait possible de compléter un raisonnement descendant de type utilitariste ou déontologique par un raisonnement analogico-inductif[46] – c'est d'ailleurs la voie suivie par MoralDM. Une telle extension permettrait d'amortir le problème du contexte – même si ce n'est pas suffisant pour le résoudre – tout en élargissant significativement le nombre de cas traitables.

| Questions |
| --- |
| **Les SMA peuvent-ils modéliser tant des approches ascendantes que descendantes ?** |
| **Présentent-ils une préférence ou un biais pour l'une ou l'autre approche ?** |
| **Qu'en est-il de l'idée d'une procédure de décision : comment est-elle modélisée dans les SMA ?** |
| **La procédure de décision est-elle propre à l'agent, ou s'agit-il d'une procédure universellement applicable ?** |
| **Plusieurs procédures (qu'elles soient ascendantes ou descendantes) peuvent-elles entrer en concurrence ?** |
| **Les agents apprennent-ils les procédures, ou celles-ci sont-elles données ?** |
| **La SBA produit-elle des résultats cohérents, répétables ?** |

[45] Pour tous les détails, voir S. L. ANDERSON et M. ANDERSON, *A Prima Facie Duty Approach to Machine Ethics*, dans IID., *op. cit.*, pp.483-490. Voir aussi, pour l'interprétation que nous en donnons ici, W. WALLACH et C. ALLEN, *Moral Machines*, pp. 128-129.

[46] M. KLINCEWICZ, *Challenges to Engineering Moral Reasoners*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, pp. 252-254.

**De quels principes méta-éthiques se réclame-t-elle ?**

## 1.6. La justification

Une dernière mise au point s'impose avant d'attaquer le vif du sujet : c'est le problème auquel renvoie le paradoxe de Polanyis inversé[47]. Le paradoxe de Polanyis désigne cette difficulté qu'a l'être humain de dire à la machine *comment* les choses doivent être faites : l'homme ne peut donner une spécification détaillée de la *fonction*, il est seulement en état de juger de la réussite du résultat final. C'est le point de départ non seulement des métaheuristiques, mais de l'intelligence artificielle en général, qui privilégie les buts à atteindre par rapport aux règles pour y parvenir.

Le paradoxe s'inverse lorsque l'ordinateur, à son tour, n'est plus capable d'expliquer comment il a obtenu les résultats qu'il fournit. Ainsi, l'apprentissage induit sur les mégadonnées (*Big Data*) ne laisse de trace que dans le calibrage des réseaux de neurones formels, sans justification intrinsèque, ou, plus précisément, sans justification exprimable dans les termes du domaine où la question posée prend son sens. C'est une situation éthiquement plus que douteuse, appelée à devenir une problématique majeure de l'éthique de l'informatique au sens large donné plus haut : au fur et à mesure que le champ d'application de ce type d'applications auto-apprenants s'étend, nous risquons de nous retrouver dans un régime technocratique où ne règne plus la loi éclairée de la raison, mais de l'obéissance aveugle à des algorithmes quasi magiques, car opaques.

L'éthique des machines se doit d'être attentive à ce problème. C'est pourquoi Susan Anderson, une des initiatrices de la discipline, insiste sur ce point : la justification doit être exprimable en langage compréhensible pour un être humain ; elle doit être évaluable en termes éthiques[48]. Une telle exigence pose la question de la théorie éthique à choisir, déontologique ou utilitariste. Si nous penchons pour un cadre déontologique, encore faut-il être clair sur le statut du devoir que nous invoquons : le devoir peut être absolu à la Kant, conditionnel à la Ross, ou empirique, auquel cas il peut encore résulter d'une généralisation par induction ou par analogie.

Liée à cette problématique de la justification, nous en trouvons deux autres. Premièrement, il y a la question de l'apprentissage des principes et des règles, c'est-à-dire des *termes* dans lesquels la justification va être formulée. À noter qu'en cette matière, selon le mécanisme retenu (déontologie, utilités, généralisation), le contenu de l'apprentissage des principes ne sera pas le même. Deuxièmement, ces règles doivent être appliquées à des cas concrets afin de servir de base à une justification probante. Une telle démarche nécessite toutefois de la machine des connaissances du monde et une prise en compte de son environnement qui sont loin d'être évidentes et posent des défis techniques considérables.

---

[47] Cf. Th. RAMGE, *Mensch und Maschine*, pp. 16-28.

[48] S. L. ANDERSON, *Machine Metaethics*, dans M. ANDERSON et EAD., *Machine Ethics*, pp. 22 et suivantes.

Notons que devant ces problèmes, les différents courants peuvent prendre des positions singulièrement divergentes. Ainsi une démarche prescriptive, surtout si elle est descendante, peut simplement réutiliser le cadre théorique de la prise de décision pour exprimer verbalement la justification. Une démarche descriptive, au contraire, se voit placée devant un problème inédit, nouveau : comment, en effet, décrire ce besoin qu'a l'être humain de se justifier ? Quel statut donner, aussi, aux productions de ce besoin de se justifier : rationalisation *post hoc*, ou description verbale d'un mobile lucide[49] ? À l'heure actuelle, le thème de la justification semble monopolisé par les études prescriptives. Nous voudrions cependant soulever la problématique qui est celle de la description fonctionnelle de la justification : que se passe-t-il lorsque l'homme se justifie ?

Un point de vue éclairant sur cette question – et sur lequel l'éthique des machines sera peut-être amenée un jour à se prononcer – est le travail des sociologues Boltanski et Thévenot[50]. Ces auteurs partent d'une interrogation qui consiste à se demander ce qui, lorsque deux êtres humains ont des intérêts divergents, rend l'accord possible en société. Comment établir un accord, surtout quand on écarte tant le recours à la force que le refus de toute confrontation qu'est le relativisme ? Les auteurs commencent par reconnaître qu'il existe une *pluralité* de manières de construire l'accord. Chacune de ces manières sont des représentations idéales de voir le vivre ensemble ou, dans la terminologie de l'ouvrage, des *mondes* ou *cités* possibles. Chacune de ces cités renvoie à un principe supérieur d'accord : une circonstance particulière doit pouvoir être rapprochée de la généralité exprimée par le principe. Afin d'intervenir dans une justification, ce rapprochement doit être communicable, c'est-à-dire commun.

Un tel principe supérieur a pour effet d'instaurer un ordre, un ordre qualifié, permettant d'ordonner les objets et les êtres du monde du plus petit au plus grand, du plus contingent au plus nécessaire, du plus particulier au plus général. Cet ordre qualifié exprime également le degré de participation au bien commun représenté par le principe. Liée à l'ordre (dit *de grandeur*) est la notion d'épreuve : une épreuve est la confrontation exemplaire prévue par la cité pour gérer le désaccord : une épreuve établit, confirme ou conteste, la place respective des êtres dans la hiérarchie de la cité.

Dans notre monde moderne, les auteurs dénombrent six de ces cités, soit six manières d'établir l'accord : la cité *domestique* d'abord, fortement structurée dans le temps et l'espace autour du foyer. On y privilégie les relations interpersonnelles et les traditions. La cité *civique* est celle des collectivités, qui exprime la volonté générale de ses adhérents. La cité *industrielle* est celle de l'efficacité technoscientifique, où sont privilégiées les mesures et les méthodes. Il y en a d'autres (citons-les pour mémoire : la cité *marchande*, organisée autour de la richesse, la cité de *l'opinion*, bâtie sur la célébrité, et la cité *inspirée* enfin, érigée autour des notions d'originalité, créativité et d'authenticité personnelle), mais ces dernières sont sans doute moins pertinentes pour notre recherche.

[49] Formulé encore autrement, la démarche descriptive se voit confrontée à la tâche de caractériser l'exigence éthique spontanée : elle cherche à élucider le processus mis en œuvre par tout homme qui, dans sa vie de tous les jours, se pose la question « *que dois-je faire ?* ». Elle doit donc décrire un phénomène dont elle ne peut remettre en cause le statut normatif. La démarche descriptive prend acte d'une normativité *déjà là*, alors que la démarche prescriptive impose la sienne propre.

[50] Ils développent ce point de vue dans leur ouvrage *De la justification*.

Cette complexité inhérente fait échec à certaines démarches simplistes ; ainsi la rigidité kantienne face au mensonge – ne pas dire toute la vérité et rien que la vérité est un affront à l'idée même de vérité – devient un choix de cité et, par là, d'épreuve qu'on fait importer. Prenons l'exemple rodé du membre de la Résistance qui cache des Juifs dans sa cave et qui reçoit la visite de la Gestapo : si nous suivons une approche descendante, nous devrions expliquer, d'une façon ou d'une autre, que le devoir de protéger les malheureux à la cave prime sur notre devoir de dire la vérité. Dans la théorie de Boltanski et Thévenot, il suffit de se rappeler que dire la vérité, créer la confiance, est une épreuve qui se joue dans le monde domestique : nous dirons la vérité à celui avec qui nous voulons cultiver des relations interpersonnelles dans la durée. Or le résistant peut, sans le dire, refuser cette épreuve, car souhaitant le retour à la République, il perçoit dans les représentants de la Gestapo les ennemis de la volonté générale du peuple français. Dans cette cité civique, les relations interpersonnelles ne sont pas valorisées, mais constituent du bruit, du contingent qui vient troubler la pure expression de la volonté du peuple[51]. L'opposition entre grandeur domestique et civique est encore celle qui anime l'exemple littéraire que nous avons donné en introduction : là où Antigone valorise les relations filiales, Créon se voit davantage obligé par le respect des lois qui expriment la volonté du peuple. Une même observation de conflits entre cités peut être faite à propos de l'utilitarisme : la fonction d'utilité maximale est souvent tirée d'un monde particulier. Or quelle que soit la façon de calculer l'utilité, l'impartialité requise par l'utilitarisme contrevient à notre attachement naturel, que nous estimons éthique, vis-à-vis de nos proches. Utiliser une justification autre que domestique – seule cité dans laquelle nos proches importent plus qu'autrui du fait même de leur proximité – serait considéré comme un calcul aliénant, pouvant mener jusqu'à la schizophrénie morale[52].

De tout ceci, il faut retenir qu'en fonction de la situation, l'être humain va mobiliser des objets, des valeurs, propres à une idéalité variable. Toute explication de la justification devra s'intéresser à ce choix possible entre typologies qui peuvent se faire concurrence, se renforcer ou même s'hybrider au gré des circonstances. L'éthique des machines – une fois qu'elle se sera dotée d'une perspective explicite sur la question de la justification – nous inviterait alors à reprendre avec un regard neuf la question du statut de la règle : le comportement éthique se laisse-t-il décrire (ou prescrire) par des règles ? Si oui, dans quelle mesure celles-ci sont formalisables ?

| Questions |
| --- |
| **Les SMA produisent-ils des justifications ? De quel type ?** |
| **La justification produite par un SMA est-elle évaluable en termes éthiques ? Peut-elle être exprimée en langage humain ?** |
| **Dans quelle mesure cette justification se base-t-elle sur des règles ?** |

[51] La théorie permet, par ailleurs, aussi d'accommoder la thèse d'Isabelle Stengers selon laquelle le discours scientifique s'arroge indûment un droit de parole exclusif sur ce qui compte : retraduite dans la terminologie présentée ici, il s'agit d'une rigidité qui ne veut voir que le monde industriel, ses épreuves, ses valeurs, ses objets et ses êtres. Nous aurons l'occasion de revenir amplement sur le dilemme de la Gestapo au troisième chapitre, dans notre section 3.3.2, consacrée à l'épreuve de la norme.

[52] Reformulé d'après Chr. GRAU, *There Is No « I » in « Robot ». Robots and Utilitarianism*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 454-461.

| |
|---|
| **Dans le choix des termes, un SMA permet-il d'exprimer tous les cadres éthiques ou est-il biaisé en faveur de certaines d'entre elles ?** |
| **Quels types de valeurs sont propices aux justifications des agents en SMA ?** |
| **Dans quelle mesure un SMA peut-il rendre compte de la complexité d'un monde topique ?** |

## 1.7. L'agentivité fonctionnelle

La notion d'agent est importante, car utilisée aussi en SMA. Or, alors que cette notion est à première vue relativement bien définie en SMA, elle est très élastique en EM. Différents auteurs énoncent différents critères, ou pire, utilisent les mêmes critères avec des significations parfois fort divergentes. En outre, selon qu'on aborde la question d'un point de vue descriptif ou normatif, les réponses ne sont pas les mêmes. C'est de cette foisonnante discussion que les paragraphes qui suivent vont tenter de rendre compte.

### 1.7.1. Approches classiques

Afin de comprendre les approches classiques de l'agentivité morale, partons du schéma conceptuel qui oppose l'homme à *l'outil*. Cette opposition se lit comme un rapport d'altérité absolue : l'homme *crée* l'outil dans un geste souverain en vue d'une finalité précise, le *contrôle* selon sa volonté, et en *connaît* le fonctionnement dans ses moindres détails : l'outil est sans surprise ni véritable consistance. En face de l'outil, « objet » de prise et de saisie, se trouve l'agent qui saisit, qui comprend et qui juge. Il se comprend comme initiateur d'une action intentionnelle, choisie librement au travers d'un processus délibératif en suivant des règles codifiées[53]. Ces critères peuvent être augmentés par une exigence de conscience de soi et de rationalité, d'avoir des projets de vie, d'avoir un système de valeurs qui ne se confond pas avec ces projets, en plus d'avoir la capacité de choisir librement entre ces différents projets[54].

Tel est, posé grossièrement, un certain schéma qui structure notre vision du monde. Or ce schéma va au-devant de difficultés de taille, tant sur le fond qu'en tant que programme de recherche. En effet, il donne lieu tout d'abord à un problème épistémologique : comment connaître les états, ou dispositions internes, d'autrui ? Même si nous nous limitons aux propriétés individuelles, les discerner chez autrui est très problématique. Ce problème, dit « des autres esprits », remonte loin ; D. J. Gunkel rappelle à cet endroit la critique kantienne de Platon[55] : nous n'avons pas un accès privilégié aux Idées, le *Ding an sich* est radicalement inconnaissable, même s'il faut « croire » à son

[53] D. J. GUNKEL, *The Machine Question*, pp. 22-23.
[54] *Ibid.*, pp. 46-47.
[55] *Ibid.*, pp. 139-141.

existence, pour éviter l'absurdité d'une apparence sans apparaissant. L'homme n'a accès qu'à son apparence, telle que nous la percevons[56].

L'épistémologie n'est cependant pas le seul point faible de la vision traditionnelle. Sur le plan des contenus, même si à première vue elle paraît assez intuitive, force est de constater qu'elle ne correspond pas tout à fait à la réalité, même en limitant l'analyse aux outils au sens le plus strict du terme. Car, l'outil a beau être créé par l'homme, il finit toujours par dépasser l'intention de son créateur. C'est ce que Gilbert Simondon a appelé les « fonctions surabondantes » de l'outil[57]. Les fonctions surabondantes sont celles qui se rajoutent à l'objet technique quand il se concrétise, se détache de la pensée, bref il s'agit de fonctions qui dépassent l'idée de leur conception. Selon Simondon, la réalisation matérielle dépasse toujours le but de l'innovation.

Si l'outil acquiert de la manière que nous avons dite une certaine épaisseur, par ses potentialités que l'homme *découvre* plus qu'il ne les crée, il en vient aussi à déborder sur l'autre versant de la dichotomie, celui du sujet. Simondon[58] analyse ainsi l'évolution de la société artisanale, où l'homme est seul à manier des outils, à la société industrielle, où la machine devient elle aussi porteuse d'outils. La machine échappe encore au contrôle de l'homme par une troisième voie, qui est celle de l'autorégulation, de sorte qu'elle s'émancipe de sa condition d'outil passif pour devenir un *individu technique* à part entière.

La fascination de l'outil finit parfois par nous contrôler : cela arrive entre autres lorsque nous sentons notre puissance décuplée par la possession d'une arme, et que notre comportement s'en trouve altéré. L'agressivité qui en résulte, est-ce vraiment nous ? Ou est-ce l'arme qui s'est imposée à nous ? Un autre exemple, tout aussi classique, concerne le sentiment de liberté procuré par la voiture. Ceci ne s'explique que si nous comprenons l'outil dans une sorte de fusion avec l'homme, l'outil comme *prothèse*[59].

À mesure que ce schéma binaire est appliqué à des artéfacts de plus en plus complexes, il se fragilise davantage : ainsi un ordinateur personnel n'a plus grand-chose de « connaissable » pour son utilisateur moyen. Ceci explique par ailleurs que dans certains cas, le PC est le « coupable » tout

---

[56] Le problème des autres esprits avancé par D. J. GUNKEL (*op. cit.*, *passim* et notamment p. 55) est lui-même non exempt de critiques. Si nous exagérons l'importance du problème, en concluant à l'impossibilité radicale de connaître l'esprit d'autrui, nous faisons par là même un sort à toute la psychologie cognitive contemporaine. Yuk HUI (*On the Existence of Digital Objects*, pp. 111-114) rappelle toute une tradition allant d'Aristote en passant par les idéalistes allemands, puis de Husserl jusqu'à Heidegger, qui dénonce cette idée en insistant sur les contraintes qu'oppose le connaissable à notre perception. Ainsi Heidegger puise la source principale de la connaissance de l'objet (ou de la chose) sur le mode *zuhanden* dans la totalité des références ou renvois que l'objet entretient avec son *Umwelt*. Par exemple, le marteau sert à enfoncer un *clou* (dans une *table*, un *mur*). Par *zuhanden*, il faut comprendre un type de rapport aux objets qui soit affaire d'expérience dans le monde : c'est le domaine de la fréquentation « quotidienne » de l'objet. Ce mode de saisie s'oppose à la saisie sur le mode *vorhanden*, qui est la saisie thématique (ou scientifique) d'un objet, la saisie qui voit l'objet dans son idéalité (ou, si l'on préfère, sa généralité).

[57] Y. HUI, *On the Existence of Digital Objects*, pp. 102-105.

[58] Cf. G. SIMONDON, *Sur le mode d'existence des objets techniques*, pp. 97-99.

[59] D. J. GUNKEL, *op. cit.*, p. 26. À noter que D. PAROCCHIA signale une fusion semblable entre l'homme et la machine, entre le pilote et son avion, aux débuts de l'aviation, l'époque dite « héroïque » dont Antoine de Saint-Exupéry s'est fait la voix littéraire (*L'homme volant*, p. 84).

trouvé[60] : quand on se trouve dans l'impossibilité de réserver ses tickets, lorsque l'imprimante « déconne », etc. Pouvons-nous nous tourner vers les spécialistes pour mieux comprendre : les programmeurs au moins connaissent les applications dont ils ont la charge ? Or là aussi, nous devons nous rendre à l'évidence : même un logiciel de taille moyenne nécessite tellement de lignes de code et de bibliothèques tierces qu'il échappe même aux membres individuels de l'équipe en charge de son développement. N'importe quelle application contemporaine, même de taille et de complexité pourtant modestes, fait appel à des dizaines, voire des centaines, de composantes logicielles préfabriquées et qui échappent largement au contrôle et même à la compréhension de l'équipe de développeurs en tant que collectif. C'est dire qu'une application donnée peut avoir une génèse, une histoire et une évolution de type complexe[61].

Pourtant, l'autonomie croissante des machines n'a pas, dans un premier temps, changé radicalement la donne : la machine va continuer à être vue comme un outil autonome par Hegel, repris par Marx, qui note cependant qu'à l'instar de l'homme, la machine va manier ses propres outils[62]. Encore dans la définition instrumentale de Martin Heidegger, l'assimilation de toute machine à un outil, quelque complexité qu'elle ait, est un fait[63]. Or la définition instrumentale nous laisse démuni pour comprendre l'autonomie d'une machine. Si la machine est un outil, son autonomie ne peut être comprise que négativement, sous l'angle d'une « perte » de contrôle[64]. Or si nous présupposons que la machine doit être sous contrôle, la question de savoir si nous exerçons assez de contrôle sur la machine pour en assumer la responsabilité éthique devient obsédante.

Telle se présente la situation du premier terme de l'opposition, la machine comme outil. Or, l'autre terme de l'opposition n'est pas moins problématique : l'agent, défini par ses propriétés individuelles, prête le flanc à toutes critiques faites à la notion de sujet de Leibniz à Ricœur. Cette unité de la conscience et de la pensée, transparente à sa propre raison et connaissable parfaitement par sa propre intuition, se comprend par altérité radicale par rapport à son autre, la machine. Cette conception est problématique à plus d'un titre : d'abord, H. Putnam a relevé que le fait d'être un artéfact, ou un système physique déterministe, ne préjuge en rien le fait d'être conscient, il n'y a pas de lien entre ces arguments, de même que l'homme en tant que création divine n'a jamais empêché de penser l'homme comme être conscient[65]. Ensuite, l'homme peut aussi être considéré comme une machine (biologique)[66]. Enfin, il est même possible d'argumenter que la différence fondamentale entre un outil et un agent (social) est avant tout une affaire de métaphore, qui crée la réalité

[60] D. J. GUNKEL, *The Machine Question*, p. 28.
[61] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 205-206.
[62] D. J. GUNKEL, *op. cit.*, p. 31.
[63] Y. HUI, *On the Existence of Digital Objects*, p. 191.
[64] D. J. GUNKEL, *op. cit.*, pp. 35-38.
[65] Rapporté dans *ibid.*, p. 53.
[66] Sous ce rapport, la thèse de Parocchia est particulièrement intéressante : au fur et à mesure que l'homme développe des instruments de perception empruntés aux autres animaux – le radar, la caméra infrarouge, etc. – il est mieux à même de se faire une idée du vécu de ces derniers. Le progrès technologique peut ainsi se lire comme un « devenir-animal » sublimé, au cours duquel l'homme conquiert des dimensions de plus en plus riches de l'expérience animale et ce, au point de faire dire à l'auteur que « le monde se récapitule de plus en plus exhaustivement dans l'homme » (D. PAROCCHIA, *L'homme volant*, p. 179).

correspondante[67]. Il vaut donc mieux recentrer le débat sur ce que « l'agent » donne à observer en situation, son comportement, en suspendant nos présupposés et nos intuitions sur nos propres motivations et sans reléguer *a priori* la machine au rang d'outil inerte, donné une fois pour toutes. Bref, il faut recourir, en la matière, à un regard fonctionnaliste.

## 1.7.2. Approche fonctionnelle

Il faut donc problématiser cette opposition binaire sujet-outil, afin d'en déduire une caractérisation probante de ce que cela implique d'être un agent : si agent il y a, quels comportements sommes-nous en droit de voir apparaître ? Dans les différents critères d'agentivité maniés par les auteurs que nous avons consultés, il serait cependant difficile de relever un consensus. Notre interrogation sera donc essentiellement ouverte, multipliant les questions que nous pourrons ensuite adresser aux SMA.

Dans la mesure du possible, nous voudrions que ces questions respectent l'engagement fonctionnaliste de l'éthique des machines, tout en restant critique sur d'éventuelles confusions ou zones d'ombre sur lesquelles nous pourrions tomber. Nous aurons également à cœur de rester le plus minimaliste possible dans nos exigences vis-à-vis d'un agent technologique. La raison en est simple : en partant de l'expérience de pensée d'un agent technologique qui serait en tous points indiscernable d'un être humain, nous pourrions soutenir qu'il serait « évidemment » éthique, mais il s'agirait d'une vérité triviale, qui n'engagerait à rien et ne nous apprendrait rien non plus.

### *1.7.2.1. Unité d'agir*

Un premier critère dont il convient de parler n'est souvent pas énoncé tel quel, il est pourtant essentiel : pour qu'il y ait une action qu'on puisse qualifier d'éthique, il faut d'abord s'arrêter sur cette notion même d'action. Posons-nous la question suivante : qu'est-ce qui distingue une action d'un évènement ? Qu'est-ce qui distingue *l'action* de férir un coup de marteau de *l'évènement* qu'est l'impact de la foudre ? Le but de la question n'est pas d'entrer dans le détail d'une théorie de l'action, nous nous contenterons ici de faire observer que la foudre *est* l'évènement, tandis que le coup de marteau trouve sa source en dehors de lui-même[68].

En d'autres termes, là où l'évènement se suffit à lui-même, parler d'action revient toujours à poser la question de la *source* de cette action. Cette source est le plus souvent tenue pour évidente, donnée dans le concept d'agent. Or ce concept ne va pas de soi : la source de l'action doit être perçue comme

[67] K. DARLING, *« Who's Johnny ? » Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, pp. 147-176. Nous aurons l'occasion de revenir à cette problématique en abordant la valeur fonctionnelle.

[68] Cette distinction entre évènement et action se trouve au cœur de la praxéologie (voir le chapitre que D. VERNANT y consacre dans son ouvrage *Introduction à la philosophie contemporaine du langage*, pp. 141 et suivantes).

étant unie[69]. Nous retrouvons le même problème en logique. Lorsque Bertrand Russell énonce la proposition « Pour tout x, si x est Roi de France, alors x est chauve », nous pouvons très légitimement nous interroger sur le statut de ce « x » : quel critère, quelle propriété, nous permet de discerner un individu, avant même de lui avoir attribué quelque propriété que ce soit[70] ? Dans quelle population est-il légitime de faire varier x ? Pour garder un sens à la proposition ci-dessus, la réponse semble évidente : « x » ne peut être interprété que comme un être humain de sexe masculin, d'extraction noble et de nationalité française. Or, contrairement à la logique formelle, l'éthique des machines ne peut espérer combler le vide par un renvoi implicite à notre connaissance du monde.

La question n'est évidemment pas fonctionnelle : plutôt, elle vient nous rappeler que la démarche fonctionnelle a beau mettre entre parenthèses les aspects ontologiques et épistémologiques des facultés éthiques[71], celles-ci ne disparaissent pas pour autant. Le critère de l'unité fait en quelque sorte office de résumé du prérequis ontologique et épistémologique. Puisqu'elles sont présupposées par la démarche, l'enquête philosophique aura en dernière analyse à y revenir. En attendant ce retour, cependant, qui ne sera pas entrepris dans le cadre de ce mémoire, il vaut mieux éviter les engagements ontologiques. Idéalement, nous voudrions que notre notion d'agent soit neutre du point de vue ontologique, eu égard à la grande diversité de sources que la pensée contemporaine semble investir de pertinence éthique : non seulement des individus humains ou animaux, mais également des collectivités comme des sociétés ou des organisations prétendent à ce titre[72].

La question de savoir quel type d'entité peut prétendre au titre d'agent a été soulevée avec acuité – et presque à son corps défendant – dans l'expérience de pensée du philosophe étatsunien John Searle, connue sous le nom de l'argument de la chambre chinoise[73] : un opérateur humain (anglophone) est enfermé dans une chambre, ayant à sa disposition une table de conversion de caractères chinois, un dictionnaire ainsi qu'un ensemble d'instructions syntaxiques. Si, de l'extérieur, on pose à la chambre des questions sur un texte chinois, les réponses données par un locuteur natif de la langue chinoise et celles de notre opérateur anglophone peuvent très bien être indiscernables. Et Searle de conclure que l'opérateur ne comprend pas le chinois, car il n'a fait que manipuler des symboles, de la même manière que ne l'aurait fait un ordinateur.

En d'autres termes, Searle affirme qu'une approche fonctionnelle de la faculté du langage n'est pas opérante, étant donné qu'en dernière analyse, l'opérateur humain ne comprend pas le chinois. L'argument, certes, présente plusieurs faiblesses, dont celle d'être impossible : sans connaissance du monde chinois et de sa culture, la chambre serait bien en peine de fournir des réponses adéquates. La syntaxe et le calcul, en gros, n'y suffisent pas. Il n'en reste pas moins que l'argument a eu un certain retentissement dans le milieu de l'intelligence artificielle et au-delà. Dans les préoccupations qui sont

---

[69] Nous retrouvons là un thème phénoménologique bien connu. En effet, selon Edmund Husserl, la conscience est un flux. « Distinguer » des unités dans ce flux, voilà une tâche qui revient à l'esprit (cf. Y. HUI, *On the Existence of Digital Objects*, p. 94).

[70] Voir, pour cette problématique, G. CHAZAL, *Les réseaux du sens*, pp. 155-159 et J.-Y. BÉZIAU, *Le Château de la Quantification et ses Fantômes Démasqués*, dans P. JORAY, *La quantification dans la logique moderne*, pp. 213-216.

[71] W. WALLACH et C. ALLEN, *Moral Machines*, p. 55.

[72] Le problème de l'engagement ontologique qu'entraîne le recours à la notion d'agent a été souligné par M. TURILLI, *Ethical Protocols Design*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 376-378.

[73] Notre discussion sur la chambre chinoise reprend celle que lui consacre G. CHAZAL, *Le miroir automate*, pp. 66-69.

les nôtres, l'objection principale qui a été faite est celle dite « du système » : certes, l'opérateur humain pris en tant que tel ne comprend pas le chinois, mais le système constitué par lui, le livre d'instructions et le dictionnaire, comprend bel et bien le chinois.

Searle, en quelque sorte malgré lui, fournit donc un bel exemple de ce que la notion de conscience individuelle n'est pas suffisante pour comprendre l'interaction qui est ici en jeu. Sur le plan de l'interaction en chinois, il faut admettre que la chambre est source d'action. C'est elle l'unité, l'opérateur humain n'en est qu'une *partie*. Le même argument vaut d'ailleurs pour la situation par trop connue de la panne informatique : le pauvre employé de la poste, de la banque…, qui se voit obligé de rejeter la faute sur l'ordinateur qui refuse d'imprimer les billets, ou d'enregistrer la transaction bancaire, se trouve dans une situation où il n'est pas maître de l'action à accomplir, il n'est que l'interface entre le client et la machine.

Mais qu'est-ce donc qu'un système ? Selon Yuk Hui[74], il s'agit d'un ensemble de relations interobjectives dont la configuration et la densité ont atteint une maturité telle qu'elles surmontent tous les obstacles dans le temps et dans l'espace. Ces relations peuvent prendre la forme de relations physiques (comme une connexion par câble) mais tous les schémas de données ou de définitions de protocoles servent le même rôle. Dans un système où les données constituent la forme matérielle la plus importante des relations interobjectives qui permettent de connecter les différentes parties, on peut parler d'un système d'information. La notion s'oppose conceptuellement à une relation intersubjective, c'est-à-dire une relation entre un sujet et son contexte, où ce dernier terme n'est rien d'autre qu'un ensemble de significations que le sujet s'est donné lui-même.

Cette notion d'interobjectivité nous pousse à réévaluer « l'argument du système » qui a été opposé à Searle. Certes, le « système » comprend bel et bien le chinois. Par extension, nous pourrions dire qu'un « système » peut être source d'unité pour un comportement qualifiable d'éthique. L'employé de banque va, en d'autres termes, établir une *relation* entre le client, d'une part, et la machine, d'autre part. Il s'agit d'une relation *interobjective*, c'est-à-dire une corrélation entre un objet et son milieu. Cette analyse n'est pas sans rappeler la définition de Floridi lorsqu'il écrit[75], en parlant d'un être humain, que celui-ci est un agent à un niveau d'abstraction donné si, et seulement si,

1. il peut compter comme un système dans un environnement ;
2. il y initie une transformation ;
3. il y produit des effets perceptibles.

Il y a cependant une nuance : Floridi cherche en effet le caractère moral d'un système d'information comme étant inhérent à sa qualité de véhicule d'information[76]. C'est une thèse beaucoup plus forte que celle présentée ici : en évoquant le travail d'Hui, nous souhaitons – beaucoup plus modestement – donner une assise raisonnable à la neutralité ontologique dont l'approche fonctionnelle se réclame. Toujours est-il que dans les deux cas, les auteurs se rejoignent dans la

[74] Y. HUI, *On the Existence of Digital Objects*, pp. 154-164.
[75] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 192-193.
[76] Cf. L. FLORIDI, *Information ethics. On the philosophical foundation of computer ethics*.

récusation d'un individualisme anthropocentrique pour épouser une idée « d'agentivité étendue »[77]. Celle-ci conçoit le sujet agissant comme un verbe plutôt qu'un nom : le fait de combiner différentes entités de diverses manières en vue d'activités différentes. Une telle vue peut même être appliquée au sujet humain. En effet, notre cerveau n'est-il pas un système modulaire dont les différentes composantes construisent ensemble une illusion d'agent unique[78] ?

De là à affirmer que l'unité, en d'autres termes, n'est pas donnée mais doit être *construite* et *perçue*, saisie par un esprit, il n'y a qu'un pas, qu'il faut se garder de franchir trop prestement, sous peine de présupposer un sujet conscient à chaque source potentielle d'action. L'exemple de la chambre chinoise nous a bien montré que l'unité n'est pas à chercher dans une propriété que l'agent posséderait intrinsèquement, mais lui est plutôt conférée dans un contexte précis. Dans d'autres contextes, par exemple, lorsque je croise John Searle dans la rue, la chambre chinoise n'aurait pas d'unité pertinente (à moins bien sûr que notre conversation ne se fasse en chinois). L'unité n'est ainsi pas intrinsèque, il s'agit d'une caractéristique qui lui est conférée par un contexte ou une structure, ou encore un faisceau de relations qui existent ou qui sont exercées en dehors de l'unité elle-même.

La notion d'interobjectivité permet par ailleurs de remplacer cette notion de « contexte » avantageusement par une autre, celle de « milieu », tissé de relations indifféremment informationnelles ou physiques, sans subordonner le lien au sujet et donc sans présupposer celui-ci à la constitution de l'unité d'agir. Le grand intérêt de la notion de milieu dans ce contexte est sa neutralité à l'égard des grandes questions liées à la conscience, au statut de la personne, etc. Il nous est ainsi possible de problématiser la notion d'agent, sans pour autant devoir abandonner notre position fonctionnelle, prudente, de type ingénieur.

| Questions |
| --- |
| **Un SMA est-il neutre d'un point de vue ontologique, c'est-à-dire, permet-il de suivre rigoureusement l'approche fonctionnelle ?** |
| **Les SMA présentent-ils des contraintes particulières sur le plan épistémologique ? L'approche fonctionnelle interdit-elle de connaître certains aspects (essentiels) de la dimension éthique ?** |
| **Qu'est-il possible de connaître à propos des entités que les SMA mettent en œuvre, que ce soit sur le plan de la connaissance commune ou scientifique ?** |

[77] Le concept provient d'un article de 2009 de la main d'Allan Hanson (cité dans D. J. GUNKEL, *The Machine Question*, p. 165).

[78] Dr. MCDERMOTT, *What Matters to a Machine?*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 108. Ce même auteur rapproche la modularité cérébrale d'un système d'exploitation, qui se compose de différents *processus* (*ibid.*, p. 99).

### *1.7.2.2. Identité*

Quoi qu'il en soit des prérequis ontologiques et épistémologiques d'une source d'agir, celle-ci doit se préserver dans le *temps*. La continuité dans le temps des agents, voilà la définition minimale de leur identité, pour autant – bien sûr – de ne pas conférer au concept « temps » un contenu trop anthropocentrique[79]. Sans identité des agents, la dimension éthique serait privée de la plupart de ses ressorts : en effet, afin d'anticiper le comportement d'un interlocuteur, de lui rappeler les promesses qu'il nous a faites, bref, afin de gérer l'interaction avec autrui, nous avons besoin de lui reconnaître une permanence, une existence continuée, bref, une identité.

Conformément à notre position fonctionnaliste, l'identité n'est donc pas quelque chose d'intrinsèque à un agent, mais une condition d'interaction : nous devons pouvoir reconnaître l'autre comme étant le même, et réciproquement, car la contrainte semble avoir un double aspect. D'une part, la source d'agir doit se reconnaître identique à elle-même, être dans la continuité de ce qu'elle fut. Le deuxième aspect est la perception de l'identité aux yeux de l'observateur : nulle promesse ne saurait être tenue, par exemple, si mon interlocuteur ne me reconnaît pas comme l'agent à qui une promesse a été faite[80].

L'identité est donc ce qui, dans un jeu d'interactions donné, permet aux interactants de se reconnaître mutuellement. Si le système requiert de ses agents d'avoir conscience d'eux-mêmes (ce qui est le cas dans une interaction éthique entre deux êtres humains), alors l'agent doit également pouvoir se reconnaître lui-même au travers des mêmes critères[81]. La difficulté, bien sûr, consiste à évaluer quels traits doivent être continus dans le temps. Les réponses traditionnelles posent en général deux contraintes majeures : continuité corporelle, d'une part, continuité mentale, d'autre part[82]. Pour ce qui est du corps, l'observateur avec qui nous interagissons doit percevoir à travers les transformations temporelles que sont le vieillissement, la régénération cellulaire constante des tissus

---

[79] P. Ricœur dégage deux conceptions majeures du temps : la première, qu'il fait remonter à Aristote, est le temps cosmologique, le temps qui ne connaît qu'un *avant* et un *après*, qui s'écoule à la manière d'un flux ou, dans le vocabulaire d'Aristote, comme un *mouvement*. La deuxième conception, qu'il attribue à Saint-Augustin et qui sera reprise et élaborée par Husserl, puis Heidegger, est le temps humain, celui qui s'articule autour de la notion de *présent*. Saint-Augustin voit en effet le temps comme un triple présent : passé, présent, futur. Si donc nous faisons référence au passé, ce que nous voulons dire, c'est que nous nous rendons « présent » le passé au moyen d'une *distension de l'esprit*. À ces deux conceptions fondamentales, le temps comme flux, le temps comme triple présent, Ricœur rajoute une sorte de passerelle, le temps calendaire (P. RICŒUR, *Temps et récit*). Ceci dit, si nous voulons concevoir le temps comme agissant dans le monde sans référence à l'être humain, Y. Hui nous fournit une troisième conception du temps : le temps comme relation interobjective, celle qui mène du terme amorphe (porteur d'énergie potentielle) vers le terme structuré (porteur de structure asymétrique). Plutôt que flux ou expérience subjective, le temps ainsi compris est une transformation du monde ; le monde est ainsi une *fonction* du temps (Y. HUI, *On the Existence of Digital Objects*, pp. 173-186).

[80] La section sur l'identité aurait fort gagné en clarté si nous avions eu connaissance plus tôt des travaux de P. RICŒUR à ce sujet. Dans *Soi-comme un autre* (pp. 39-54), il distingue – bien plus clairement que nous l'avons fait ici – entre individualisation, identité comme mêmeté et identité comme ipséité. L'individualisation renvoie à notre discussion sur l'unité d'agir : il s'agit d'un échantillon indivisible à l'intérieur d'une espèce, passible d'identification. L'identité comme mêmeté est quant à elle possibilité de *réidentification* : la chose reste la même dans les limites d'un cadre spatio-temporel donné. L'identité comme ipséité, pour finir, est une identité réflexive, une *autodésignation*.

[81] Nous retrouvons ici, le lecteur l'aura compris, l'ipséité qui au centre des analyses de Ricœur mentionnées dans la note précédente.

[82] Pour tout le paragraphe sur l'identité, voir J. DIGIOVANNA, *Artificial Identity*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, pp. 307-321.

organiques… le même agent. Le deuxième aspect, celui de l'identité psychologique, s'appuie en général sur la mémoire (épisodique). L'exigence de la permanence de la mémoire épisodique est notamment mise en relief par la perception identitaire de soi d'un agent[83]. La mémoire, bien sûr, n'est qu'un moyen par rapport à la finalité, qui est celle de pouvoir reconstituer l'historique d'interaction avec l'autre. C'est ainsi que l'homme peut également faire appel à une prothèse technique (tel un agenda) pour renforcer sa mémoire[84].

Ces divers critères sont cependant problématiques étant donné que nous pouvons attribuer une identité continuée à des êtres incorporels, comme l'Allemagne, la République, IBM et ainsi de suite. Il est patent, dans ces différents cas, que l'attribution de l'identité se fait autant dans le chef de l'observateur qu'elle n'est due aux structures intrinsèques de l'être considéré. Ainsi nous pouvons reconnaître une continuité à un État tant que ni son territoire, ni sa constitution ne changent trop radicalement, et que la forme présente de cet État continue à se réclamer de la forme passée de ce « même » État.

Tout particulièrement, la continuité corporelle semble une contrainte à première vue difficile à transposer sur une machine. Après tout, tout le paradigme informatique est construit sur une pensée dualiste, qui dissocie radicalement la matérialité du logiciel. La machine de Turing – qui représente un ordinateur en toute idéalité – permet d'exécuter toute suite de calcul, sans acception de son implémentation matérielle. Et dans les faits, un robot peut subir des changements radicaux dans le temps. Quand nous pensons que l'histoire de l'informatique a commencé il y a à peine 60 ans, force est de constater que tout ce matériel est particulièrement peu robuste : alors que plus aucune machine de cette époque n'est encore en service, les machines biologiques humaines nées à la même époque dirigent aujourd'hui notre monde.

Afin d'honorer l'esprit fonctionnel, il faut donc se tourner vers d'autres critères de continuité. Et nous pouvons nous appuyer sur le fait que si un être humain change radicalement de valeurs au cours de sa vie, il nous est difficile de le reconnaître en tant que *même* personne. Cet état de fait nous autoriserait donc à nous pencher sur la continuité des *projets de vie* et des *valeurs* pour déterminer la permanence d'un agent. Nous retrouvons cette idée chez plusieurs auteurs, dont N. Bostrom, qui parle de l'intégrité des objectifs (*goal-content integrity*) :

> *If an agent retains its present goal into the future, then its present goals will be more likely to be achieved by its future self.*[85]

Il s'avère que ce critère est une condition *sine qua non* d'un comportement cohérent, peu importe le type d'agent auquel nous faisons face[86]. Cette voie semble donc assez prometteuse, dans la mesure où elle permet de subsumer tout type d'agent sous le dénominateur commun de « fil téléologique » (*teleological thread*). Nous retrouvons la même idée dans la littérature consacrée aux voitures autonomes, où la synergie qui s'instaure entre l'homme et sa voiture peut être considérée comme

[83] Dr. McDermott, *What Matters to a Machine?*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, p. 105.
[84] Y. Hui, *On the Existence of Digital Objects*, pp. 147-148.
[85] N. Bostrom, *Superintelligence*, p. 132.
[86] S. Petersen, *Superintelligence as Superethical*, dans P. Lin, R. Jenkins et K. Abney, *Robot Ethics 2.0*, p. 329.

un système hybride ou mieux, comme un « sujet pluriel ». Ce sujet est constitué par un objectif commun, c'est-à-dire par une connaissance partagée par les divers sous-systèmes.

| Questions | |
|---|---|
| **Dans quelle mesure une identité doit-elle reposer sur la mémoire : ne peut-elle pas être simplement expliquée par la persistance des moyens des (sous-)processus/systèmes ?** | |
| **Dans un jeu collaboratif, il ne suffit pas qu'un joueur soit « le même », l'ensemble des interactants doit le reconnaître tel. Comment modéliser la contrainte d'interaction qu'est l'identité ?** | |
| | Suffit-il d'avoir une clef unique dans une table ? |
| | Si l'identité est obtenue par une clef unique, elle équivaut à l'unicité. Les SMA ont-ils d'autres moyens de garantir l'identité ? |
| **Les simulations à base d'agents doivent-elles nécessairement faire appel à l'identité ? Cest-à-dire, peuvent-elles obtenir des résultats sans que les agents puissent se reconnaître entre eux ?** | |
| **Les SMA permettent-ils de problématiser l'identité, de faire varier les critères de continuité ?** | |
| **Comment un SMA représente-t-il le temps, sa progression ?** | |

#### *1.7.2.3. Autonomie*

Le critère d'autonomie se retrouve dans la plupart des listes[87]. Or il se fait que sous ce vocable, les auteurs rangent des choses assez diverses. Il conviendra donc d'être attentif aux nuances. Dans un premier sens, un robot sera dit « autonome » par opposition à un robot téléguidé, ou assisté[88]. Ainsi compris, le robot est autonome de la même manière qu'un enfant qui apprend à s'habiller, s'apprêter, etc., sans l'aide de son entourage. Bref, l'agent s'oppose simplement à l'outil. Cette conception semble très réduite, mais elle permet déjà de voir des agents technologiques se servant eux-mêmes d'outils (pensons à notre propre ordinateur personnel qui se sert de périphériques d'entrée-sortie telle une connexion Internet pour accomplir les recherches que nous lui confions). C'est ce sens minimal de l'autonomie qui a été qualifiée d'autonomie d'action[89].

Un deuxième sens donné à la notion est la faculté d'être opérationnel dans un environnement évolutif. En termes informatiques, nous pourrions dire que l'autonomie est ici le contraire de la dépendance sur l'hypothèse d'un monde clos. Une telle faculté peut encore recouvrir beaucoup de choses, nous remettons une discussion de ce critère à la section consacrée à l'adaptabilité

[87] Une exception notable étant peut-être W. WALLACH et C. ALLEN, *Moral Machines*, p. 32, où ces auteurs distinguent deux axes dans le développement des machines : l'axe de l'autonomie est ainsi totalement indépendant de l'axe de la sensibilité éthique.
[88] J. P. SULLINS, *When Is a Robot a Moral Agent?*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 157-158.
[89] Cf. C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 76.

(§ 1.7.2.10). À ce stade, convient simplement d'observer que ces deux sens – l'autonomie s'opposant à la téléguidance et l'autonomie comme faculté de s'adapter – ne se recoupent pas. Selon le premier critère, une chaîne de montage traditionnelle est plus autonome qu'un drone téléguidé. Selon le deuxième critère, en revanche, le robot téléguidé est plus autonome que la chaîne de montage, dans la mesure où il pourra être utilisé dans des environnements différents, à des tâches variées, grâce justement à l'intelligence humaine qui le dirige.

Dans un troisième sens, qui surenchérit sur le deuxième, l'autonomie se comprend comme une interactivité riche entre l'agent et son environnement : l'autonomie, dès lors, devient la capacité d'évaluer continûment (*monitor*) l'environnement, d'évaluer l'effet de ses actions sur cet environnement, et finalement de s'en servir pour décider de l'action suivante[90]. Un système autonome peut être automodifiant, c'est-à-dire qu'il peut enrichir – ou pour tout le moins ajuster – son répertoire d'actions avec effet sur l'environnement. Compris dans ce sens, le critère de l'autonomie dite rationnelle se rapproche fort de celui de l'interactivité. Si nous nous avançons encore davantage sur la même piste, l'autonomie finit par se confondre avec la prise de décision tout court.

Lorsque nous nous penchons sur l'étymologie de la notion – la faculté de se donner ses propres lois – il est clair que nous avons affaire ici à un concept éminemment politique. Une communauté (et non un individu) sera dite « autonome » si elle est « libre » de se donner elle-même les contraintes légales qui la régissent[91]. Tirons deux remarques de cette étymologie : premièrement, l'autonomie n'est pas la liberté qu'aurait un individu de faire *n'importe quoi*, acception du terme qui est dépourvue de sens et à plus forte raison de pertinence éthiques. Au contraire, l'autonomie revient ici à se doter de ses propres contraintes de fonctionnement interne, à s'auto-organiser en somme. Deuxièmement, ces contraintes prennent la forme d'un élément discursif, la loi, qui peut être abstraite de la communauté qui l'a édictée[92]. Vu sous cet angle, nous retrouvons l'exigence, pour un agent éthique, d'être muni d'une justification *préalable* à son action. L'autonomie, ainsi, est une affaire de la collectivité qui dépasse les individus qui la forment.

Nous nous arrêtons un bref instant sur le sens le plus minimaliste de l'autonomie, celui qui oppose l'agent autonome à l'outil (télé)guidé, pour faire deux observations. Premièrement, l'autonomie est éminemment une affaire de *gradation* : de la simple tondeuse à herbe manuelle – où tout le principe actif se trouve dans l'être humain qui s'en sert – à la tondeuse robot autonome en passant par la tondeuse électrique, des degrés plus ou moins grands d'autonomie existent. Or dès que nous admettons l'idée que la caractéristique n'est pas binaire, notre regard change, et nous nous posons

---

[90] *Ibid.*, p. 159.

[91] Pour une discussion sur l'autonomie et ce qui l'oppose à la liberté ou l'indépendance individuelle, voir l'article d'A. RENAUT, *Liberté*, dans Ph. RAYNAUD et S. RIALS, *Dictionnaire de philosophie politique*, pp. 406-409.

[92] L'originalité de Kant en la matière a été de déplacer l'autonomie législative – apanage d'un peuple – sur l'individu doué de raison : l'homme qui, plutôt qu'à obéir à autrui s'impose ses propres lois, peut être dit dès lors pleinement autonome. L'autonomie ainsi entendue est d'ailleurs la clef de voûte de l'édifice moral kantien : est moral, l'être qui plie sa volonté à l'empire des lois, pour autant que celles-ci sont inspirées par la seule Raison (cf. P. RICŒUR, *Soi-même comme un autre*, p. 245). Nous reviendrons sur l'autonomie kantienne et ses implications éthiques au troisième chapitre, lorsqu'il s'agira de dénouer le nœud gordien entre finalisme (ou téléologie) et déontologie, notamment dans la section consacrée à l'internalisation de la norme (§ 3.3.1).

des questions différentes : quel est le degré *minimum* d'autonomie qu'un agent doit avoir pour être qualifié d'éthique ? Peut-on acquérir toujours *plus* d'autonomie, de façon continue, ou y a-t-il des effets de palier, des seuils qualitatifs infranchissables ?

Une deuxième observation, plus importante encore, est que nous ne pouvons pas perdre de vue notre premier critère – l'unité d'agir – en évaluant l'autonomie : un être humain peut être pleinement autonome pour accomplir une tâche particulière, par exemple il n'a besoin de personne pour nouer ses lacets. En revanche, lorsque la même personne se retrouve guichetier aux chemins de fer, il peut se découvrir très peu autonome pour délivrer des titres de transport lorsque l'ordinateur tombe en panne. L'unité d'agir n'est pas la même dans les deux cas : dans le premier, l'être humain est la source de l'agir ; dans le deuxième, il n'en est qu'un auxiliaire subalterne.

Cette observation peut être portée plus loin. En effet, les outils informatiques sont réputés nous conférer plus d'autonomie pour l'exécution d'un grand nombre de tâches, même de rendre envisageables des tâches qu'auparavant un individu n'aurait jamais pu commencer seul. Or dans la perspective que nous venons d'esquisser, l'évaluation peut être tout simplement renversée ; car l'être humain n'est pas devenu plus autonome : il n'avance plus qu'à l'aide de ses prothèses, de ses béquilles en somme. En dernière analyse, il se pourrait que la question de l'autonomie soit, dans un certain nombre de situations au moins, mal posée. La vraie question serait alors plutôt : quelle est la configuration de relations interobjectives nécessaires pour qu'une unité d'agir soit à même d'accomplir une tâche qui soit attendue d'elle dans une situation donnée ?

Ce déplacement du regard requiert de se concentrer sur l'action à entreprendre, sur le résultat à obtenir, plutôt que sur l'agent en tant que tel. Nous retrouvons ce type de regard dans la modélisation par Matteo Turilli du degré d'autonomie d'un agent[93]. Dans sa modélisation, le degré d'autonomie est déterminé en s'intéressant à chacune des opérations qui composent un processus donné – terme que l'auteur préfère à celui d'action. Une opération, comprise ici comme la relation qui lie un état initial à un état final dans un système donné, peut dépendre de plusieurs processus. L'ensemble des processus dont elle dépend – ne fût-ce que pour une information – est appelé le périmètre de contrôle (*control closure*) de l'opération. Le degré d'autonomie dont dispose un processus sur une opération donnée n'est alors rien d'autre que son rapport, inversement proportionnel, au périmètre de contrôle de cette dernière : si le processus y figure seul, son autonomie peut être dite maximale. Nous trouvons ici une définition de l'autonomie qui soit à la fois graduelle, attentive aux fins à accomplir et aux relations (interobjectives) existantes.

| Questions |
| --- |
| **Les SMA peuvent-ils représenter des cas d'autonomie collective ?** |
| **Est-il possible, en SMA, de mesurer le degré d'autonomie de ses agents ? Si oui, sur base de quels critères cela se fait-il ?** |

[93] M. TURILLI, *Ethical Protocols Design*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 385-387.

### *1.7.2.4. Intentionnalité*

Parmi toutes les notions abordées jusqu'ici, celle d'intentionnalité est peut-être la plus difficile à replacer dans un cadre fonctionnel. Après tout, ne renvoie-t-elle pas, sans reste, à une intériorité dont l'accès nous est par définition barré ? Michel Foucault vient nous rappeler à quel point ce que nous appelons « intentionnalité » ici puise ses origines dans une fabrique historique et culturelle parfaitement retraçable :

> *L'individu s'est longtemps authentifié par la référence des autres et la manifestation de son lien à autrui (famille, allégeance, protection) ; puis on l'a authentifié par le discours de vérité qu'il était capable ou obligé de tenir sur lui-même.*[94]

Calquée comme elle l'est sur la pratique chrétienne de l'examen de conscience, cette pratique qui témoigne d'une fixation sur ce qui n'est pas là, sur ce qui aurait pu ou aurait dû se produire, l'intentionnalité semble vraiment aux antipodes d'une approche qui part du comportement observable ! De surcroît, la pertinence éthique de l'intentionnalité est clairement à la baisse : les théories d'inspiration déontologique ou utilitariste ne s'y intéressent qu'indirectement. En revanche, des théories qui en font une pièce maîtresse, telle que la doctrine du double effet[95], sont loin d'être en odeur de sainteté. L'approche fonctionnaliste a donc quelque raison d'aborder ce critère avec prudence.

Et pourtant, même une théorie fonctionnaliste ne peut faire l'économie du noyau dur de sens véhiculé par l'idée qu'un comportement doit être « délibéré », c'est-à-dire que pour être qualifié de moral, il doit être calculé, organisé, de façon à atteindre un but particulier. Ce calcul, ou cette organisation, serait de préférence explicite, ou retraçable. Bref, un comportement intentionnel renvoie donc à un agent qui poursuit un objectif, mais dans un sens psychologique faible[96]. Explicitons, en effet, ce que « avoir un but à poursuivre » veut dire : tout d'abord, l'agent doit avoir une représentation explicite de son but. Ensuite, Il doit avoir les moyens de mesurer dans quelle mesure les actions qu'il pose le font progresser vers le but qu'il s'est donné. Enfin, il doit être en mesure de prendre des actions correctives si l'atteinte du but est compromise. Si l'on suit ces critères, nous pouvons très bien dire qu'un missile « a l'intention » d'atteindre sa cible[97].

Maintenant que nous avons quelque peu tiré au clair ce qu'un comportement « intentionnel » pourrait vouloir dire dans un contexte fonctionnel, l'ayant dépouillé d'interprétations plus riches peut-être mais peu propices à éclairer le comportement d'un agent, précisons que nous ne posons nullement ici *l'exigence* d'une telle caractéristique. Nous nous contentons de dire qu'un auteur se

---

[94] M. FOUCAULT, *Histoire de la sexualité I*, p. 78.

[95] L'art de la guerre semble être le seul domaine qui continue à s'en réclamer. Ainsi chez R. ARKIN, *Governing Lethal Behavior in Autonomous Robots*, où l'auteur se penche sur l'usage des drones tueurs sur le champ de bataille : aux pages 46-47, une variante de la doctrine du double effet est présentée.

[96] J. P. SULLINS, *When Is a Robot a Moral Agent?*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 158.

[97] Ces critères proviennent de Dr. MCDERMOTT, *What Matters to a Machine?*, dans *op. cit.*, p. 100. Remarquons au passage que cet auteur préfère le vocabulaire de la volonté et la volition à celui de l'intentionnalité.

réclamant d'une approche fonctionnaliste pourrait légitimement y faire appel, non qu'il devrait le faire[98]. La notion est ainsi disponible pour servir à notre examen des SMA et il nous incombera d'en examiner l'usage et la fécondité dans ce cadre.

Et de fait, un auteur comme Luciano Floridi s'oppose plutôt frontalement à l'importance donnée à la « téléologie ». Selon ses dires, ce concept ne convient qu'à une certaine classe d'individus et convient mal à des situations de moralité répartie, qui impliquent des agents éthiques tels que des groupes ou des collectivités. Cela étant dit, parler de l'intentionnalité d'un artéfact peut avoir un sens, quitte à concéder que cette intentionnalité soit étroitement liée à celles de son producteur et de son utilisateur, qui se l'approprient. Ainsi chez Deborah Johnson[99], lorsqu'elle prend l'exemple d'une mine antipersonnel qui tue un enfant au Cambodge : les actions morales les plus pertinentes ici sont la production de la mine, ainsi que son utilisation dans un certain endroit à un certain moment, dans un champ de blé en contexte de guerre civile. L'auteure note que personne n'avait l'intention de tuer tel enfant en particulier. L'intention de l'utilisateur humain ne fait, en quelque sorte, qu'amplifier celle, toute simple et automatique, de la mine elle-même. L'intentionnalité de l'artéfact, en d'autres termes, prolonge l'action morale de l'être humain ; son efficacité produit des effets dans le monde moral, même en l'absence d'indépendance par rapport aux intentionnalités liées. L'utilisation, la création et l'appropriation des artéfacts ne sont pas neutres moralement !

La critique de l'action volontaire n'est d'ailleurs pas absente non plus dans la philosophie des techniques. Simondon[100], par exemple, note que la fin pratique ne dit pas le tout de l'objet technique, car celle-ci peut être réalisée par des structures, par des dynamiques, très différentes les unes des autres. Le *résultat du fonctionnement effectif* d'un individu technique doit ainsi être soigneusement distingué de son effet escompté ou voulu. L'auteur développe surtout les implications épistémologiques qui découlent de cette césure lorsqu'il soutient que pour pénétrer le sens propre de l'objet technique, le chemin d'accès qui y mène ne pourraient être des relations d'usage ou de propriété, ni même de connaissance scientifique ; il y faut le point de vue de l'ingénieur qui, pour ainsi dire, orchestre l'ensemble. D'autres implications, cependant, sont d'ordre proprement éthiques et signifient, en même temps que la limite de l'intentionnalité, son dépassement : comme nous le verrons plus loin, dans la section consacrée à la responsabilité, l'étude anticipatoire des effets non voulus d'une innovation technique peut devenir un devoir éthique à part entière.

| Questions |
| --- |
| **Comment les SMA représentent-ils les intentions des agents, comprises ici comme des objectifs exécutables ?** |
| **Les SMA en ont-ils besoin ? Quels comportements peuvent en faire l'économie ?** |

[98] Voir L. Floridi, *On the Morality of Artificial Agents*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, pp 199-201.
[99] D. G. Johnson, *Moral Entities but Not Moral Agents*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, pp. 178-182.
[100] G. Simondon, *Du mode d'existence des objets techniques*, pp. 14-21.

### *1.7.2.5. Liberté*

Lors d'une promenade, nous entendons quelqu'un crier à l'aide. En nous approchant, nous voyons un malheureux qui se noie dans le fleuve. Quelle attitude allons-nous adopter ? Comme tout un chacun, lorsqu'un tel cas est rapporté, nous aimerions nous voir en sauveur : nous plongerions dans l'eau, nagerions vers le naufragé, le ramènerions sur la terre ferme. Or force est de reconnaître que l'issue pourrait être très différente : nous pourrions être pris de panique, et nous enfuir. Pour la discussion qui nous occupe ici, la question est de savoir si nous étions libre d'agir autrement. La question est pertinente pour les deux issues : étions-nous vraiment libre de sauter dans l'eau, n'avons-nous pas agi de façon irréfléchie ? Et dans l'autre branche de l'alternative, avions-nous *vraiment* la possibilité de ne pas céder à notre peur ? Il importe de souligner que dans les deux cas, nous aurions agi de façon *réflexe*.

Le type de liberté dont il vient d'être question porte un nom : c'est le *libre arbitre*. Le libre arbitre relie la question de la liberté à celle de la *volonté* : est libre, celui qui ne fait pas seulement ce qu'il veut, mais qui est libre de vouloir autre chose que ce qu'il veut actuellement. Même si un tel concept de la liberté a été fort décrié pour être hors expérience, hors preuves, et par là indécidable[101], il n'est pas dénué de pertinence éthique, car d'une part, aucun doute plane sur la qualification éthique de notre liberté de vouloir. Ainsi, pour reprendre l'exemple de la noyade, nous serons le héros du jour dans le premier cas et dans le deuxième, juste un lâche.

D'autre part, vouloir autre chose que ce que nous voulons, cela s'entraîne ! Depuis la plus haute Antiquité en effet, d'Épicure à Sénèque, les différentes éthiques de la vertu ont recours à l'image de l'homme esclave de ses pulsions (ou « passions de l'âme »). Des philosophes contemporains comme Peter Sloterdijk[102] attirent l'attention sur le fait que l'être humain n'existe que sur fond d'habitudes, où l'habitude est définie comme la répétition d'un geste, d'une action, de manière à améliorer la capacité à réaliser l'action à son exécution suivante[103]. La liberté, dès lors, n'est rien d'autre que la faculté de remplacer des mauvaises habitudes par des meilleures. Cette idée est séduisante dans la mesure où elle permet d'expliquer la pertinence éthique de la liberté : le comportement courageux s'entraîne, s'exerce : dans mille occasions de la vie quotidienne, nous aurons eu l'opportunité de mettre l'intérêt d'autrui au-dessus du nôtre propre, ou au contraire de ne penser qu'à nous. Cet « apprentissage » influencera, l'heure venue, les réflexes que nous adopterons. Nous retrouvons ici une idée très ancienne, remontant à Aristote :

> *La [vertu] éthique provient de l'habitude, et c'est d'ailleurs de ce mot (éthos), légèrement modifié, qu'elle tire son nom[104].*

Le libre arbitre n'est toutefois pas le seul type de liberté. Ainsi le concept kantien d'initiative ne fait-il pas référence à la volonté mais à la *causalité* et la puissance d'agir : parce qu'il conjugue, en son agir, plusieurs sortes de causalité, l'homme peut être considéré libre d'intervenir dans le monde.

---

[101] A. COMTE-SPONVILLE, *Petit traité des grandes vertus*, p. 186.
[102] P. SLOTERDIJK, *Du mußt dein Leben ändern*, pp. 639-646.
[103] « […] üben heißt: ein Aktionsmuster so wiederholen, daß infolge seiner Ausführung die Disposition zur nächsten Wiederholung verbessert wird » (*ibid.*, p. 643).
[104] Cité dans Fr. Woerther, *Aux origines de la notion rhétorique d'èthos*, p. 90.

L'initiative donne ainsi le coup d'envoi à une nouvelle chaîne d'évènements causalement déterminés[105]. Plus près de chez nous, des auteurs se sont penchés sur le problème de réconcilier la liberté avec l'idée d'une nature déterminée par des lois, donnant ainsi naissance à des thèses dites compatibilistes. Citons Harry G. Frankfurt, pour qui la liberté comprend à la fois des souhaits ou désirs de second ordre (c'est-à-dire discursifs, avec qui nous nous identifions en tant que personne) et la capacité d'agir selon ces désirs[106]. Ou encore Daniel Dennett, pour qui il s'agit en fin de compte de la capacité de considérer diverses options et de choisir entre elles[107].

Il est donc possible de dégager une notion minimale et fonctionnaliste de l'idée de liberté : le pouvoir de distinguer entre plusieurs actions possibles et de choisir l'une d'elles au terme d'une délibération éclairée par la connaissance, en l'absence d'une contrainte extérieure. Cette notion est minimaliste ; elle est fonctionnelle aussi : en effet, avant de choisir, parmi les pistes possibles, celle qui servira le mieux notre but, il faut que nous soyons en mesure d'avoir « à l'esprit » tout l'éventail des alternatives que nous pourrions choisir.

Cette notion minimale garde son sens même pour des auteurs qui privilégient l'idée de la répétabilité de l'action éthique : des auteurs comme Anderson tiennent en effet la position que la liberté n'est pas requise pour résoudre un dilemme éthique : dans une approche fonctionnelle, où l'on juge par les résultats de l'action, dans les mêmes circonstances le même geste sera toujours considéré bon ou mauvais de la même manière. C'est – selon elle – une simple reformulation de l'exigence de la cohérence[108]. Même dans un tel cadre, où l'auteure prône le déterminisme, l'agent qui se veut éthique ne doit pas avoir connu une seule solution. Avant de lancer son algorithme, plusieurs alternatives doivent être évaluées, quitte à ce que dans un certain contexte, un certain jeu de paramètres donne toujours la même issue à l'algorithme.

#### *1.7.2.6. Responsabilité*

De toutes les notions abordées ici, la « responsabilité » est peut-être celle qui a connu la plus grande évolution dans son champ d'application[109] : pour nous en convaincre, il convient d'abord de rappeler que la responsabilité se représente logiquement par une relation ternaire : *X est responsable de Y devant Z*, où X est le *porteur* de la responsabilité, Y *l'objet* de la responsabilité et Z *l'origine* ou le *fondement* de la responsabilité. Or parmi les porteurs de responsabilité ne figuraient pendant longtemps que des sujets (« un homme responsable »). Ensuite, la notion a été étendue pour pouvoir désigner l'action d'un sujet (« un achat responsable »), pour finir par pouvoir être prédiqué d'un objet quelconque (d'où l'emploi de « matériaux éco-responsables »). Laissons de côté les responsabilités des matériaux et des actions – qu'il est toujours possible de tenir pour des emplois dérivés – afin de nous concentrer sur celles qui s'appliquent aux seuls sujets. Un sujet peut être responsable d'une

[105] Cf. P. RICŒUR, *Soi-même comme un autre*, p. 124-136.
[106] Cf. C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 123.
[107] Cf. W. WALLACH et C. ALLEN, *Moral Machines*, p. 60.
[108] S. L. ANDERSON, *Philosophical Concerns with Machine Ethics*, dans M. ANDERSON et EAD., *Machine Ethics*, p 164.
[109] La remarque est faite par J. VAN DEN HOVEN, *Value Sensitive Design and Responsible Innovation*, dans R. OWEN, J. BESSANT et M. HEINTZ, *Responsible Innovation*, p. 81.

action concrète, d'une tâche, d'un rôle, ou même d'une autre personne[110]. Son champ d'application est donc très vaste.

En plus d'impliquer une grande variété d'êtres, la responsabilité est une notion qui a cours dans des domaines de discours divers : en dehors de l'éthique, elle est d'usage très courant dans le monde juridique. Face à une telle richesse d'acceptions et d'emplois, la prudence s'impose : quel noyau de sens est-il souhaitable de considérer pour notre propos ? Afin de répondre à cette question, nous avons heureusement un guide sûr dans la personne de Paul Ricœur, qui a consacré plusieurs pages à fouiller l'archéologie du concept[111]. Ricœur part de la question qui nous a déjà retenue dans la section consacrée à l'unité d'agir (§ 1.7.2.1) : comme « ascrire » une action à sa source ? *L'ascription* est une opération qui consiste à attribuer un prédicat – psychique ou physique – à une personne. Elle n'est pas de nature discursive. Lorsque nous ajoutons une appréciation éthique à l'ascription, nous jugeons un agent louable ou blâmable pour une action permise ou non permise. Ce type de jugement qui s'ajoute à l'ascription est une *imputation*. En plus d'être une énonciation, l'imputation présuppose un lien causal : l'action doit dépendre de l'agent. L'imputation doit à son tour être distinguée de *l'incrimination* : alors que l'imputation est pour ainsi dire un acte privé, l'incrimination est d'emblée un acte institutionnel, qui rend l'agent jugé passible de récompenses ou, au contraire, de châtiments et de dommages-intérêts.

La responsabilité, quant à elle, ne surenchérit pas sur l'incrimination, mais sur l'imputation, tout en se déployant sur un horizon temporel. En effet, il n'y a responsabilité que grâce à la fidélité du sujet à soi[112] : ce n'est que parce que nous souhaitons stabiliser notre identité qu'il est possible d'investir éthiquement les répercussions de nos actions dans le monde. Si la responsabilité se projette dans le passé, elle nous invite à assumer un héritage qui nous affecte sans être pour autant notre œuvre. Une telle conception rétrospective de la responsabilité a partie liée avec l'idée d'une dette[113]. Cependant, c'est dans l'idée d'une responsabilité prospective que la notion reçoit probablement, de nos jours, son principal intérêt éthique : elle nous invite d'assumer les conséquences de nos actions dans le futur, sans tenir compte des intentions qui ont présidé au choix de nos actions. Il faut penser ici à l'œuvre de Hans Jonas, qui a fait de la responsabilité pour l'avenir le principe moteur de son éthique[114]. Nous avons la responsabilité non pas primairement de notre comportement passé et de ses conséquences présentes, mais des choses qui, *parce qu'elles sont dans la sphère de notre puissance d'agir*, revendiquent notre agir. Dès l'instant où le contexte nous confère un pouvoir ou un contrôle sur autrui, nous avons l'obligation d'assurer sinon son bien-être, pour tout le moins la possibilité de son existence continuée.

---

[110] Le lecteur intéressé peut se référer à H. LENK, *À propos de risque et de responsabilité*, dans C. KERMISCH et G. HOTTOIS, *Techniques et philosophies des risques*, pp. 39-44, pour une typologie complète de la responsabilité.
[111] Voir P. RICŒUR, *Soi-même comme un autre*, pp. 109-136, 337-344.
[112] Fidélité à soi que Ricœur appelle *ipséité*. Nous avons déjà rencontré la notion lors de la discussion consacrée à l'identité (§ 1.7.2.2), nous la verrons encore à l'œuvre en abordant la question de la valeur (§ 1.8.1).
[113] Dans une même optique, E. LEVINAS définit la responsabilité comme une sensibilité, ou une possibilité de sensibilité, à l'égard de malheurs qui ne commencent pas dans notre liberté ou notre présent. (*Autrement qu'être ou au-delà de l'essence*, pp. 183-184).
[114] Nous nous fondons ici sur H. JONAS, *Le principe responsabilité*, pp. 182 et suivantes.

Deux exemples de l'auteur nous intéressent au premier chef, car il s'agit de métiers automatisables : le chauffeur, à qui les circonstances ont confié la garde d'un certain nombre de passagers, exerce par là même un certain pouvoir sur eux ; il a dès lors le devoir de les transporter en toute sécurité. Une remarque similaire peut être faite pour le capitaine d'un navire : il doit mener toutes les personnes à bord à bon port. Sachons gré à Jonas d'avoir si logiquement arrimé la responsabilité au pouvoir causal d'intervenir dans le monde. Il s'ensuit que tout agent qui est capable de tenir un rôle social, à comprendre comme des *devoirs* ou des *charges*, peut être dit exercer des responsabilités. Suivant cette ligne de pensée, il n'est pas absurde d'attribuer « l'agentivité morale » à un robot qui remplit la condition de pouvoir causal :

> *We can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behavior by assuming it has a responsibility to some other moral agent(s).* [115]

C'est peut-être aller un peu vite en besogne ; en tout cas, aller plus loin que le texte de Jonas nous autorise à nous aventurer, car l'auteur a en vue, de bout en bout de son texte, des agents humains, qu'il convient de préserver des affres d'un progrès technologique en surchauffe. De surcroît, le pouvoir n'est pas le seul ressort de la responsabilité chez Jonas : un être responsable doit aussi mettre tout en œuvre pour prévoir les conséquences de son action, d'anticiper toutes les possibilités, même celles qui lui paraissent de prime abord peu probables. Pouvoir causal et *savoir anticipatoire* vont de pair.

Et pourtant, le critère de la responsabilité sociale que nous rencontrons en éthique des machines semble toucher quelque chose de profondément juste. Pour le voir, faisons un pas en arrière et reconsidérons l'idée même d'agentivité : peut-elle être considérée comme une propriété fonctionnelle, à l'instar du vol, partagé entre l'oiseau et l'avion ? Dans un certain sens, la réponse est clairement affirmative. Prenons, pour nous en convaincre, l'exemple du cabinet du Ministre du budget : peu importe sa composition ou son lieu d'installation, il continuera à émettre des accords budgétaires quand bien même tous les cabinettards, jusques et y compris le Ministre lui-même, auront été remplacés par d'autres. Dès lors, est un agent fonctionnel, tout « système » (biologique, social, ou autre) qui *remplit un certain rôle dans un environnement*. L'exigence d'une responsabilité sociale renvoie ainsi en dernière analyse à une compréhension fonctionnelle de l'être humain.

| Question |
| --- |
| **Quels types de rôle social les agents d'un SMA peuvent-ils endosser ?** |

[115] J. P. SULLINS, *When Is a Robot a Moral Agent?*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 159.

### *1.7.2.7. Interactivité*

L'interactivité exige que l'agent et son environnement puissent agir l'un sur l'autre. À cette définition sommaire, il convient d'apporter immédiatement une précision cruciale : l'interaction entre l'agent et l'environnement doit se situer au niveau des observables du niveau d'abstraction considéré[116]. Ainsi entre notre corps et l'environnement, un va-et-vient incessant de vies minuscules peut bien venir troubler, aux yeux de l'observateur muni d'un microscope, les frontières entre les systèmes biologiques, il serait pourtant bien curieux qu'un malade s'en prenne aux bactéries pour leur discerner un blâme moral. Aussi, par rapport à l'unité d'agir considérée, ce critère exclut-il tant le « trop petit », comme les bacilles, que le « trop grand », comme les collectivités, la nature, ou encore le climat.

Ce critère est important, car il s'inscrit en ligne droite dans la visée principale d'une certaine conception de l'intelligence artificielle, qui est de multiplier les possibilités d'interaction entre l'homme et la machine. Pour le cerner au mieux, abordons ce critère par le type d'interaction dont nous avons l'expérience la plus familière, le cas de l'interaction d'être humain à être humain, celle-là même que nous avons l'habitude de qualifier de sociale. Une première remarque à faire, c'est que nous devons assumer la sociabilité de l'autre, que nous lui accordons « par défaut » tant que l'autre continue à se montrer « normal », c'est-à-dire, socialement, à nos yeux.

Une telle attribution peut être étendue dans le cas de l'interaction entre homme et machine, pour autant que le robot se montre socialement normal. Précisons qu'une telle attribution de sociabilité par l'homme à la machine peut être purement fonctionnelle : l'être humain n'attribue aucunement, par-dessus le marché pour ainsi dire, vie ou conscience, émotion ou intelligence, au robot. Sur le plan du vécu, l'attribution de sociabilité suffit pour qualifier une interaction de sociale, qui devient dès lors une sorte de *jeu*[117]. Ce phénomène répond au doux nom de « l'effet Eliza », en référence à l'application du même nom conçue par le docteur Weizenbaum pour simuler une séance de thérapie. Cet épisode révèle qu'il peut être plus facile de parler à une machine parce qu'elle ne comprend *pas* ce qui lui est dit : le fait même de parler, en d'autres termes, crée du sens[118].

Le jeu interactif a un sens, cela paraît évident : dès que la machine nous montre de l'intérêt, c'est une façon de pousser sur nos « boutons darwiniens » : elle se signale comme apte à l'interaction, comme entité relationnelle. Ce pouvoir d'évocation tient cependant autant, voire plus, à nos vulnérabilités qu'aux capacités intrinsèques de la machine. Des simulations de gestes de sollicitude commandent des réactions affectives très vite et très fortes. Si la machine se montre dépendante, réclame des soins, elle peut aller jusqu'à nous construire comme ses parents. Un autre exemple est celui des robots persuasifs[119], dans un contexte de vente. Un tel robot peut persuader sans être lui-même persuadé, et adopter intentionnellement certaines stratégies de persuasion. Ainsi, il pourrait simuler la sympathie en prenant une voix plus douce, ou en jouant sur la proximité corporelle. Ces effets de

---

[116] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p 193.

[117] Tout le développement qui précède est basé sur S. PAYR, *Towards Human-Robot Interaction Ethics*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, p. 33.

[118] Sh. TURKLE, *Authenticity in the Age of Digital Companions*, dans M. ANDERSON et S. L. ANDERSON, *op. cit.*, pp. 63-70.

[119] Exemple repris à J. HAM et A. SPAHN, *Shall I Show You Some Other Shirts Too? The Psychology and Ethics of Persuasive Robots*, dans R. TRAPPL, *op. cit.*, pp. 65-77.

persuasion, mettant en œuvre des moyens simples, peuvent se faire sentir même si l'interlocuteur du robot persuasif ne lui reconnaît qu'un très faible degré d'indépendance, ou d'agentivité. Un robot peut même aller jusqu'à se faire pardonner ses erreurs ou à réclamer un traitement aménagé, pourvu qu'il fasse montre de réactions sociales plutôt que factuelles : donner des signes d'épuisement, d'inconfort, ou simplement s'excuser après une mauvaise prédiction… voilà autant de moyens pour faire passer un message bien plus sûrement que par des arguments rationnels ! Il est important de souligner que ces effets sont préconscients : aucun avertissement, aucune mise en garde n'y peut rien. L'effet de persuasion peut être grand quand bien même nous ne nous en rendons pas compte.

De tout ceci, il appert toutefois avec netteté que l'interactivité affective ne naît pas tant des capacités intrinsèques du robot que de nos besoins. Au contraire même, un robot peut nous inciter à répondre d'autant plus affectivement qu'il paraît plus artificiel : sa gaucherie peut être estimée attachante, et moins il paraît réel, moins il paraît menaçant, pour peu que nos représentations symboliques soient satisfaites : il lui faut tout de même des yeux, des mimiques, etc., quitte à ce que tout ceci soit stylisé au plus haut point[120]. Toujours est-il que la flatterie, les expressions faciales… toutes ces choses suscitent quasi « automatiquement » des réponses sociales de la part des interlocuteurs humains.

La question se pose : faut-il comprendre ce jeu d'interactivité asymétrique comme simulation ou duperie, manque d'authenticité et consommation narcissique, ou comme un jeu au sens véritable du terme, une occasion d'apprentissage social, voire comme une thérapie ? Car il est manifeste que l'être humain a un besoin vital de se sentir accepté, d'appartenir à un groupe : la relation à l'autre prime sur le besoin de l'autonomie[121]. Le sens qui en découle est attribué unilatéralement. Ceci n'empêche évidemment pas de faire justice à l'idée intuitive que l'interactivité est aussi affaire de degré : nous pouvons *plus* interagir avec notre Eliza, caché(e) à l'intérieur d'un écran, qu'avec un poisson rouge dans toute la transparence de son bocal.

Élargissons maintenant le propos : il semble bien difficile de dire qu'un agent, humain ou non, quel que soit le degré d'abstraction de son unité, soit *intrinsèquement* interactif. L'interactivité se joue dans la relation entre deux systèmes, relation que les deux systèmes en présence l'un de l'autre vont probablement investir de sens chacun à leur manière. Certes, ce n'est pas la première fois que nous constatons qu'un critère ne prend tout son sens que dans ce qu'il convient d'appeler une *relation*, un rapport de l'agent que nous considérons vers ce qui est *hors* de lui, mais cette spécificité ne nous est jusqu'à présent point parue avec autant d'acuité.

Ceci étant dit, il ne faut pas perdre de vue que l'interactivité, dans un environnement donné, pose des difficultés non négligeables à la machine et ceci pour une raison simple : alors que l'interactivité a tout pour paraître évidente aux êtres interactifs que nous sommes, elle est tout sauf naturelle à la machine : celle-ci doit avoir conscience non de soi, mais de son environnement. Considérons une voiture à conduite autonome : afin d'interagir correctement dans son environnement, elle doit tout d'abord être équipée d'une myriade de capteurs pour enregistrer l'activité extérieure. Cette information ne suffit cependant pas : elle doit ensuite faire le tri entre toutes ces informations pour identifier les entités pertinentes : d'autres voitures sur la route, par exemple, doivent être évaluées

[120] G. CHAPOUTHIER et Fr. KAPLAN, *L'homme, l'animal et la machine*, pp. 102-103.
[121] S. PAYR, *Towards Human-Robot Interaction Ethics*, dans R. TRAPPL, *op. cit.*, p. 40.

sur leur vitesse, leur direction, etc., alors que l'oiseau conquérant le ciel peut être ignoré. Toutes ces entités doivent ensuite être évaluées par rapport au code de la route et aux intérêts du conducteur. Ainsi faut-il reconnaître son environnement, l'investir d'un certain sens et le prendre en compte en vue d'une décision[122]. C'est dire que plus la machine intelligente prend conscience de son environnement, plus elle pourra faire jouer les ressorts de l'interactivité.

| Questions |
| --- |
| **Dans quelle mesure les agents en SMA interagissent-ils ? Dans quelle mesure leur comportement est-il influencé par la présence d'autrui ?** |
| **À quels types d'information les uns sur les autres les agents en SMA accèdent-ils ?** |

#### *1.7.2.8. États internes*

Selon un mot fameux de Derrida, un automate peut réagir, non répondre[123]. En des termes empruntés à la théorie de l'action, Deborah Johnson[124] affirme qu'une action ne peut être comprise que comme un comportement motivé intentionnellement, c'est-à-dire par des états mentaux internes. Parmi ces états internes, citons l'intention d'agir (*intending to act*), qui, plus que l'intentionnalité à proprement parler (c'est-à-dire l'intention de réaliser un certain but), se trouverait à la base de la responsabilité sociale.

Face à ce genre d'argument, des auteurs fonctionnalistes ont bien sûr beau jeu de mettre en garde contre ce qu'ils appellent des « spéculations » psychologiques[125]. Une autre objection est évidemment de contester l'usage plus qu'hâtif qui est fait ici de la théorie de l'action, à qui l'auteure fait tenir un rôle de policier que celle-ci ne prétend pas nécessairement occuper. La vraie difficulté est cependant plus profonde : c'est qu'on dénie à la machine une chose, des états internes, alors que la notion même d'état mental interne ou, pour parler avec les psychologues cognitivistes, de *représentation*, est en réalité une hypothèse inspirée d'une métaphore computationnelle de l'esprit[126]. Alors que la notion fait débat en psychologie[127], la pertinence de celle-ci est établie de longue date dans des disciplines telles que la physique ou l'informatique, où elle se trouve même au centre des préoccupations.

Là où l'approche fonctionnelle peut se permettre de mettre entre parenthèses l'intention d'agir pour se concentrer sur les comportements, d'autres approches vont plus loin dans leur critique des états

---

[122] Cf. Th. RAMGE, *Mensch und Maschine*, pp. 13-16.
[123] Ce mot est rapporté par D. J. GUNKEL, *The Machine Question*, 2012, p. 60.
[124] D. G. JOHNSON, *Moral Entities but Not Moral Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 173-175.
[125] Par exemple, L. FLORIDI, *On the Morality of Artificial Agents*, dans *op. cit.*, pp 199-200.
[126] Voir les pages extrêmement intéressantes que consacre D. ANDLER au *projet des sciences cognitives*, dans ID., A. FAGOT-LARGEAULT et B. SAINT-SERNIN, *Philosophie des sciences I*, p. 256-296.
[127] Retracer, en ces pages, la querelle psychologique qui entoure la représentation déborde le sujet du présent mémoire. Le lecteur intéressé par cette question peut se référer à J.-M. GALLINA, *Les représentations : un enjeu pour les sciences cognitives*, dans N. BAULT et autres, *Peut-on se passer de représentations en sciences cognitives ?*, pp. 20-23.

internes, allant jusqu'à nier qu'il y ait autre chose que le comportement. Prenons le cas des soins à la personne (*care*)[128] : les soins dispensés par une infirmière sont constitués d'un ensemble de gestes et de mouvements qui expriment une attention particulière aux besoins et vulnérabilités du patient. Il arrive souvent que le malade investisse cette attention du sens de « sollicitude », et il convient alors de s'entendre sur le statut et l'origine de ce sens. Dans une première approche, parfois qualifiée d'énactiviste, le sentiment de sollicitude émane d'une création de sens, privée et interne au sujet qui l'éprouve. C'est dire que le courant énactiviste soutient que le sujet projette, par une inférence analogique, ce qu'il ressent sur celui (ou celle) qui dispense les soins.

Une deuxième forme de critique est plus radicale encore : la condition pour que ce sens puisse voir le jour ne réside pas dans la réciprocité des émotions ou des états internes, ni non plus dans la projection d'un état interne sur l'autre, comme dans le courant énactiviste. Au contraire, la colère *est* le comportement qui lui est associé. En suivant ce raisonnement, la sollicitude ne découle pas des attitudes et gestes d'assistance, mais n'existe que dans le contexte même des soins. L'approche, dès lors, n'est ni fonctionnelle ni même énactiviste, mais phénoménologique[129] : elle prend appui sur une thèse de Merleau-Ponty, qui soutient que la colère ne se trouve pas, de façon somme toute mystérieuse, dans l'esprit de celui qui agit colériquement.

Ces critiques vont clairement plus loin que l'approche fonctionnaliste adoptée ici : la différence principale se situe dans le statut accordé au primat du comportement. Ce primat est purement méthodologique dans le cas du fonctionnalisme ; l'approche phénoménologique, en niant qu'il y ait autre chose qu'un comportement (fût-il en contexte) l'érige en primat de principe. L'approche phénoménologique est, à ce titre, compatible avec le fonctionnalisme, et permet de radicaliser l'interrogation : dans la modélisation des agents, pouvons-nous reporter toute la signification sur le contexte, et vider les agents de toute notion d'état interne ?

Les systèmes multi-agents ont peut-être de quoi nourrir cette réflexion. C'est ainsi que Luciano Floridi modélise la notion d'agent en recourant à une métaphore de machine d'états[130]. Un état y est qualifié d'*externe* lorsqu'il y a une entrée de, ou une sortie vers, l'environnement. Si tel n'est pas le cas, l'état est qualifié d'*interne*. La notion d'état interne est ainsi opérationnalisée, formellement définie, de manière à nous permettre d'interroger un système multi-agents : quelle est la portée des états

---

[128] D. MEACHAM et M. STUDLEY, *Could a Robot care? It's all in the Movemen*t, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, pp. 97-104. Le terme anglais *care* recouvre (au moins) deux choses que nous aimerions pourtant distinguer : la première est un comportement, les soins donnés à une personne dans le besoin, la deuxième est un état d'esprit, que nous traduisons ici par « sollicitude ».

[129] Certains lecteurs pourraient se sentir perdus, car la problématique mentionnée ici ressemble à celle du comportementalisme logique inaugurée par Gilbert Ryle, issue du courant de la philosophie du langage ordinaire et notoire détracteur de la phénoménologie. Selon sa doctrine, un sentiment comme la sollicitude n'est rien d'autre qu'une disposition à agir selon une certaine façon en fonction des données environnementales reçues. Nous retrouvons donc un souci similaire à Merleau-Ponty, dans la mesure où il refuse la dichotomie cartésienne du corps et de l'esprit et veut réinscrire le sens dans le corps (voir, à ce propos, Gabrielle JACKSON, *Skill and the Critique of Descartes in Gilbert Ryle and Maurice Merleau-Ponty*, pp. 63-78). La référence à l'école du langage ordinaire s'impose certainement pour les linguistes pragmaticiens, qu'il ne faut pas convaincre que le code linguistique est, en soi, sous-déterminé, et que les interlocuteurs construisent l'essentiel du sens en essayant de mettre le code et le contexte en adéquation.

[130] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp 191-192.

internes, quel type de comportement est, ou n'est pas, modélisable par des stimulations directes venant de l'environnement extérieur ?

| Questions |
| --- |
| **Quel type de comportement est – ou n'est pas – modélisable par des stimulations directes venant de l'environnement extérieur ?** |
| **Y a-t-il des types de comportements qui peuvent se passer de toute représentation d'état interne ? Les SMA auraient-ils les moyens de mettre cette idée à l'épreuve ?** |

#### *1.7.2.9. Conscience de soi*

Il peut paraître très imprudent, de prime abord, pour une approche fonctionnelle, de se risquer au traitement d'un thème aussi chargé que la conscience de soi. Avant d'entrer dans le vif du sujet, faisons un pas en arrière et demandons-nous quelle serait une approche fonctionnelle non de la conscience *de soi*, mais de la conscience *tout court*.

Or il se fait que sous l'influence des théories modulaires de l'esprit – dominantes aujourd'hui dans la psychologie cognitive et les neurosciences – la conscience a acquis un rôle fonctionnel bien précis : si une information « accède » à la conscience, toutes les fonctions cognitives supérieures y ont accès et peuvent – doivent – en faire l'objet de leur attention. C'est ainsi que nous pouvons entendre tous les jours mille bruits autour de nous sans y faire attention. Cependant, dès qu'une voix humaine nous interpelle, ou un bruit suspect, tel un coup de frein sec d'une voiture un rien trop près, tout notre esprit est en éveil : nous nous retournons sur la source du bruit, tendons l'oreille, mobilisons toutes nos ressources intellectuelles et affectives pour décoder le message d'autrui et pour y répondre au mieux.

Dans les travaux en intelligence artificielle, cette idée à reçu une place de choix dans le cadre de ce qui est appelé les architectures cognitives[131]. Il s'agit de systèmes qui veulent se calquer le plus fidèlement possible sur le fonctionnement de l'esprit humain. Le pari, propre aux démarches descriptives, est d'approcher ainsi au plus près le comportement humain ordinaire. Dans un tel système, la conscience est une sorte *d'état global*, dont le contenu est rendu disponible à l'ensemble des processus cognitifs, dont les informations sont par ailleurs cloisonnées (*encapsulated*). Sans entrer ici dans des détails excessivement techniques, contentons-nous de quelques mots quant au fonctionnement concret d'une telle architecture, en l'occurrence LIDA. Des codelets d'attention créent des structures dites de coalition à partir de contenus dans la mémoire de travail. Il s'ensuit une véritable compétition pour arriver à la conscience basée sur le contraste (*salience*) : l'information

[131] Toute la présentation technique de l'architecture cognitive LIDA provient de T. MADL et St. FRANKLIN, *Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, pp. 141-149. Les retombées pour l'éthique des machines sont tirées de W. WALLACH et C. ALLEN, *Moral Machines*, pp. 172 et suivantes.

la plus urgente, importante, nouvelle, inattendue… « gagne ». L'espace de travail global se charge ensuit de diffuser (*broadcast*) les contenus gagnants.

Ce qu'il convient de retenir de cette manière de voir, c'est que la conscience peut jouer un rôle fonctionnel, influer concrètement sur nos représentations et comportements quotidiens. Il reste cependant à déterminer dans quelle mesure la conscience, ainsi définie, est pertinente d'un point de vue éthique : n'est-elle pas reléguée à un processus d'attention d'un niveau cognitif finalement assez bas ? De fait, l'exemple de l'architecture cognitive LIDA pointe dans ce sens : les contraintes que nous y observons restent au niveau strictement réactif. De l'aveu même des auteurs, implémenter des règles utilitaristes dans un tel système se heurterait très vite aux limites computationnelles même des machines contemporaines. Le comportement de l'aide-soignante qui implémente LIDA – qui porte le nom de *Carebot* – n'a en dernière analyse rien de spécifiquement éthique. L'exemple montre seulement qu'imiter un comportement humain, quel qu'il soit, constitue un problème « IA-complet ». Dit simplement, pour s'acquitter correctement d'une telle tâche, une machine doit être dotée d'un appareil cognitif au-delà de ce qui est actuellement disponible.

La conscience, en tant que telle, est donc peut-être bien un processus cognitif de base. Il demeure que la conscience *de soi* seule peut toujours être créditée d'une pertinence éthique. Or la conscience de soi, dans le cadre esquissé ici à larges traits, reviendrait pour une machine à avoir une représentation d'elle-même dans son état global. Cette représentation devrait en outre être reliée à la mémoire épisodique de la machine. Il s'en suivrait un type de proto-identité embryonnaire, qui permettrait à la machine d'évaluer des questions d'un genre nouveau : que s'est-il passé dans « mon » histoire, que *m*'est-il déjà arrivé à moi-même ?

Une telle acception de la conscience de soi permet d'opérationnaliser le concept en contournant le problème dit « des autres esprits » : hérité de Kant, ce problème prend parfois des formes très radicales, notamment chez Gunkel, qui y a recours pour condamner sans appel toute interrogation sur ce qui se passe dans l'esprit d'autrui. Il va sans dire que formulé de façon aussi dogmatique, le problème revient à rejeter toute la science cognitive contemporaine. En outre, contre Kant, il est tout à fait possible d'affirmer, avec Aristote, que notre perception et notre conscience des choses sont contraintes par la chose elle-même : une science des contraintes que la chose nous oppose est donc possible.

C'est à cet endroit qu'il convient d'invoquer la distinction faite par le philosophe Ned Block entre la conscience d'accès et la conscience phénoménale[132]. La conscience d'accès est de nature cognitive : c'est la sorte de conscience que nous avons d'une information lorsqu'elle fait l'objet d'une délibération explicite, d'un contrôle rationnel, en vue d'une prise de décision quelconque. En revanche, la conscience phénoménale concerne notre vécu subjectif : que cela nous fait-il de voir des chaussures rouges plutôt que noires, ou d'éprouver une sensation de douleur ? Des deux formes de conscience, c'est bien la deuxième, la conscience phénoménale, qui est frappée, en vertu du problème des autres esprits, d'ineffabilité.

[132] Distinction reprise par C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 35.

Si tant est qu'un système multi-agents soit à même de donner corps à un type d'accès de conscience de soi, des interrogations très riches s'ouvriraient dès lors à notre examen : la conscience de soi donne-t-elle lieu à des comportements qualifiables – par métaphore si l'on veut – d'égoïstes ? Ou, au contraire, pourrait-elle contribuer à ouvrir la machine à une sensibilité éthique décuplée ?

| Question |
| --- |
| **Un SMA peut-il tirer profit d'une histoire personnelle de ses agents ?** |

### *1.7.2.10. Adaptabilité et apprentissage*

Dans *Phèdre*, Platon nous fait part d'une réflexion intéressante de Socrate[133] : celui-ci, lorsqu'un disciple lui demande si le texte écrit est intelligent, répond par la négative, car s'il est vrai qu'en apparence il contient des trésors de sagesse, il reste muet quand on l'interroge : il répétera inlassablement la même chose à l'infini. Ce type de réflexion sur l'intelligence a sa contrepartie en éthique. L'enjeu est alors le suivant : la machine pourrait-elle faire autre chose que ce qu'elle fait ? Précisons de suite la question : faire autre chose, non par lubie ou caprice, mais parce que la machine a obtenu, d'une manière ou d'une autre, l'information que son action n'avait pas le résultat désiré ?

Lorsqu'il s'agit de rendre compte de la capacité des machines en la matière, la discussion est parfois close un peu hâtivement, même avant d'avoir été ouverte. La capacité d'apprendre est déniée à la machine sous prétexte que son code applicatif ne change pas. L'argument est spécieux, car il repose sur une méprise de niveau d'abstraction : si adaptation il y a, cela se manifeste au niveau du comportement, ou pour parler en termes informatiques, à l'exécution du programme. Même si au niveau du code, aucune adaptation n'a lieu, cela n'empêche qu'au niveau de l'exécution, la machine peut faire montre d'apprentissage. Il convient de voir qu'il s'agit là de deux niveaux d'abstraction différents[134].

À la vérité, beaucoup de réussites de l'intelligence artificielle, même si elles sont capables de certaines prouesses de l'esprit comme battre le champion du monde aux échecs, ne font preuve que d'une capacité limitée d'apprentissage. Pour rester sur l'exemple des joueurs d'échecs artificiels[135], ils reçoivent le jeu « clefs en mains » : toutes les règles ont déjà été codifiées. Ils ne savent rien faire, ni rien apprendre d'autre. Au contraire, l'être humain est obligé de passé par un apprentissage et des efforts d'abstraction, ne fût-ce que l'identification des entités des échecs à partir des indices physiques qui peuvent être extrêmement variables. De la même façon, en empruntant le même chemin qui de l'intelligence mène à l'éthique, les concepts éthiques peuvent être codés de la même façon (*stick-built*), c'est-à-dire entièrement sous contrôle humain.

---

[133] Rapportée par D. J. GUNKEL, *The Machine Question*, p. 138.
[134] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 196.
[135] L'exemple et son interprétation sont empruntés à J. STORRS HALL, *Ethics for Self-Improving Machines*, dans *op. cit.*, pp. 513-514.

Comme à l'accoutumée, commençons par nous pencher sur le sens précis de l'apprentissage. À un niveau très minimal, c'est la capacité de collecter l'effet de ses actions sur l'environnement et de s'en servir pour décider de l'action suivante[136]. Ce sens d'apprentissage est assez restrictif : il s'agit finalement de l'inhibition ou de la stimulation de certains comportements à l'aide de punitions ou de récompenses. Il est cependant possible d'aller plus loin. Ainsi, un système peut être automodifiant, en enrichissant son répertoire d'actions avec effet sur l'environnement, c'est-à-dire que ses interactions peuvent l'amener à changer les règles de transition par lesquelles il change d'état[137]. En ce sens, nous voyons bien que l'adaptabilité se rapproche d'un sens que nous avons déjà donné à l'autonomie individuelle et que, dans beaucoup de cas, l'apprentissage se comprend comme une forme un peu particulière d'interactivité. Interactivité, car l'apprentissage ne se fait jamais seul : sans prétendre épuiser ici cette question, nous pouvons cependant observer qu'un aspect essentiel de l'apprentissage humain consiste à *imiter* le comportement d'une *personne de référence* (un parent, un entraîneur)[138], dans l'espoir d'égaler, voire de surpasser, sa compétence. Imiter quelqu'un, parce que nous avons le *désir* de ressembler à ce quelqu'un. Vu sous cet angle, un apprentissage – éthique ou non – a aussi partie liée avec l'émotion : changer de comportement, cela ne se fait-il pas toujours sous l'influence de quelqu'un qui souffre ou qui punit ? Nous sommes loin d'une capacité solipsiste du sujet humain, qui tirerait des comportements inouïs de son propre fonds.

Nous pouvons nous poser la question des *contenus* nécessaires à un apprentissage éthique. Dans le comportement éthique, quelle est la part de l'apprentissage de règles éthiques explicites, de nature discursive, et quelle est la part d'un apprentissage implicite ? La question est loin d'être triviale. Lorsque nous apprenons à jouer d'un instrument ou aux échecs, très souvent l'explication des règles est vite faite. Cependant avant de devenir « bon » dans l'activité considérée, de grands efforts de pratique et d'entraînement sont nécessaires. Ce type d'apprentissage est moralement significatif, également. Il s'agit en effet d'une pierre angulaire de toute éthique de la vertu de facture traditionnelle : apprendre à se comporter de façon vertueuse, vaincre ses passions, voilà qui nécessite un entraînement, des exercices spirituels continus.

C'est ce qui fait le succès de l'approche connexionniste en intelligence artificielle : l'apprentissage en lui-même est un processus relativement opaque, où l'activation des neurones formels selon des patrons statistiques ne laisse en rien préjuger du résultat qui sera finalement obtenu. De fait, le parallèle entre l'approche connexionniste et les éthiques de la vertu se fonde sur l'importance accordée à l'entraînement, à l'exercice[139]. Un auteur comme Peter Sloterdijk tirera d'ailleurs toutes les conclusions de cette observation en faisant de l'exercice une des figures primordiales de la condition humaine.

Une dernière source de questions concerne les limites de l'apprentissage. Relativement à l'apprentissage de valeurs éthiques, se pose le problème de savoir s'il n'est pas plus efficace, plus

---

[136] M. ANDERSON et S. L. ANDERSON, *Case-Supported Principle-Based Behavior Paradigm*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, p. 159.
[137] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 193.
[138] Si l'apprentissage est ici compris comme un processus primitif qui résulte d'un jeu d'interactions avec l'environnement, il n'y a toutefois *aucune nécessité* à ce que partenaires impliqués dans l'apprentissage soient *de même nature* (J. SALLANTIN, *Ce que peut apprendre une machine*, dans Fr. TINLAND, *Ordre biologique ordre technologique*, pp. 197-198).
[139] J. GIPS, *Towards the Ethical Robot*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 250.

économe, de charger d'emblée les valeurs pertinentes, plutôt que de laisser à la machine le soin de les découvrir par essai et erreur[140]. Plus généralement, la question des rapports entre complexité et éthique est posée : un robot capable de gérer la complexité du réel peut-il avoir une moralité « parfaite » ? Y a-t-il incompatibilité ? Plus il peut traiter des situations inattendues, plus son comportement pourrait devenir étonnant, exactement à la façon de la prise de décision humaine[141]. Une telle incompatibilité pourrait nous éclairer sur l'idéalité de la réflexion éthique.

| Questions |
| --- |
| **Une machine peut-elle apprendre à devenir meilleure ? Y a-t-il un apprentissage éthique spécifique ?** |
| **Comment apprenons-nous les règles éthiques ? Quelles règles apprenons-nous au juste ?** |
| **Les apprentissages pour le jugement et le comportement éthiques sont-ils de nature différente ?** |

#### *1.7.2.11. Souffrance*

La question de la souffrance a été mise à l'agenda éthique par les mouvements des droits des animaux. Ils se réclament souvent de Bentham[142], qui a affirmé que la capacité de souffrance prime lors des considérations morales, non les facultés de raisonnement ou de langage. L'importance de ce changement de perspective a été soulignée par Derrida, qui y a vu une remise en question de toute la tradition philosophique occidentale. Toutefois, donner une caractérisation adéquate de la souffrance est un exercice délicat. Gunkel n'a pas tort de souligner que la souffrance est un concept protéiforme, recouvrant un large éventail de réalités, allant de la maladie à la souffrance psychique et émotionnelle[143]. Selon le même auteur, il n'est pas évident que même une réalité aussi simple que la douleur physique se laisse réduire à un cas d'excitation nerveuse (*just adverse nerve stimulus*). Il convient donc de s'entendre.

Au plus bas niveau[144], le niveau physique, nous trouvons ladite excitation nerveuse. Dans le vocabulaire des neuroscientifiques, il est question de « nociception » : si nous faisons l'expérience d'une douleur (par exemple une source de chaleur excessive) un mécanisme tout à fait automatique se déclenche qui nous fait retirer notre main le plus prestement possible. Le dispositif est facilement traduisible en termes fonctionnels comme un mécanisme de conditionnement un peu à la façon de Pavlov. C'est ainsi qu'il est possible de faire adopter à un chien robotique des comportements de plaisir ou de tristesse, en comptabilisant le nombre de messages de succès et d'échec reçus[145]. À un

[140] C'est la distinction que fait N. Bostrom entre *value-loading* et *value learning* (cf. *Superintelligence*, pp. 226-229).
[141] Cf. W. Wallach et C. Allen, *Moral Machines*, pp. 177-179.
[142] Pour la filiation entre Bentham et l'activisme en faveur des animaux, voir D. J. Gunkel, *The Machine Question*, pp. 111-112.
[143] *Ibid.*, pp. 114-115.
[144] La caractérisation de la douleur à trois niveaux provient de G. Chapouthier et Fr. Kaplan, *L'homme, l'animal et la machine*, pp. 62-68.
[145] D. J. Gunkel, *The Machine Question*, p. 135.

niveau supérieur, celui de l'émotion, nous pouvons à proprement parler de « douleur ». Enfin, nous pouvons véritablement parler de souffrance lorsque la douleur reçoit un traitement cognitif, ou conscient : c'est le moment où l'expérience douloureuse peut accéder à notre mémoire épisodique afin de faire partie de notre vécu, de devenir partie de nous-mêmes en quelque sorte.

Quelques remarques s'imposent par rapport à cette caractérisation de la douleur que nous venons d'esquisser. Tout d'abord, alors que les niveaux physique et cognitif sont fonctionnellement bien qualifiables et universalisables à tout système agent, le niveau émotionnel en revanche semble plus récalcitrant à une explication fonctionnelle.

La composante émotive de la souffrance, en effet, ne se laisse que partiellement expliquer en termes fonctionnels. Il n'est même pas sûr que l'émotion puisse être décrite adéquatement comme un comportement : selon Jean-Michel Salanskis[146], l'émotion (ou l'affect) se vit sur un mode passif – on éprouve, on *subit* l'émotion – qui rend problématique une lecture en termes de comportement ou d'action. Nous pouvons aussi penser, avec Sartre, que l'affect s'interprète de deux manières[147]. La première est une réaction de type magique qui transforme le monde de manière à le simplifier, rayer la nécessité d'un comportement trop difficile à être tenu pour le sujet ému : c'est ainsi qu'une femme pleure parce que l'aveu à faire est trop pénible pour être dit. Comprise en ce sens, l'émotivité est une possibilité d'action, jamais réalisée, du sujet. Elle aurait alors une signification tout en étant dépourvue de fonction. Toujours selon Sartre, une deuxième manière de vivre l'émotion, c'est sur un mode irrationnel de conscience passivée : le monde se révèle magique là où nous le croyions déterminé. Et Sartre de citer à l'appui la réaction de frayeur que nous éprouvons lorsque, contre toute attente, un visage hostile apparaît soudainement à la fenêtre : l'être hostile est encore loin, mais notre peur abolit la distance qui le sépare de nous. Le sens de ce visage réside dans notre conscience, et l'émotion est alors comprise comme une synthèse irrationnelle de passivité et de spontanéité. Une fois de plus, l'émotion aurait un sens, non une fonction.

On le voit, l'affect est une question difficile. En outre, il paraît plus difficile à généraliser en dehors des individus biologiques, au point de faire considérer à certains auteurs qu'il justifie la création d'une catégorie distincte d'entités morales : le « patient » moral. Il n'est pas extrêmement clair, toutefois, en quoi un tel ensemble d'entités s'oppose à l'ensemble des agents, car selon la théorie éthique envisagée, différentes relations entre agent et patient peuvent prévaloir. Ainsi, selon une tradition que Luciano Floridi fait remonter à Kant[148], la classe des patients coïncide avec celle des agents. Des écologistes radicaux, en revanche, confèrent un statut de patience jusqu'aux pierres et aux minéraux. Pour eux, les agents forment un sous-ensemble des patients. De ce fait, la patience morale peut simplement être vue comme une question de rôles[149]. Ainsi un même individu peut être considéré sous un aspect de producteur moral ou de récepteur moral (*recipient*). Un producteur moral serait le

---

[146] Voir sa discussion du pâtir dans J.-M. SALANSKIS, *Le monde du computationnel*, pp. 176-184.

[147] Voir J.-P. SARTRE, *Esquisse d'une théorie des émotions*.

[148] L. FLORIDI, *On the Morality of Artificial Agents*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 184-185. Ricœur, quant à lui, la fait remonter davantage encore et la rattache à Descartes. Dans ses vues, la réversibilité des rôles agent et patient est au cœur de la réciprocité et de la pluralité, véritables fondements de sa conception de l'éthique (*Soi-même comme un autre*, p. 382).

[149] S. TORRANCE, *Machine Ethics and the Idea of a More-Than-Human Moral World*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 117 et suivantes.

« responsable éthique » d'un effet moral sur le récepteur, à ne pas confondre avec la cause de l'effet, qui peut être d'un tout autre ordre. Ainsi, la cause d'un incendie peut être une unité électrique défaillante, tandis que le producteur à blâmer est par exemple un électricien négligent. Une telle terminologie n'est certes pas indiscutable, mais elle a le mérite de rendre compte de la distinction entre agents et patients sans multiplier inutilement les classes d'entités morales[150].

En l'absence de définition fonctionnaliste complète, se pose également le problème de la *vraie* souffrance : qu'est-ce qui distingue une réaction d'authentique souffrance d'une suite de simagrées ? Il faut qu'un agent se laisse toucher pour que la souffrance prenne sens en tant que telle ; le mouvement semble profondément intersubjectif. Pour que tel geste prenne un sens moral, il faut que, dans une situation particulière, agent et patient se mettent pour ainsi dire d'accord. Une fois encore, agent et patient ne sont pas à comprendre ici dans un sens étroitement individuel mais plutôt générique, ou collectif : de même que nous ne décidons pas souverainement en tant que « sujet » à qui nous accordons la patience morale, de même le patient individuel n'est pas libre dans son choix d'agents moraux dans le blâme qu'il discerne, sous peine de non-sens, ou de courir le risque d'être taxé de déviant.

Finalement, mettre l'accent sur la souffrance nous mettrait en position délicate vis-à-vis de doctrines éthiques selon lesquelles des entités autres que biologiques peuvent être prises en compte dans les considérations morales. Il n'en demeure pas moins que le critère de la souffrance, pour qui sait différer son jugement, soulève une série de questions très intéressantes.

---

[150] Ce n'est qu'après la rédaction de ce chapitre que nous avons eu connaissance du petit ouvrage de Paul Ricœur, où une distinction rigoureuse est faite entre la lamentation de celui qui souffre (« pourquoi donc y a-t-il souffrance ? »), la plainte (« pourquoi est-ce moi qui souffre ? »), puis l'imputation de responsabilité ou l'accusation (« je souffre à cause de toi ! ») : « Prise […] dans la rigueur de son sens, la souffrance se distingue du péché par des traits contraires. À l'imputation qui centre le mal moral sur un agent responsable, la souffrance souligne son caractère essentiellement subi : nous ne la faisons pas arriver ; elle nous affecte. De là, la surprenante variété de ses causes : adversité de la nature physique, maladies et infirmités du corps et de l'esprit, affliction produite par la mort d'êtres chers, perspective effrayante de la mortalité propre, sentiment d'indignité personnelle, etc. ; à l'opposé de l'accusation qui dénonce une déviance morale, la souffrance se caractérise comme pur contraire du plaisir, comme non-plaisir, c'est-à-dire comme diminution de notre intégrité physique, psychique, spirituelle. Au blâme, enfin et surtout, la souffrance oppose la lamentation ; car si la faute fait l'homme coupable, la souffrance le fait victime : ce que clame la lamentation. » (P. RICŒUR, *Le mal*, p. 23) Ricœur montre notamment comment un travail de deuil peut être mené non pas en redressant les torts (où le patient changerait de rôle, deviendrait agent), mais en se dépouillant de toute velléité d'accusation (d'autrui), de plainte (en cherchant la cause de la souffrance en soi-même), au point de renoncer à l'idée même de lamentation : « L'horizon vers lequel se dirige cette sagesse [de la théologie de la Croix] me paraît être un renoncement aux désirs mêmes dont la blessure engendre la plainte : renoncement d'abord au désir d'être récompensé pour ses vertus, renoncement au désir d'être épargné par la souffrance, renoncement à la composante infantile du désir d'immortalité, qui ferait accepter la propre mort […] » (*ibid*., p. 64).

### *1.7.2.12. Émotions*

Ce qui précède pourrait faire croire que l'émotion est non seulement irréductible à l'analyse fonctionnelle, mais ne se laisserait de surcroît pas même éclairer par elle. Évidemment, des interprétations résolument fonctionnelles de l'émotion existent. C'est ainsi que le neurologue Antonio Damasio, dans son succès de librairie *L'erreur de Descartes*, confère à l'émotion une fonction régulatrice importante dans la prise de décision[151], dont la plus importante peut être qualifiée de « fonction de tri rapide »[152]. Et de fait, dans l'intelligence artificielle existe l'espoir de trier entre informations pertinentes et non en se basant sur une émotion artificielle. L'émotion, alors, crée un cadre à l'intérieur duquel la rationalité peut s'exercer. C'est le thème bien connu de la rationalité limitée[153]. Le thème a des implications éthiques évidentes. Ainsi, contrairement à une tradition qui remonte à la Stoa, Aristote défendait l'idée que les émotions peuvent jouer un rôle significatif lorsqu'il s'agit de déterminer quelles actions sont vertueuses.

À l'inverse, Aristote attendait également, de la part d'un homme vertueux, une retenue émotive. C'est concéder que l'émotion excessive peut avoir des effets néfastes, tout en lui reconnaissant en définitive un rôle positif. Si l'émotion crée l'espace dans lequel la raison peut se déployer, l'aveuglement causé par l'émotion n'est qu'un dysfonctionnement périphérique, dont une machine pourrait s'affranchir. C'est ce que veut éprouver l'architecture cognitive ALEC[154], doté d'un processus de décision à deux niveaux : d'une part, le processus émotionnel est très étendu et prend la plupart des décisions en toute autonomie ; d'autre part, le processus cognitif exerce une influence correctrice sur la décision émotionnelle, mais uniquement en cas de besoin. Le processus cognitif est nécessairement beaucoup plus limité que le processus émotionnel, car le type de calculs qu'il implique seraient beaucoup trop lourds à mettre en œuvre de façon continue.

---

[151] Selon C. MISSELHORN (*Grundfragen der Maschinenethik*, p. 43), qui les a apparemment recensées, Damasio trouve non moins de 12 fonctions à l'émotion. Pour notre part, nous nous contenterons de faire état de la « fonction » qui consiste à *marquer le corps* (A. DAMASIO, *L'erreur de Descartes*, les chapitres 8 et 9) : dans une situation de prise de décision, notre psychisme rejoue en quelque sorte les émotions liées aux conséquences des actions envisageables. Même si ces émotions, dites *secondaires*, sont détachées de leur contexte d'origine, elles provoquent les mêmes réactions corporelles que les émotions dites *primaires*. Ces réactions, telles que le ventre qui se noue, ou une transpiration accrue, Damasio les désigne par le terme de *marqueurs somatiques*, où il faut bien comprendre que le dynamisme des marqueurs somatiques est bidirectionnel. En effet, dans un premier temps, le cerveau, se projetant les conséquences possibles d'une action, libère des neurotransmetteurs qui vont donner lieu aux marqueurs somatiques ; dans un deuxième temps, le marquage va inhiber certaines pistes si celles-ci sont associées à un contenu émotionnel trop désagréable ou menaçant. Ainsi, les alternatives trop pénibles sont écartées avant de devenir accessibles à la délibération consciente. Pour utiliser un vocabulaire qui n'est pas celui d'un neurologue, nous pouvons dire que les marqueurs somatiques viennent élaguer, de façon computationnellement économe, l'espace de recherche.

[152] M. DOMINICY, *L'épidictique et la théorie de la décision*, dans ID. et M. FRÉDÉRIC, *La mise en scène des valeurs*, pp. 76 et suivantes. Cet auteur établit un parallèle intéressant entre les patients avec lésions cérébrales suivis par Damasio et le thème de *l'acrasie*, connue d'Aristote : pour faire bref, disons que l'acrasie aristotélicienne est l'impossibilité dans laquelle se trouve un sujet d'arrêter une délibération pour prendre une décision et de passer à l'action. L'arrêt de la délibération doit être provoqué par un acte de la volonté du sujet acratique, ce dont celui-ci se montre malheureusement bien incapable.

[153] W. WALLACH et C. ALLEN, *Moral Machines*, pp. 143 et suivantes.

[154] *Ibid.*, pp. 160-161.

L'émotion nous éclaire non seulement sur l'action à entreprendre, mais aussi sur la situation à juger, puisqu'elle met en relief ce à quoi nous accordons de la valeur [155]. L'émotion peut alors créer une disposition à l'action, ayant une composante motivationnelle forte. L'émotion est ainsi doublement liée à la prise de décision. La question concernant le statut des émotions dans le monde moral est donc posée. Il n'est pas rare d'entendre dire qu'une machine pourrait prendre des meilleures décisions éthiques car dépourvues d'émotions, amour-propre, intérêts personnels, etc. Or il n'est pas impossible qu'elle en soit, en vérité, un ingrédient indispensable.

| Questions |
| --- |
| **Le rôle de l'émotion ( et autres heuristiques ascendantes) est-il de procéder à un élagage « implicite » de l'espace de recherche des solutions éthiques ?** |
| **Si élagage il y a, comment cela entre-t-il dans la justification donnée ? Comment éviter que celle-ci devienne une ratiocination, une rationalisation *post-hoc* ?** |
| **Quelles alternatives sont considérées à la réaction émotive ? Comment celles-ci sont-elles collectées ?** |
| **Il pourrait, à ce titre, également être intéressant de se demander dans quelle mesure l'émotion intervient dans le passage à l'acte : tant dans l'action que dans l'inaction, il peut y avoir des retombées éthiques fâcheuses. En situation d'incertitude, vaut-il mieux s'abstenir ? L'émotion joue-t-elle un rôle ?** |
| **Enfin, une méthode formelle telle qu'elle est à l'œuvre dans un système multi-agents est-elle seulement *capable* de modéliser des émotions ? Comment s'y prendrait-elle ?** |

#### *1.7.2.13. Résumons…*

S'il nous est loisible de faire un premier point, nous dirions que les critères d'agentivité, dans leur ensemble, présentent une tendance individuelle et psychologisante très marquée : ils analysent le comportement éthique comme une série d'aptitudes cognitives[156] plutôt que comme réalité sociale. Or, même si certaines aptitudes cognitives sont, sans nul doute, un substrat nécessaire au vivre-ensemble éthique, du moins en nous limitant aux êtres biologiques, elles n'en constituent pas les composantes définitoires propres, de même qu'en linguistique, la langue ne se résume pas aux facultés phonologiques, sémantiques et syntaxiques d'un individu quelconque, mais ne peut ultimement être dite opérer qu'en situation, dans un contexte social.

[155] C. MISSELHORN, *Grundfragen der Maschinenethik*, pp. 41-44.

[156] Le psychologisme est par exemple très marqué chez N. BOSTROM, lorsqu'il attribue à une « superintelligence » (le terme n'étant jamais vraiment conceptualisé) la faculté de développer un sens moral tout à fait hors pair. Or dans son propre ouvrage, il fait référence à l'intelligence collective, c'est-à-dire la spécialisation du travail qui rend possible le monde moderne. Cette spécialisation présuppose des structures et des institutions, industrie, éducation… C'est dire qu'un ensemble dense de relations interobjectives permet la spécialisation, bien davantage que le « contenu » de notre boîte crânienne.

Lorsque nous nous intéressons explicitement au contexte social d'une interaction éthique, la question devient donc la suivante : un statut moral peut-il être décidé sur la base de qualités individuelles[157] ? La source principale du comportement éthique est-elle à chercher dans « l'individu » ou dans la collectivité ? Plusieurs auteurs se sont intéressés à cette question. Là comme ailleurs, les avis divergent sur le statut exact à accorder à ce contexte social. Une vue – extrême – est de considérer que le contexte prime à tel point sur les individus que l'agentivité s'analyse mieux en termes de construction sociale : l'agentivité est ainsi attribuée par un locuteur en vue d'une efficacité sociale. Attribuer ainsi une agentivité devient, par là même, une action éthique. Il s'ensuit que l'agentivité devient une question secondaire : la relation sociale précède et prescrit ce que nous sommes. En effet, comme la qualité de personne est négociée et construite socialement, la personne morale devient un idéal normatif et donc un *explicandum* de l'analyse éthique, plutôt qu'un *explicans*. Sans aller aussi loin, il est indéniable que la moralité naît d'interactions sociales, d'où d'ailleurs toute la pertinence de considérer des systèmes multi-agents, capables de prendre en compte la dynamique sociale[158].

De fait, notre examen de l'agentivité a bien montré que cette notion n'est pas *nécessairement* liée à la possession intrinsèque de certaines qualités individuelles. Non seulement le statut même d'agent moral peut être attribué en contexte, mais d'autres qualités constitutives – pensons à l'autonomie – ne peuvent être évaluées qu'en fonction de l'action considérée : nous pouvons être très autonome pour nouer nos lacets ou préparer à manger, nous ne serons cependant pas pour autant plus fiers lorsqu'il s'agira de sauver des vies. Reprenons encore l'exemple de l'unité d'agir : ce n'est que dans un contexte social bien précis que nous accordons un degré d'unité à la chambre chinoise en tant que telle[159].

À ce stade, il faut peut-être préciser un point : soutenir la thèse selon laquelle les attributions morales ne proviennent pas de facultés intrinsèques de certains individus ne préjuge en rien l'existence de ces facultés. La thèse revient simplement à s'interroger sur leur pertinence sur le plan de l'analyse éthique. En effet, si l'individu doit être comparé à un rouage qui produit un certain effet dans une machinerie, il est bien sûr loisible à tout un chacun d'étudier la forme et la matière du rouage. Or, il paraît évident que ni la forme, ni la matière du rouage individuel ne produit le moindre mouvement. Ce qui est premier ici, c'est *l'emboîtement* des rouages, conçus pour fonctionner ensemble : nous pouvons changer à souhait la forme des roues – remplacer la denture droite par une denture hélicoïdale par exemple – tout en conservant un engrenage équivalent, pourvu bien sûr que nous changions tous les rouages au même moment, que nous respections la structure et la logique du système en place.

S'il y a une chose pour laquelle un consensus entre auteurs fonctionnalistes se dessine, c'est bien sur la question de savoir quels seraient les critères *minimaux* afin qu'une unité puisse être considérée comme agent. Afin de faire preuve d'un sens minimal d'éthique dans sa prise de décision, il doit tout d'abord faire le « bon » choix, c'est-à-dire un choix éthiquement souhaitable dans un contexte

[157] Pour un survol rapide des thèses constructivistes, voir D. J. GUNKEL, *The Machine Question*, pp. 163-175.
[158] W. WALLACH et C. ALLEN, *Moral Machines*, p. 133.
[159] Rappelons-nous à ce propos la thèse d'Isabelle STENGERS dans les Cosmopolitiques : « ce qui compte » ne peut être défini que lors d'une délibération, sinon nous risquons toujours d'écraser autrui et ce qui compte pour lui.

donné. Cette première condition tombe sous le sens, vu le primat accordé au comportement, vu le sens commun aussi, mais elle ne suffit pas. Il est, en outre, essentiel que l'agent puisse *justifier* ce choix en fonction de considérations (principes, valeurs…) éthiques. Ce sens minimal *inclut* la possibilité de poser une action moralement correcte mais *exclut* le *vécu* d'avoir, *d'éprouver*, un problème éthique[160]. Dans le prochain chapitre, nous aurons amplement l'occasion de revenir sur une question qui, jusqu'ici, est restée dans l'ombre : quelles *valeurs* un agent artificiel pourrait-il bien poursuivre ?

Une deuxième critique du psychologisme est plus fondamentale : il privilégie la vue d'une machine comme un robot anthropomorphe, alors qu'il est loin d'être évident que la plupart des machines intelligentes qui nous entoureront à l'avenir auront la forme canonique d'un corps à quatre membres et une tête. Au contraire, il y a tout lieu de croire que la plupart des agents artificiels à venir, intelligents ou non, resteront comme aujourd'hui au stade d'applications à l'individualité mal définie. Leurs implications éthiques n'en seront pas moindres pour autant, ni leur faculté de prendre des décisions autonomes, d'interagir avec leur environnement ou d'inculquer un sens au monde dans lequel elles évolueront.

## 1.8. La valeur fonctionnelle

L'éthique des machines est une discipline relativement nouvelle, nous l'avons vu. Il est cependant remarquable qu'une question centrale de l'éthique reste étrangement périphérique à ce champ de recherche, nous voulons dire celle de la valeur. Précisons cette affirmation. Du côté des études descriptives, la valeur est prise pour une donnée brute : ainsi dans l'étude que nous avons déjà citée sur MoralDM, où la présence de valeurs protégées diminue la sensibilité des participants aux retombées quantifiables de leur (in)action[161]. Les auteurs se contentent de constater la présence de la valeur, un peu à la manière d'un facteur statistique qu'il s'agirait d'extraire de la gangue des observations encore informes. Ce point de vue peut se montrer fécond dans certains types d'études d'éthique appliquée. Cependant, pour d'autres types de questions que nous verrons dans ce chapitre, il s'avère vite limité.

Si nous tournons le regard vers les études d'inspiration prescriptive, la « bonne » valeur est présentée comme le fruit du consensus entre spécialistes de l'éthique (*trained ethicists*). Nulle interrogation sur la façon dont ces éthiciens parviennent à ce consensus ; là encore, la valeur est donnée. En d'autres termes, et en ne forçant guère le trait, la valeur y devient un optimum à maximiser aux mains d'une poignée de technocrates. Si ceux-ci ne sont pas (encore) parvenus à un consensus, les auteurs continuent, point d'approche fonctionnelle possible, ni même souhaitable.

[160] Dr. McDermott, *What Matters to a Machine?*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, pp. 111-112.
[161] M. Dehghani, K. Forbus, E. Tomai et M. Klenk, *An Integrated Reasoning Approach to Moral Decision Making*, dans *op. cit.*, pp. 422-441.

Ce point de vue résonne étrangement avec le scientisme dénoncé par Isabelle Stengers, qui se fonde sur la violence interne à un savoir qui se veut « enfin » objectif :

> *Tant que nous serons hantés par le modèle idéal d'un savoir rationnel, objectif, susceptible de mettre d'accord tous les peuples de la terre, que ce soit pour le promouvoir ou pour le déconstruire, nous resterons incapables de nouer avec ces autres peuples des rapports dignes de ce nom.*[162]

Selon la philosophe, une question d'intérêt public ne peut être livrée sans reste au jugement d'une discipline scientifique, quelle qu'elle soit, car souvent la question posée au préalable contient déjà sa propre réponse. Pour peu qu'on ouvre le débat, la question sera reprise, retravaillée, transformée, afin de prendre en compte les intérêts du plus grand nombre, et non plus seulement de ceux qui se trouvaient à l'origine de la question initiale. La question de savoir ce qui compte, ce qui a de la valeur est donc essentiellement ouverte :

> *On peut donc rêver d'une autre histoire, où les scientifiques auraient cultivé ce qui fait leur spécificité, c'est-à-dire où ils auraient pu se présenter de manière civilisée à d'autres, cultivant d'autres spécificités. On peut rêver d'une histoire où ce qui aurait essaimé n'aurait pas été l'autorité des faits mais le caractère exigeant de ce qui signifie pertinence. Si la pertinence – l'engagement à créer des situations qui donnent à ce à quoi un scientifique s'adresse le pouvoir de faire une différence cruciale en ce qui concerne la valeur de ses questions – avait été le trait commun des sciences, le nom du jeu aurait été aventure et non conquête.*[163]

Cette « civilisation » du travail scientifique vaut autant pour le travail politique :

> […] *que nul ne puisse être autorisé à définir « ce qui importe vraiment* ». *Cet interdit n'est pas moral mais condition d'une culture de la symbiose, d'une culture de la capacité de chaque protagoniste à se présenter avec ce qui* lui *importe et à savoir que ce qu'il apprendra de l'autre devra toujours être compris comme réponses aux questions qui,* pour lui*, importent. Questions dont la valeur tient certes à la pertinence, condition pour que la réponse ne soit pas extorquée, mais c'est précisément la pertinence qui interdit le rêve de l'extraction de ce qui est « vraiment important ». On ne s'empare pas de ce dont on dépend.*[164]

De cette (trop) brève discussion, retenons qu'une position technocratique en matière éthique est difficile à tenir : un consensus entre éthiciens ne vaut pas valeur. Un examen plus approfondi est donc nécessaire.

---

[162] I. STENGERS, *Une autres science est possible !*, p. 124.
[163] *Ibid.*, p.127.
[164] *Ibid.*, p. 80.

### 1.8.1. Un système de valeurs

Nous ne pouvons pas faire l'économie d'une discussion sur les conditions et l'efficace d'un système de valeurs. S'il nous est permis de faire un pas en arrière en nous en tenant un instant au cas d'un être biologique, celui-ci va, dans son comportement, éviter les déplaisirs et chercher les plaisirs. Les plaisirs et les déplaisirs dépendent du type d'organisme : ainsi beaucoup d'organismes craignent la faim, la soif, la solitude, etc. Cet ensemble de plaisirs et de déplaisirs n'est dans le fond rien d'autre qu'un système de *valeurs*, certes assez fruste, mais déjà capable de guider le comportement de l'agent biologique. Des valeurs les plus élémentaires aux plus complexes, tous les raffinements sont envisageables. Ainsi dans le cas de l'être humain individuel, celui-ci supporte mal des dégradations de son image de soi, ou, pour parler en termes aristotéliciens, de son *ethos*. Inversement, les gestes les plus désintéressés se laissent encore interpréter comme une façon qu'a l'homme de valoriser sa propre image, ne fût-ce qu'à ses propres yeux.

La question de l'éthos ne doit pas être confondue avec celle de la *survie* : la lutte pour la survie est certes un thème bien connu en biologie, mais ce n'est pas ici notre problème. Comme l'a si bien résumé Michel Dubois, chaque être persévère dans son être… au point même d'en mourir[165] ! Ne survivent en réalité, dès que les conditions de vie deviennent plus difficiles, que les opportunistes, ceux qui ont sacrifié une part de leur « essence », qui ont renoncé à une part d'eux-mêmes afin d'assurer leur présence physique continuée. Cette « essence », Ricœur la qualifie *d'ipséité*[166] : nous tendons à conserver dans le temps nos propres traits de personnalité ainsi que nos habitudes (notre *caractère*), nous honorons nos promesses aussi, nous sommes fidèles à la parole donnée, justement parce que nous assurons ainsi un certain maintien de soi. Mais ce maintien de soi, ce n'est pas précisément la perpétuation du même ; il arrive d'ailleurs à Ricœur[167] de le rapprocher du *conatus* spinoziste : le conatus, chez Spinoza, désigne la persévérance de l'être à être, l'effort de conserver, voire d'augmenter, sa puissance d'être à l'étant. Cet effort est son essence – si toutefois le terme d'essence peut avoir un sens pour une conception essentiellement dynamique de l'être. L'implication éthique du conatus[168] est la vertu de l'affirmation de soi, de la puissance vitale de l'être victorieuse de la mort, de l'ignominie et de la peur. En deux mots : « la vie veut vivre »[169].

De telles lectures « autotéliques » – où la conservation *de soi* au sens large est à l'origine du comportement éthique – ne sont pas réservées à l'individu biologique, mais peuvent être appliquées à des collectivités. Nous en voulons pour preuve le débat écologique : y a-t-il un poids éthique parce que « Gaia souffre », ou parce que nous renvoyons les conséquences de nos actions « à nos enfants » ? À coups d'images – métaphores et métonymies – nous donnons corps à une appartenance qui transcende nos propres limites dans l'espace comme dans le temps. Les êtres biologiques individuels ne sont donc pas nécessairement le bon niveau d'analyse éthique ; car ce

---

[165] M. DUBOIS, *La métaphore et l'improbable*, p. 46. L'auteur désigne cette observation par « la métaphore d'Icare ».
[166] P. RICŒUR, *Soi-même comme un autre*, pp. 140-150, 193-198. Voir aussi nos sections consacrées à l'identité (§ 1.7.2.2) et à la responsabilité (§ 1.7.2.6).
[167] *Ibid*., p. 365.
[168] A. COMTE-SPONVILLE, *Petit traité des grandes vertus*, p. 148.
[169] Ce sont les paroles d'une professeure de biologie : alors que son discours scientiste sert d'ordinaire des propos misanthropes, il lui arrive par moments aussi d'accéder à des perles d'une touchante lucidité (J. SCHALANSKY, *Der Hals der Giraffe*, p. 129).

pour quoi l'écologiste se bat n'est pas la conservation de son corps individuel, il se bat pour une certaine image de l'homme qu'il voudrait voir vivre dans un certain rapport d'équilibre avec les créatures environnantes, alors qu'il a des raisons de le voir – *se* voir – plutôt en parasite… ou pire encore, en prédateur vorace.

Dans le cas de l'homme, la provenance de ces valeurs peut être organique, apprise socialement… les sources sont diverses. Nous retrouvons là un débat très vaste. Qu'en est-il cependant dans le cas des robots ? L'idée de les doter d'un sens de conservation de soi a déjà été explorée dans la pièce de théâtre *R.U.R* de 1920, qui vit la naissance du terme « robot »[170] : pour des raisons purement industrielles, des robots y sont pourvus de « sentiments de douleur » afin de fournir une garantie de longévité à leur fabricant. Cette valeur est donc tout à fait extérieure, surimposée par son concepteur humain. De tels mécanismes ont entretemps été effectivement implémentés[171] : il est possible d'inculquer à un robot quels états sont désirables, lesquels ne le sont pas. Muni d'un tel système de valeurs et d'une capacité de prédiction, le robot peut alors adapter ses stratégies pour réaliser certaines actions, sans qu'il y ait besoin de programmer explicitement les comportements que le robot doit tenir.

Cependant, de telles valeurs peuvent également être intrinsèques[172] : à partir de certains « méta-principes » très généraux, tels que le « désir » de maximiser les occasions d'apprentissage ou encore le principe de réversibilité (« ne fais pas ce que tu ne peux pas défaire »), un robot peut être amené par lui-même à découvrir la nécessité de se conserver. En effet, une action qui l'endommage ne peut être défaite. Il apprendra par exemple de façon tout à fait autonome d'éviter les collisions. Ainsi, afin d'assurer sa propre intégrité physique et son autonomie de fonctionnement, le principe abstrait peut devenir un moteur d'actions « bonnes », sans aucune programmation spécifique *ad hoc*. Les valeurs se trouvent ainsi donc au cœur du comportement, sans qu'il y ait besoin d'engagement métaphysique appuyé. Et même dans le cas des machines, une valeur ne doit pas être d'origine individuelle, mais peut aussi émerger grâce à une interaction sociale, et être passée de génération en génération.

Nous devons cependant émettre une réserve importante par rapport à la conversation de soi appliquée à des agents robotiques ; la réserve que nous allons faire ici s'inspire d'une remarque qu'a faite Ronald Arkin[173] sur l'utilisation de drones tueurs sur le champ de bataille. Dans de telles conditions, l'intérêt éthique d'une telle utilisation réside dans le prodigieux quant-à-soi du drone : il ne perd jamais son sang-froid, il n'agit jamais par frayeur. S'il n'est pas en mesure de distinguer un civil d'un militaire, un ennemi d'un allié, il s'abstient du recours à la force. Ce n'est vraisemblable, bien sûr, que quand le drone ne valorise pas outre-mesure son propre maintien. Ceci a une conséquence décisive pour notre problématique : dans certains cas, un robot ne peut être utilisé éthiquement qu'à condition de *ne pas ressembler à l'homme* ; en l'occurrence, qu'il soit dépourvu de cet instinct de conservation dont l'homme a un sens tellement aigu. Nous nous attendons – nous devons nous y attendre – à ce qu'il renonce à son être propre au profit d'autrui, ce qui explique aussi

[170] D. J. GUNKEL, *The Machine Question*, p. 134.
[171] G. CHAPOUTHIER et Fr. KAPLAN, *L'homme, l'animal et la machine*, pp. 31-33, 56.
[172] *Ibid.*, pp. 94-95.
[173] Cf. R. ARKIN, *Governing Lethal Behavior in Autonomous Robots*, pp. 29, 45-48.

pourquoi la casse est particulièrement importante parmi les drones. L'auteur fait d'ailleurs état de deux autres endroits où la *dissemblance* entre homme et robot est résolument salutaire : d'une part, l'absence chez le robot d'une faculté d'intentionnalité d'ordre supérieur le rend inapte au mensonge ; d'autre part, la dissemblance physique : dans la plupart des situations de la vie réelle, un robot à huit pattes ou sur chenilles sera autrement plus stable, aura des mouvements bien plus sûrs, qu'un robot humanoïde à deux jambes, qui est une invention peut-être condamnée à ne jamais dépasser le stade de gadget.

Cette réserve étant faite, tournons-nous vers les aspects fonctionnels de la valeur. D'un point de vue fonctionnel, son intérêt est double : la fonctionnalité d'un système de valeurs a partie liée avec son rôle dans la prise de décision d'une part et a une composante motivationnelle forte, d'autre part : le système de valeurs – tout comme l'émotion – a une importance dans notre disposition à l'action. Nous avons déjà vu plus haut (§ 1.4.1) qu'à utilité égale, l'être humain préfère la passivité à l'action. L'être humain doit être motivé afin de passer à l'action, il doit estimer que son action *en vaille la peine*. Un système de valeurs ne doit pas être vu comme un ensemble d'interdictions ou de contraintes, mais comme une condition *sine qua non*, une structuration, non seulement de notre réalité sociale et de notre vivre ensemble, mais même comme le socle de notre rationalité pratique : notre système de valeurs ne limite pas, mais rend possible le débat politique et éthique. En effet, la valeur permet l'argumentation, elle peut rendre problématique un penchant qui sans elle ne relèverait que du goût de chacun :

> *On peut discuter de l'argent, de son usage social, de son affectation au sein de la société, mais cela n'a pas de sens de critiquer quelqu'un dont toute la vie tourne autour de l'argent. Si on le fait, c'est sur la base de valeurs qui dépassent les choix individuels, et qui servent même à en évaluer les conséquences, mais les passions elles-mêmes ne permettent pas de juger les passions.*[174]

Il y a cependant une condition majeure que doit remplir un système de valeurs efficace. Un système de valeurs n'est fonctionnel que s'il est *partagé* : les lois et normes de la cité, à condition d'être bien implantées, deviennent des lieux communs, ou topiques. Dès lors, avoir raison, c'est se ranger avec la collectivité, faire preuve d'*homonoia*, la cohésion sociale devenant ainsi garde-fou de la rationalité[175]. Penser la rationalité comme avant tout collective n'est en rien propre à la Grèce antique : nous retrouvons cette idée – avec des nuances importantes – chez des penseurs venant d'horizons divers, allant de Stengers à Habermas. La théorie des *implicitations* (*implicatures*) de Paul Grice a également mis en relief que la construction de sens dans un dialogue n'est possible que si les interlocuteurs assument une rationalité partagée[176]. En d'autres termes, la rationalité partagée précède les individus.

Cet état de choses nous inspire une autre réflexion, portant sur les rapports entre agents, pris en tant qu'unités d'agir individuelles, et le groupe ou la collectivité. Dans la mesure où les valeurs ne sont

---

[174] L'observation a été faite par M. MEYER, *Principia Rhetorica*, p. 195.
[175] E. DANBLON, *La rationalité du discours épidictique*, dans M. DOMINICY et M. FRÉDÉRIC, *La mise en scène des valeurs*, p. 28.
[176] S. PAYR, *Towards Human-Robot Interaction Ethics*, dans R. TRAPPL, *A Construction Manuel for Robots' Ethical Systems*, p. 48.

pas une affaire de l'individu d'abord, mais du groupe dont il émane, il est pertinent de problématiser le primat implicite que l'éthique des machines accorde à l'individu, qu'il soit humain ou animal, biologique ou robotique, physique ou logiciel. Plus précisément, l'interrogation porte sur le poids éthique du choix individuel : dans quelle mesure le monde qui nous entoure nous impose ses choix, ou plutôt, dans quelle mesure ce monde façonne-t-il la palette de choix qui s'offre à nous ? Dans quelle mesure cette palette de choix est-elle éthique ? Nous comprenons, dès lors, que toute simulation de la dimension éthique ne peut faire l'économie de formaliser rigoureusement le système de valeurs qui l'inspire. En d'autres termes, un système multi-agents ne devra pas se limiter à définir et modéliser les agents, mais aussi le système de valeurs qui les anime.

| Questions |
| --- |
| **Comment les SMA modélisent-ils les valeurs sur lesquelles les agents peuvent fonder leurs choix ? Dans quelle mesure les valeurs informent-elles les choix que les agents peuvent faire ?** |
| **Comment évaluer le poids respectif de l'éventail des choix offert à l'agent et les choix individuels faits par l'agent dans cet éventail ?** |

### 1.8.2. Les valeurs en action : le blâme et la reconnaissance

Le système de valeurs a partie liée avec l'autotélisme, nous l'avons vu. L'autotélisme est parfois attribué au « rôle » du patient en éthique. Pourtant, disposer d'un système de valeurs n'est pas le propre du patient, il guide également l'agent dans ses actions. Formuler un blâme verbalement, dans le langage, c'est encore une action. Laissons donc là l'autotélisme, et reformulons de façon neutre dans un système de valeurs quelconque, potentiellement généralisable à tous les contextes : lorsqu'un agent réalise une action qui a une incidence, négative ou positive, sur les valeurs d'un système interactant, celui-ci peut émettre un blâme ou une reconnaissance[177].

L'interactant patient va émettre un blâme ou une reconnaissance à l'encontre de qui ? Le seul fait de soulever la question revient à reconnaître qu'à son tour, le patient définit qui il accepte comme agent. « Le patient » doit ici être compris génériquement, le patient non comme individu mais comme chaînon social. Illustrons ce point par un exemple : nous pouvons dénier individuellement tel ou tel

[177] En relisant ce paragraphe, nous nous sommes rendu compte qu'il y manque une précision importante quant à l'usage de l'opposition entre blâme et louange. C'est que le blâme et la louange peuvent se prédiquer de deux types d'êtres différents : ils peuvent – premier emploi – s'appliquer aux actions elles-mêmes ; ils cherchent alors à distinguer entre ce qui est *permis* et ce qui ne l'est pas. Visant l'objectivité, le jugement qu'ils contiennent est passible d'être vrai ou faux. Dans un deuxième emploi, le blâme et la louange ne se prédiquent plus des actions, mais des agents qui les posent. Ils ambitionnent dès lors de distinguer entre qui est *coupable* et qui ne l'est pas. Dans ce deuxième emploi, ils ne jugent plus un état du monde (en tant que prédicat de premier ordre), mais l'état d'esprit d'un agent, en tant que prédicat de second ordre et à ce titre, le vrai et le faux des logiciens ne sont plus à invoquer (P. RICŒUR, *Soi-même comme un autre*, pp. 340-341). La précision est importante, mais il nous semble que, dans tout le paragraphe, le deuxième sens est utilisé de bout en bout.

titre à telle ou telle entité, mais ce déni n'est pas forcément performatif, car il peut retomber sur nous qui risquons d'être taxé de sociopathe, de cruel… Ce n'est donc qu'au travers du regard d'un Autre que nous nous constituons en tant qu'agent. Cependant, afin que le blâme ou la marque de reconnaissance soit reconnu, il faut que l'agent ait accepté le système émetteur… comme interactant. Pour le dire autrement, les interactants sont le produit de l'interaction, ils ne lui précèdent pas[178]. Nous pouvons compléter en disant qu'il en est ainsi parce que la nature de l'interaction impose son système de valeurs.

Revenons un instant sur les mécanismes qui peuvent conduire un patient à choisir un type d'agents plutôt qu'un autre. À ce propos, rappelons que le blâme et la reconnaissance sont attribués non (seulement) en vertu d'une capacité de prise de décision, mais aussi (voire surtout) en fonction d'un pouvoir de *réparation*[179]. Or, et quoiqu'une telle vue puisse être précieuse lorsqu'il s'agit de tirer des conséquences légales de considérations éthiques, une telle définition risque de donner lieu à des confusions avec l'ordre juridique, tout en étant vulnérable à des attaques d'inspiration nietzschéenne : tous les beaux discours éthiques ne serviraient en dernière analyse qu'à sublimer des rapports de pouvoir.

Le critère premier devrait donc plutôt s'énoncer ainsi : nous discernons un blâme et accordons notre reconnaissance à celui que nous estimons le plus capable de *répondre*, celui en somme avec qui nous sommes le plus à même d'interagir. Cependant, une réserve s'impose. Les raisonnements qui précèdent présupposent que l'objet du blâme, l'individu blâmé, est également celui à qui le blâme est adressé. Or à prendre les choses sous l'angle de la rhétorique, il appert qu'un discours de blâme prend souvent pour public un auditoire autre que la personne blâmée, comme si, sans témoins tiers, le blâme restait sans effet. Dès lors, en discernant un blâme, il faut tenir compte de la possibilité que celui que nous cherchons à influencer ne soit pas la personne blâmée : nous la donnons en (anti-) exemple à d'autres. Le blâme et la reconnaissance, dans cette optique, sont donc principalement attribués en vertu de leur pouvoir d'*évocation*.

Un deuxième critère peut être trouvé dans la réduction à l'absurde que constitue l'exemple de l'outil. L'usage de l'outil pour augmenter notre capacité physique et/ou intellectuelle est très naturel à l'être humain, voire même, cet usage prétend au titre d'être le propre de l'être humain[180]. L'outil est en soi inerte : sa prise d'initiative est néant. Allant un pas plus loin, un agent technologique tel qu'un robot de chaîne de montage n'est pas inerte, mais son action est rigoureusement planifiée. Une fois activée, il pose toujours les mêmes gestes. Si matière à blâmer il y a, celle-ci est nécessairement transférée sur les agents qui interagissent avec lui, vu sa prévisibilité.

La prévisibilité et son pendant, l'opacité, forment le cœur d'une éthique d'inspiration lévinassienne : l'opacité de l'autre n'est pas une limitation mais la condition même de la relation éthique. L'égo découle de l'Autre, de ce qui me résiste[181]. Hâtons-nous de préciser que l'opacité de l'autre n'est pas une qualité intrinsèque d'autrui, mais ne peut prétendre à la pertinence que sous le regard de l'autre,

---

[178] Comme l'a formulé Donna Haraway (cité dans D. J. GUNKEL, *The Machine Question*, p. 124).
[179] G. CHAPOUTHIER et Fr. KAPLAN, *L'homme, l'animal et la machine*, p. 138.
[180] Y. HUI, *On the Existence of Digital Objects*, p. 147.
[181] D. J. GUNKEL, *The Machine Question*, pp. 176-177.

du point de vue donc de nos interactants. Selon la grille d'analyse que nous venons de voir, où nous opposons responsabilité et prévisibilité, le blâme et la reconnaissance seront attribués au chaînon « opaque » du système considéré. L'exemple du télérobot est à cet égard éclairant : nous aurons tendance à blâmer l'opérateur pour toute action répréhensible du robot, même si sans ce robot l'opérateur n'aurait jamais eu le pouvoir nécessaire pour faire le mal[182]. Nous pourrions blâmer le robot, ou au moins son producteur, mais nous choisissons en premier lieu le sous-système dont nous soupçonnons les intentions troubles, ou la négligence coupable.

Quoi qu'il en soit, le blâme et la reconnaissance, principes premiers en quelque sorte de l'éthique, doivent être rigoureusement distingués d'une notion extrêmement ambiguë ayant pour nom la « responsabilité éthique ». Il s'agit là d'une invitation permanente à toutes les confusions avec la responsabilité juridique[183], tout en formant un cadre trop étroit pour juger des conséquences éthiques d'un acte[184]. Pour s'en convaincre, il suffit de rappeler que dans une perspective utilitariste au moins, la responsabilité morale n'a qu'une pertinence toute secondaire[185]. Il en va d'ailleurs de même dans l'éthique essentiellement tournée vers l'avenir de Hans Jonas, où la responsabilité traditionnelle, tournée vers le passé, passe au second plan. La question du statut de la notion est donc ouverte. Ceci est d'autant plus vrai quand la machine entre en ligne de compte : il semble assez évident que les machines en tant que telles ne se qualifient pas pour la responsabilité morale, mais influencent celle des hommes, notamment pour les systèmes sociotechniques, ceux qui comprennent à la fois une composante technique et une composante humaine (que ce soit à titre de constructeur, d'utilisateur, ou d'exploitant)[186].

Alors que l'utilité de la notion est douteuse, il est en revanche absolument certain qu'elle donne lieu à de fréquents abus dans les discours tant publics que privés : il semble suffire d'évoquer la responsabilité pour trouver l'aubaine de s'en décharger. Ainsi, Misselhorn signale à quel point le petit personnel exécutif qui doit manier les drones militaires est minutieusement encadré[187], alors que ce ne sont manifestement pas eux qui ont décidé de la guerre, ni qui ont libéré du budget pour acheter des drones. Ce sont bien là, évidemment, les décisions morales les plus significatives.

---

[182] Pour l'exemple du télérobot, voir J. P. SULLINS, *When Is a Robot a Moral Agent?*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 152-154.

[183] Les confusions avec l'ordre juridique semblent malheureusement plutôt la règle que l'exception dans les travaux que nous avons pu consulter. Ainsi D. J. GUNKEL déduit-il du fait que l'esclave, sous l'empire romain, était la propriété du *pater familias* qu'il n'avait aucun statut éthique (*The Machine Question*, p. 40). Ou encore, plusieurs auteurs semblent puiser dans la « personnalité juridique » de certaines organisations et sociétés un argument convaincant pour le statut éthique de ces dernières. Il serait intéressant de comparer ce type de confusions à la fascination contemporaine pour le droit et au paradigme de la société de contrôle, où l'État institutionalise l'attribution du blâme et, dans une moindre mesure, celle de la reconnaissance. Dans le vocabulaire ricœurien, le blâme – même de second ordre – ne doit pas être confondu avec *l'incrimination*.

[184] W. WALLACH et C. ALLEN, *Moral Machines*, p. 204.

[185] C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 169.

[186] *Ibid.*, pp. 128-129.

[187] *Ibid.*, p. 189. Nous aurons l'occasion de voir une telle déresponsabilisation à l'œuvre dans le dernier cas pratique du troisième chapitre, consacré aux voitures autonomes.

### 1.8.3 Les valeurs en mutation : la négociation

La valeur ne doit pas être expliquée, elle sert de garantie non décomposable à la discussion éthique. Très souvent, la valeur est ce au-delà de quoi la raison pratique refuse d'aller[188]. Même abstraction faite des effets rhétoriques qui présentent la valeur comme immuable, c'est aussi l'intuition du sens commun : la valeur est tellement prégnante qu'elle semble détachée du temps et du lieu qui l'ont vu naître. La valeur paraît atemporelle. Et pourtant, nul ne saurait ignorer que l'éthique est toujours en mouvement : les normes doivent être établies de façon coopérative, dans un travail toujours à recommencer. C'est ce qui a fait dire à Putnam que l'éthique est une affaire de décision de nature collective – décision toujours provisoire d'ailleurs – et non de découverte [189].

À ce mouvement perpétuel de l'éthique, il est cependant possible de donner deux sens qui, tout en étant interdépendants, ne se recoupent pas. Dans un premier sens, que nous pourrions qualifier de diachronique : les valeurs changent dans le temps. Un deuxième sens serait plutôt l'aspect synchronique : entre individus, quelle valeur appliquer dans telle situation concrète ? Comment la catégoriser ? La vie de tous les jours, en effet, nous montre pléthore d'exemples où la valeur, justement, est au centre des préoccupations. Le cas le plus fréquent tient aux conflits de jugement : tel dira que tel acte est un vol, tel autre déniera cette qualification. À titre d'anecdote exemplaire, il y a quelque temps, nous étions confronté à un consultant en *coaching* éthique. Celui-ci soutenait que l'entreprise était un peu comme la maison d'un ami : donc, prendre des stylos, des classeurs ou du papier à l'entreprise, c'était *comme* voler des fourchettes ou des couteaux de la domesticité de notre ami. Nous avions bien noté, alors, que cette mise en équivalence de deux registres *a priori* distincts n'avait pas vraiment convaincu l'auditoire : en d'autres termes, la négociation que le consultant a proposée est restée lettre morte, et la mise en équivalence – exprimée sur le mode du « comme » – a pu être dénoncée comme n'étant que figure de style[190]. Pour exprimer les choses encore autrement, en empruntant la terminologie de la théorie de Boltanski et Thévenot vue précédemment (§ 1.6), nous pourrions dire que le consultant a proposé une épreuve venant de la cité domestique, donnant lieu de la sorte à un *différend* avec son auditoire. En effet, celui-ci ne l'a pas suivi, lui préférant en l'occurrence une épreuve industrielle, portant sur une organisation efficace du travail.

Autour des transgressions de la valeur – ici, le respect de la propriété privée – nous négocions dans de tels cas ses limites d'applicabilité. Dans des cas plus extrêmes, nous pouvons tellement rétrécir les limites d'applicabilité que la valeur en perd son caractère atemporel : ainsi la virginité des jeunes filles a vu son champ d'application se restreindre à la seule progéniture des classes dominantes[191]. Une fois qu'une telle restriction est opérée, la valeur devient raffinement, avant de sombrer ultimement dans le ridicule. En définitive, les conceptions éthiques ne sont pas statiques, mais en

---

[188] Voir le chapitre *Logique, dialectique, philosophie et rhétorique* de Ch. PERELMAN, *L'empire rhétorique*, pp. 16-25.
[189] D. J. GUNKEL, *The Machine Question*, pp. 213-214. Soulignons que nous retrouvons ici les valeurs en tant que principe constitutif.
[190] C'est l'analyse que fait I. STENGERS de ce qu'elle appelle des « concepts nomades » en sciences lorsque ceux-ci échouent leur opération de capture (cf. *La propagation des concepts*, dans EAD., *D'une science à l'autre*, pp. 14-23).
[191] M. FOUCAULT, *Histoire de la sexualité II*, pp. 265-278.

renégociation constante. Ainsi, il faut prévoir que les machines intelligentes influenceront nos conceptions éthiques. Reste à savoir – et à décrire – comment elles le feront.

| Questions |
| --- |
| **Les agents peuvent-ils négocier entre eux ? Peuvent-ils pour cela recourir à des valeurs différentes ?** |
| **Les SMA peuvent-ils modéliser l'évolution des valeurs, dans le temps et/ou dans leur champs d'applicabilité ?** |

#### *1.8.3.1. Négociation et image de soi*

Le système de valeurs auquel nous adhérons en dit long sur la façon dont nous nous voyons nous-mêmes. Sans entrer ici dans le débat – particulièrement difficile – de la cause et de l'effet, contentons-nous de l'évidence qu'une renégociation des valeurs s'accompagne d'une redéfinition de soi. À cet égard, les développements en robotique revêtent une importance particulière, car depuis toujours, l'homme demande à la machine de définir qui il est :

> *L'homme n'est pas une machine, c'est une machine plus « quelque chose ». Et c'est ce « quelque chose » qui définit l'homme. Dès lors, plus notre reflet artificiel est ressemblant, plus nous en savons sur nous-mêmes et plus nous sommes vexés de nous voir ainsi représentés.* […] *Aujourd'hui plus que jamais, notre conception occidentale de l'homme est entièrement ancrée sur notre appréciation des performances et des limitations des machines. Nous nous observons dans le miroir des machines que nous savons construire et dans ce reflet nous évaluons notre différence*[192].

Et ces mêmes auteurs de poursuivre que la peur de la machine provient de sa fonction de miroir : elle nous force à une éternelle redéfinition du « quelque chose » dans l'équation de l'homme à une machine plus ce « quelque chose » quand bien même nous nous aimons comme nous sommes. Mais ce qu'il faut bien voir, c'est nous-mêmes qui conférons à la technologie ce pouvoir exorbitant de redéfinition perpétuelle.

Les développements récents en robotique nous livrent un exemple éclatant de cette dynamique. C'est ainsi que Sherry Turkle[193] a pu mettre en lumière un changement dans la conception que les

[192] G. CHAPOUTHIER et Fr. KAPLAN, *L'homme, l'animal et la machine*, pp. 122-123. Un même enseignement peut se lire chez G. CHAZAL, *Le miroir automate*, pp. 30-46. La question de savoir ce qui serait exactement cette différence dépasse de loin le cadre du présent mémoire ; il n'est cependant pas interdit de penser que, pour l'homme occidental au moins, cette différence réside en quelque sorte dans le sexe, à qui, depuis toujours, nous demandons qui nous sommes (cf. les thèses développées par M. FOUCAULT dans *Histoire de la sexualité I* : l'Occident – faute d'avoir élaboré un *ars erotica*, a pourtant produit un corps de textes important qui se classe plutôt comme *scientia sexualis*).

[193] Sh. TURKLE, *Authenticity in the Age of Digital Companions*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 62-76.

enfants se font de l'état d'être en vie (*aliveness*). Dans une étude de Piaget des années '60, la notion avait une connotation nettement physique, car basée sur la motion autonome. Dans les années '80, le critère devient psychologique : l'objet informatique est considéré « comme en vie » s'il peut penser par lui-même, l'autonomie requise étant donc de nature psychologique ou intelligente. À ce stade, la personne diffère de la machine parce qu'elle est émotionnelle. Cette barrière va être franchie avec les artéfacts relationnels, dont l'apparition des Tamagotchi en 1997 annoncent la couleur. Le critère d'être en vie devient alors la connexion émotionnelle que l'enfant peut établir avec l'objet, et les fantasmes qu'il peut entretenir vis-à-vis des sentiments de l'objet à son égard. L'objet devient un sujet, un bébé.

Les Tamagotchis seront bientôt suivis par des *Furbies*, des *Aibos*, des *Real Babies*, dans un succès jamais démenti. Comme le montre bien l'exemple du Tamagotchi, les liens que l'être humain peut tisser avec ce dispositif somme toute rudimentaire sont très forts, même si à l'évidence, toujours dans l'exemple du Tamagotchi, ils sont manifestement à sens unique. Nous retrouvons là « l'effet Eliza » : le désir d'interactivité suffit pour nourrir le phantasme de la réciprocité, par une attitude de projection, dans des échanges dont il est avéré qu'ils changent notre image de soi[194].

Il faut dire que Turkle pose le problème de ses propres observations en termes *d'authenticité*, plus précisément, en termes de *perte* de ladite authenticité. Or il est légitime de penser que nous avons affaire ici non pas à un problème d'authenticité mais à l'illustration d'une renégociation de notre perception de nous-mêmes, de ce que c'est pour nous d'être un homme, de ce qui compte comme être vivant. À ce propos, rappelons le thème fondamental du désir : l'être humain n'a pas seulement besoin de (et le droit à) l'autonomie, comprise ici comme absence de contrainte extérieure, mais aussi et surtout à la création de relations et de liens sociaux. Ce besoin fondamental d'appartenance, de se sentir accepté, voilà qui constitue le point de départ de l'éthique de la sollicitude (*care ethics*)[195].

Nous ne pouvons pas nous empêcher d'évoquer à ce propos une nouvelle d'Hugues Bersini, dont l'argument ne manque pas de piquant : au volant de sa voiture, Mme Yen constate avec horreur que son compagnon digital est sur le point de mourir de faim ; elle en oublie la route et cause ainsi un accident qui entraîne la mort d'un enfant[196]. À partir de ce qui n'aurait pu être qu'un fait divers tragique, Bersini va développer une réflexion sur la solitude et, de façon plus frappante encore, sur cette nouvelle possibilité qu'a l'être humain de créer un sentiment (une simulation ?) de lien de filiation avec une machine. Ce faisant, il redéfinit en même temps ce que la filiation veut dire. Bersini suggère clairement cet aspect par la mise en scène d'une panoplie de relations filiales différentes : celle de Mme Yen à sa fille biologique, femme d'affaires surmenée qu'elle voit rarement ; celle de Mme Yen à sa progéniture virtuelle, qui est devenue le centre de ses préoccupations et de sa sollicitude ; celle de la maman anonyme à son enfant fauché sur la route ; toutes celles enfin que nouent des couples japonais avec des enfants robotiques qu'ils préfèrent à des enfants biologiques.

---

[194] J. HAM et A. SPAHN, *Shall I Show You Some Other Shirts Too? The Psychology and Ethics of Persuasive Robots*, dans R. TRAPPL, *A Construction Manual for Robots' Ethical Systems*, p. 73.
[195] S. PAYR, *Towards Human-Robot Interaction Ethics*, dans *op. cit.*, pp. 40-58.
[196] H. BERSINI, *Le Tamagotchi de Mme Yen et autres histoires*, pp. 5-26.

Comme la postface de la nouvelle vient nous le rappeler, non sans ironie, cette redéfinition risquerait bien d'entraîner la démographie du pays à sa perte.

Le robot nous montre peut-être ici une face cachée de l'homme, une face plus humble de cet être qui se voudrait maître de son corps, de son esprit, de sa volonté, guidé (et appelé) par la seule raison à être la cerise sur le gâteau de la Création. Nous le voyons subitement en quête constante d'attention, débordant de besoins affectifs qu'il est prêt à assouvir avec n'importe quel objet un tant soit peu sensible à sa présence. Bref, ce fier seigneur devient un petit chien, qui a peur de la solitude, de l'abandon, et qui passe sa vie à se serrer au plus près de la meute.

En guise de conclusion, précisons tout de même que toute renégociation n'est pas bonne à prendre. Ainsi certains auteurs férus de posthumanisme proposent une renégociation radicale de notre humanité : partant du principe qu'une machine sera éthiquement meilleure que nous, êtres humains, il est possible de déduire en toute logique que nous avons le devoir de construire ce type de machines avant de « sortir de scène avec quelque dignité »[197]. Cette logique a cependant pour prémisse que les valeurs de l'être humain, ce qui compte pour nous, auront entretemps fort changé. De fait, pour l'auteur posthumaniste en question, les seuls acquis de quelque valeur de l'homme se réduisent aux mathématiques et à « la » science.

Plus généralement, il faut prendre garde que les renégociations que nous propose l'éthique des machines, même si elle n'en est évidemment pas seule responsable, ne vont pas dans le sens d'une instrumentalisation de l'être humain, ou dans le sens d'une déresponsabilisation. Cette dernière menace est exemplifiée par la place de l'agir humain en contexte de guerre, dont Misselhorn nous rappelle qu'elle s'est réduite comme peau de chagrin suite à la façon industrialisée de faire la guerre[198].

#### *1.8.3.2. L'obéissance à la règle reste la norme…*

Ce qui précède pourrait faire croire que les règles sont faites pour être renégociées. Quand bien même la renégociation est essentielle pour expliquer leur appropriation, quand bien même aussi les règles changent en fonction du milieu et du temps, cela n'implique pas que la transgression d'une règle soit un concept vide de contenu : si elle peut être transgressée, cela veut dire qu'on soit tenu en principe de la respecter, faute de quoi le concept même de règle serait dénué de sens. Même si nous admettons des exceptions aux règles, celles-ci doivent être soigneusement délimitées[199].

Il convient évidemment de s'entendre : le respect de la règle – ou de la valeur – implique d'identifier d'abord quelle règle est pertinente. Prenons pour exemple le petit enfant : son espace de négociation

---

[197] E. DIETRICH, *Homo Sapiens 2.0. Building the Better Robots of Our Nature*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, p. 537.

[198] C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 176. Dans le même ordre d'idées, R. ARKIN note l'explosion du nombre de troubles psychiatriques chez les combattants au XX^e^ siècle, alors que la plupart d'entre eux n'ont jamais tué personne (*Governing Lethal Behavior in Autonomous Robots*, pp. 29-36).

[199] H. PUTNAM, *Le Réalisme à visage humain*, p. 371.

est nul. Ceci, non seulement parce qu'il se trouve dans une situation de dépendance affective et matérielle, mais aussi parce que ses connaissances du monde ne sont pas assez étendues pour qu'il puisse proposer une *justification concurrente* de son comportement. Son choix est simple : obéir et suivre le comportement « éthique » imposé, ou donner libre cours à sa pulsion immédiate. Il n'a pas de « dilemme » entre valeurs conflictuelles.

La question de l'obéissance dans ce sens semble évacuée dans le cas d'un agent technologique : si son architecture cognitive comporte un certain nombre de contraintes à respecter, nous voyons mal comment il pourrait les transgresser afin de donner libre cours à ses pulsions. Et pourtant, si comme nous l'avons vu (§ 1.8.1), l'agent technologique est investi d'un système de valeurs (rappelons-nous, il peut avoir reçu la curiosité comme valeur fondamentale, ce désir de découvrir le monde et d'interagir avec lui), des conflits entre telle ou telle contrainte particulière et cette « pulsion » fondamentale deviennent plausibles : imaginons-nous seulement toute la tension qui pourrait naître d'un robot « curieux » censé respecter la vie privée ou « protéger » des données à caractère sensible. En définitive, il appert donc que le respect de la règle s'apprend – s'entraîne même – et ce, non seulement dans le cas des enfants et des animaux domestiqués, mais aussi dans celui des machines.

Faut-il cependant en conclure avec McDermott[200], que la question de l'obéissance est centrale dans le dilemme éthique ? La question mérite d'être posée, d'autant plus qu'une explication évolutionnaire des principes moraux met fortement en lumière le rôle de l'obéissance[201]. Vu sous cet angle, le code éthique permet au groupe de prospérer, même au détriment de l'intérêt individuel. La prudence (ou le conservatisme, si l'on y tient) de ces codes sert souvent de garde-fou contre des entreprises innovatrices trop téméraires, même si elles semblent cautionnées par le sens commun. Ceci expliquerait que dans une culture dominante, où la sélection basée sur le respect de la règle n'opère plus, l'affaiblissement éthique comporte des aspects autodestructeurs.

Quoi qu'il en soit du pouvoir explicatif à accorder aux thèses évolutionnistes, l'idée maîtresse ici est que la morale est plus « intelligente » que l'individu. Pour reformuler l'idée sans faire usage d'une notion contestable, disons que l'obéissance aux règles et principes peut être plus importante que le calcul individuel explicite. L'éthique se ferait donc plutôt *malgré* que *grâce à* notre intelligence[202]. La thèse éclaire par un jour nouveau certains aspects de la réalité. Ainsi, un être intelligent a plus de ressources pour manipuler son entourage ou adopter un comportement immoral. À l'inverse, si nous nous tournons vers l'attribution du blâme et de la reconnaissance, il est clair que la bêtise n'est pas un motif de mansuétude. Il éclaire aussi autrement le cas des nouveau-nés : ceux-ci ont un statut éthique particulièrement prégnant[203], malgré, ou même grâce à, leur infériorité en matière d'individualité et d'intelligence.

---

[200] Cf. Dr. McDermott, *What Matters to a Machine?*, dans M. Anderson et S. L. Anderson, *Machine Ethics*, pp. 88-114.
[201] J. Storrs Hall, *Ethics for Machines*, dans *op. cit.*, p. 38.
[202] Cf. ce mot de M. Dubois : « L'intelligence sans moyen naturel de prédation, c'est l'invention du mal ; prédation multiforme, non canalisée par la biologie. L'intelligence s'affute dans la prédation de ce qui est accessible, y compris de son prochain, par l'invention d'outils toujours plus élaborés pour dominer. » (*La métaphore et l'improbable*, p. 45).
[203] Là-dessus aussi, les évolutionnistes ont leur mot à dire : « Un bébé crocodile, au sortir de l'œuf, mord. Un bébé tigre, lui, assoiffé de lait, avide d'un corps chaud et ami, veut avant tout aimer, être aimé. Mamelles à téter, première innocence des mammifères. Plus tard, reconversion brutale. Maintenant, tout à la douceur. » (H. Michaux, *Poteaux d'angle*, p. 13).

Bref, la louange ou le blâme ne sont jamais attribués à cause d'un comportement que nous qualifierions de particulièrement « intelligent ». Pour prendre un exemple peut-être un peu forcé, imaginons le cas d'un flâneur qui, en se promenant le long du canal, voit un malheureux se noyer. S'il se décide à le sauver, celui-ci sera reconnaissant – il aurait même quelque part le devoir d'être reconnaissant. La raison en est qu'il sait que le premier homme a mis en péril sa propre vie, a dû affronter la peur de la mort, à laquelle il sait par projection que l'autre est sujet. En d'autres termes, la reconnaissance est attribuée par le patient moral en fonction de l'autotélisme projeté sur l'agent. Faire de cette attribution une affaire d'intelligence n'est pas permis : non, l'homme sauvé n'est pas reconnaissant parce que l'autre a correctement analysé la situation. Et non encore, une IA ne développera pas spontanément un intérêt propre qu'elle pourrait le cas échéant mettre en péril au profit d'autrui. Là encore, il semble que l'intelligence soit relativement neutre dans nos jugements éthiques. Le point de vue de l'ingénieur en la matière reste certes légitime : un minimum d'intelligence pour comprendre la situation est un prérequis obligé, dont l'éthique des machines ne saurait faire l'économie et qui constitue même un des enjeux majeurs de la discipline en l'état actuel de son développement.

Pour conclure, disons que la tension qui existe entre les aspirations individuelles et les règles collectives sont tout à fait dignes d'intérêt de l'examen éthique, et doivent retenir notre attention pour la suite : comment l'agent technologique fera-t-il face en cas de conflit entre mobiles individuels profonds et les exigences de la situation ? Saura-t-il seulement reconnaître le conflit ? Quelle priorité accordera-t-il aux deux termes ? Comment justifiera-t-il ces priorités ? Il y a là un terreau fertile aux interrogations les plus diverses.

#### *1.8.3.3. Négociation et langage*

Nous en avons déjà brièvement touché un mot : la négociation présuppose la possibilité de confronter des *justifications concurrentes* d'un même comportement. Or, pour l'être humain, la justification passe nécessairement par le langage. Les liens qui unissent l'homme à *son* langage sont profonds et bien connus. Il semble presque inutile de rappeler, dans ce contexte, la position de Descartes, qui fait de la faculté de langage le fondement même non seulement de notre humanité mais de la raison tout court[204].

Au centre de la négociation langagière, nous trouvons l'extension de sens : si nous disons que le robot « veut » quelque chose, nous attribuons au robot une volition. Or il est clair que la volition du robot n'est pas la même que celle d'un écrivain qui veut remporter le prix Goncourt ou d'un chien qui veut attraper un os. Tout en voulant attribuer ce qu'évoque le mot « vouloir » au robot, nous enrichissons, chemin faisant, le sens du mot : l'un ne va pas sans l'autre. Au centre de cet enrichissement du

[204] D. J. GUNKEL, *The Machine Question*, p. 59.

langage, nous trouvons la métaphore. Selon Ricœur, la métaphore s'analyse comme un acte de prédication – x est P, ou P(x) – qui crée du sens, ou plus précisément, une *ressemblance*[205].

Or il se fait que l'informatique est une grande consommatrice de métaphores : déjà le mot « programme » en était une lorsque l'ordinateur a fait surface[206]. Cet apport n'a jamais été démenti par la suite : algorithme génétique, colonies de fourmis, réseaux de neurones, pare-feu… nous en passons, et bien des meilleurs ! Dans le vocabulaire d'Aristote, nous pouvons dire que les modélisations informatiques fournissent des définitions formelles. La définition formelle s'est avérée être une véritable fabrique de ressemblance. En retour, l'efficacité du dispositif à s'acquitter de sa fonction sert de preuve, ou plus exactement, de garantie. Une formalisation, ou spécification, d'un comportement d'un état 1 à un état 2 n'est en fait rien d'autre qu'une *fonction*. La fonction est en principe toujours informatisable[207], c'est-à-dire que le comportement observé, au prisme de la fonction, peut être porté sur le plan « techno-logique »[208]. Entre la colonie de fourmis « physique » et la colonie de fourmis techno-logique, l'informatique a enrichi la langue d'une nouvelle métaphore : la colonie de fourmis étant un dispositif techno-logique qui permet de trouver « le plus court chemin » entre deux « points »[209].

En retour, grâce à la métaphore, la fréquentation de cet autre niveau de réalité auquel renvoie le monde applicatif et logiciel, vient enrichir notre vécu du monde physique. C'est cet effet que Stengers a qualifié d'entre-capture : nous investissons une machine de sens, et en retour, elle nous modifie également. Ce n'est d'ailleurs pas un hasard si Stengers prend comme exemple de capture la notion informatique de « programme »[210]. Et la question se pose si la métaphore doit être comportementale pour avoir quelque efficace. Comportementale, la métaphore l'est très clairement dans l'exemple du vol, donné par de Wallach et Allen[211] : le vol est à l'évidence une propriété fonctionnelle, partagée entre l'avion et l'oiseau. L'extension de sens se fait ici autour d'un noyau de sens de nature comportementale « se déplacer dans les airs, le cas échéant au mépris du vent » : *voler*, en effet, n'est pas *planer*.

La négociation langagière peut bien sûr échouer : toujours selon Stengers, nous dirions alors que *ce n'était qu'une métaphore*. Or si un terme n'est « que » métaphore, il s'agit d'un effet de sens qui a échoué à s'imposer. Il y a eu refus de (re)négociation, refus d'une capture qui, une fois admise, nous

---

[205] C'est tout le propos du chapitre *La métaphore et la sémantique du discours* de P. RICŒUR, *La métaphore vive*, pp. 87-128. Pour l'usage argumentatif de la métaphore, qui y voit une sorte de schéma d'inférence abductif, voir le chapitre *Analogie et métaphore* de Ch. PERELMAN, *L'empire rhétorique*, pp. 145-157.

[206] Pour l'histoire du mot, ainsi que les idées qu'il véhicule, voir J.-P. SÉRIS, *La double origine de la notion de programme*, dans Fr. TINLAND, *Ordre biologique ordre technologique*, pp. 133-148. Dans cet article, l'auteur note d'ailleurs la concomitance de l'emploi de la même idée, sinon de la même notion, en informatique et en génie génétique.

[207] C'est au moins ce que dit la thèse ou plutôt la conjecture de Church-Turing, selon laquelle toute fonction « intuitivement » calculable peut être calculée par une machine de Turing (cf. C. MISSELHORN, *Grundfragen der Maschinenethik*, p. 19).

[208] Le néologisme est dû à J.-M. SALANSKIS (*Le monde du computationnel*, pp. 123-162) et désigne le niveau de réalité où les idéalités informatiques opèrent.

[209] Tous les guillemets sont de rigueur, car « le plus court chemin » devient en informatique à son tour métaphore dans la famille des algorithmes métaheuristiques. Tout problème à optimiser peut être décrit comme un « espace », où les points sont des états. Parcourir l'espace des états (ou « des solutions ») ne doit donc pas être pris au pied de la lettre.

[210] Voir I. STENGERS, *La propagation des concepts*, dans EAD., *D'une science à l'autre*, pp. 12-23.

[211] W. WALLACH et C. ALLEN, *Moral Machines*, p. 67.

aurait obligé à regarder autrement le réel. C'est par ailleurs aussi la thèse de Gunkel[212], lorsqu'il affirme que pour introduire des nouvelles manières de penser, il faut faire violence à la langue. Toutefois, l'échec ponctuel de la médiation langagière ne doit pas conduire à nous masquer son efficace très réelle. Pour faire sentir ce point, citons l'exemple rapporté par Zdenek d'un agent immobilier virtuel[213] : il permet à des personnes intéressées par un achat immobilier de faire des suggestions dans une base de données d'offres, de faire des tours guidés virtuels de propriétés, etc. Afin de faciliter l'interaction entre l'agent et l'utilisateur humain, ses concepteurs l'ont personnifié : ils lui ont donné un nom (REA), un sexe (le présentant comme une femme), un visage et une voix. En réalité, cet agent est un ensemble matériel particulièrement complexe ; sa personnification est sans nul doute dans une bonne mesure un effet de discours. En effet, la métaphore personnificatrice[214] présente REA comme un agent qui « utilise » d'autres composants comme une base de données, des écrans, des microphones… alors que, en réalité, la ligne de démarcation entre les différents composants du système est loin d'être aussi nette. Cette personnification sert le but de ce que Gilbert Hottois[215] a appelé l'accompagnement symbolique de l'innovation technique : loin d'être l'essence de la création technique, elle n'est qu'une mesure visant à faciliter son appropriation par l'homme.

| **Question** |
| --- |
| **La négociation passe-t-elle nécessairement par le langage, ou y a-t-il d'autres codes symboliques – éventuellement transposables en langage – qui peuvent également servir ?** |

[212] D. J. GUNKEL, *The Machine Question*, pp. 208-210. À noter que l'auteur préfère le terme emprunté à Derrida de « paléonymie », c'est-à-dire la réutilisation et le recadrage de notions anciennes pour explorer une nouvelle contrée de sens.

[213] Cf. l'article de S. ZDENEK, *Artificial intelligence as a discursive practice*.

[214] Notons l'emploi très lâche du terme « métaphore » tel qu'il est utilisé par l'auteur de l'article, car si celui-ci insiste particulièrement sur les effets discursifs qui se donnent à voir dans les manœuvres rhétoriques des articles scientifiques qui présentent REA, force est de constater que la personnification est induite autant – sinon plus – par *l'image* : or une *image* (fut-êlle de synthèse) n'est pas une médiation symbolique.

[215] G. HOTTOIS, *Généalogies philosophique, politique et imaginaire de la technoscience*, p. 80.

# Chapitre II. Le paradigme multi-agents

*Si tu traces une route, attention, tu auras du mal à revenir à l'étendue.*

Henri MICHAUX

Il y a quelques années, dans le cadre d'un travail de fin d'études, nous avons été amené à programmer une simulation à base d'agents. Lors de ce travail, nous avons eu l'occasion de consulter quelques travaux sur le paradigme multi-agents. Nous nous sommes alors posé pas mal de questions, à des degrés d'élaboration toutefois très variables. Ainsi, la question de *l'unité* du paradigme : après tout, tant de pratiques diverses se réclamant du paradigme, il n'était pas clair pour nous si le terme « multi-agents » désignait autre chose qu'un effet de mode, un mot-clef dont les moteurs de recherche académiques s'étaient entichés. Nous nous proposons donc de parcourir très brièvement ces pratiques, avec une attention particulière pour les passerelles qui existent entre elles.

Parmi ces pratiques cependant, le sujet principal demeure, pour nous, la simulation à base d'agents. Assurément, il s'agit d'une pratique scientifique relativement nouvelle, en plein essor de surcroît. Comme nous le verrons, la pratique de la SBA soulève la question du *modèle* en science : quelle peut donc être la valeur d'une construction qui cherche à exister à côté – *malgré*, peut-être ? – des autres pratiques scientifiques que sont le prélèvement de données dans le monde sensible, la théorie et l'expérimentation ? En compulsant des exemples de cas scientifiques recourant à la SBA, nous sentions confusément comme un glissement dans le centre de gravité des préoccupations ; une façon de faire science à mille lieues des expériences de physique sous vide ou des dissections de pigeons tant prisées par les écoles.

Une troisième question est celle des pratiques discursives qui entourent la SBA : nous venons déjà de dire qu'un facteur d'unité, une passerelle entre pratiques, est la *métaphore* de l'agent. Or comment une science peut-elle accueillir une pratique dont la base métaphorique, tellement dense, semble à première vue si difficile à concilier avec une approche dénuée des mirages de la subjectivité ?

Enfin, une dernière question a été directement inspirée par le sujet de ce mémoire : si nous demandons à la SBA d'éclairer l'éthique – ce sera l'objet du troisième chapitre – quel rôle peut-elle jouer dans une *justification éthique*, dont nous avons vu l'importance au premier chapitre ? La justification pourrait se présenter de trois façons : par la visualisation immédiate des résultats et par la discursivité métaphorique, mais nous nous interrogerons aussi sur les possibilités, la pertinence et les limites d'une mathématisation de la simulation.

Pour répondre à ces questions soufflées par la curiosité, hélas tout nous faisait défaut à l'époque : le temps, les bagages théoriques… En plus d'être résolument hors sujet, elles soulevaient une recherche bien au-dessus de nos moyens ! Ce chapitre se propose tout d'abord de formuler ces questions, de façon moins naïve qu'à l'époque espérons-nous. Nous aurons à cœur de les articuler – dans la mesure du possible – en fonction des enjeux présentés au premier chapitre, mais une mise en garde s'impose : les travaux se réclamant d'une recherche éthique au moyen de simulations multi-agents ne seront abordés qu'au troisième chapitre. Ainsi, le chapitre qui va suivre ne fait que nous donner les moyens d'une fin encore à venir.

## 2.1. L'agent comme métaphore

Avant d'être une architecture robotique, un paradigme de programmation ou un algorithme, l'agent est avant tout une *métaphore*. Encore faut-il se hâter d'ajouter la précision suivante, capitale : ce qui sous-tend ce large éventail de technologies, c'est d'abord une *métaphore efficace*. Ceci ne devrait pas nous étonner : comme nous l'avons vu au premier chapitre (§ 1.8.3.3), l'informatique souvent procède ainsi. Métaphore, certes, mais notre pari – au risque de la répétition – est d'y voir un reflet de la réalité humaine dont elle s'est abstraite. Les exemples en sciences ne manquent pas : allant de la sélection darwinienne des biologistes à la main invisible des économistes, la pensée métaphorique précède en quelque sorte les calculs mathématiques dans la compréhension des faits humains[1].

Aussi pouvons-nous dire que la métaphore constitue un premier niveau de *spécification*, pour utiliser un vocabulaire propre à l'informatique. Avançons dans la spécification en donnant la définition formelle d'un agent[2]. Définissons l'environnement – provisoirement – comme un ensemble fini d'états instantanés et discrets :

$$E = \{e, e', \ldots\}$$

Définissons ensuite le répertoire d'actions possibles pour l'agent comme suit :

$$Ac = \{\alpha, \alpha', \ldots\}$$

Une toute première définition – provisoire – de l'agent est alors la suivante :

$$Ag : E \rightarrow Ac$$

Ainsi défini, l'agent associe un état de l'environnement à une action.

Ce petit ensemble de définitions – et plus particulièrement, la définition de l'agent – appelle quelques remarques. Tout d'abord, nous verrons plus tard dans ce chapitre que ce type d'agent est, en réalité, une très forte simplification. Pour l'heure, contentons-nous de dire que cette simplification correspond à une sous-classe des agents, dits *purement réactifs*. L'exemple classique d'un tel agent

[1] M. J. F. DUBOIS, *La métaphore et l'improbable*, p. 59.
[2] Adaptée (en simplifiant quelque peu) d'après M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 34-38.

est un thermostat : si la température ambiante est satisfaisante au regard de certains critères, il enclenche le chauffage ; sinon il le désactive.

Ensuite, attirons l'attention sur la nature de cette définition. Pour un informaticien, il est habituel de définir la réalité qui lui est soumise comme une *fonction*, au sens mathématique le plus strict. Une telle définition – rigoureuse – est également fonctionnelle dans le sens où ce terme a été pris au premier chapitre : l'agent n'est conçu que comme un comportement, quelque chose qui agit sur autre chose, en l'occurrence son environnement. Une telle définition préserve la possibilité de nous interroger sur l'identité des agents, sans toutefois l'imposer. La définition ne dit rien non plus sur les façons d'obtenir un tel agent : en principe, tous les outils de l'IA peuvent être mis à sa disposition.

Retenons de tout ceci qu'un agent, en SMA, repose sur une définition purement fonctionnelle : l'agent puise, plus ou moins autonomement, dans son répertoire d'actions pour produire un effet dans l'environnement. Par voie de conséquence, l'agent ne peut être dissocié de l'environnement pour lequel il est créé, et qui est pourtant hors de lui.

### 2.1.1. L'agent en quête d'identité ?

L'agent réactif, très simple – pour ne pas dire simpliste – que nous venons de voir, correspond à ce que nous avons appelé, au premier chapitre (§ 1.7.2.1), une unité d'agir. Continuons maintenant en donnant la définition complète d'un agent. Il nous faut pour cela introduire la notion d'exécution *r* (*run*) d'un agent :

$$r : e_0 \xrightarrow{\alpha_0} e_1 \xrightarrow{\alpha_1} e_2 \xrightarrow{\alpha_2} \dots \xrightarrow{\alpha_{u-1}} e_u$$

Une exécution n'est rien d'autre qu'une alternance d'états de l'environnement entre lesquels les actions de l'agent viennent exercer leur influence. Soit *R*, l'ensemble de toutes les exécutions possibles d'un environnement. Soit $R^E$, le sous-ensemble de ces exécutions qui se terminent (momentanément) par un état de l'environnement. Nous sommes maintenant armé pour la définition finale de l'agent dit « standard » :

$$Ag : R^E \rightarrow Ac$$

Formulé simplement, l'agent est la fonction qui associe à une action non pas seulement l'état présent de l'environnement, mais *tout l'historique d'exécution* qui lui est accessible. Nous retrouvons ici, fût-ce métaphoriquement, la mémoire épisodique dont il a été question au premier chapitre (§ 1.7.2.2) : sans mémoire des évènements passés, aucune interaction digne de ce nom n'est possible. La définition de l'agent standard incorpore donc la composante principale de la forme d'identité que nous avons qualifiée de « psychologique ».

Or, l'historique des changements d'état d'un environnement indifférencié ne dit encore rien des interactions de notre agent avec les autres agents. Pour accéder à l'identité sociale, il faut que l'autre soit reconnu, au cours même de l'interaction ; l'environnement doit être *perçu* comme *composé* d'agents avec qui il est possible d'entretenir un contact dans la durée. Afin de clarifier ce point,

donnons quelques exemples de simulations où ces notions revêtent un intérêt particulier. Écartons toutefois d'emblée un malentendu possible : dans le langage des informaticiens, la seule notion connue est celle d'un *identifiant*. Vues sous l'angle technique, en effet, toutes les distinctions que nous pouvons être amené à faire peuvent s'implémenter de la même façon ; ce n'est qu'à *l'usage* que des différences apparaissent.

La première simulation sur laquelle nous voulons nous pencher est GRANULAB[3], simulation conçue pour mettre en lumière la mécanique des matériaux granulaires. Exprimé simplement, la simulation s'intéresse aux grains de sable qui, suite à la pression exercée des uns sur les autres, vont donner lieu à des mouvements parfois importants dans les milieux désertiques. Dans ce contexte, plusieurs concepts utilisés dans le domaine scientifique à modéliser pourraient se voir discerner le « rôle » d'agents : les grains de sable eux-mêmes bien sûr, mais aussi les empilements, ou encore les lignes de force jouant sur eux. Pour des raisons à la fois de *précision* des prédictions et de *performance* des calculs (rappelons-nous que la métaphore, en informatique, se doit toujours d'être efficace), le grain de sable a été privilégié comme agent principal dans GRANULAB.

Un grain de sable est un « objet » très simple, n'ayant que trois propriétés : une masse, un rayon, ainsi que la position de son centre. Son environnement est constitué des grains de sable en contact avec lui, soit au-dessus de lui (ses entrées) ou en-dessous de lui (ses sorties). Pour calculer le réseau des forces de poids et de frottement qui vont déterminer si une configuration donnée de grains est en équilibre ou non, un mode de résolution individuel est utilisé où, partant du haut de l'empilement, les grains se voient répartis en trois catégories :

- les grains « résolus », qui ont pu trouver un équilibre en proposant un jeu de forces à leurs sorties tenant compte des forces appliquées sur eux par leurs entrées ;
- les grains « à résoudre », dont les forces d'entrée sont connues mais qui n'ont pas encore une proposition d'équilibre sur leurs sorties ;
- les grains « en attente », dont au moins une des entrées n'est pas encore connue.

Si un grain « à résoudre » ne trouve aucun équilibre, il devra « demander » à un ou plusieurs de ses voisins déjà résolus de se recalculer, afin de lui proposer un nouveau jeu de forces en entrée. L'efficacité de l'algorithme de résolution se mesure alors au nombre de résolutions individuelles nécessaires à l'obtention d'un équilibre de l'ensemble.

Dans cette simulation, le grain de sable est à la fois un support de calcul, nécessaire à la résolution graduelle des forces d'équilibre, et un support d'information : sans lui, il faudrait recourir à une méthode statistique globale, perdant par là les moyens d'exprimer la surface de contact entre grains où encore l'orientation des forces. Le grain de sable est donc clairement, eu égard à la terminologie du premier chapitre, une unité d'agir, jouant un rôle pivot dans l'algorithme de résolution, *comme si* les grains de sable contribuaient activement à la recherche de leur propre équilibre.

Pouvons-nous pour autant dire que les grains de sable ont une *identité* ? Ce n'est pas une affirmation absurde dans l'absolu, dans la mesure où elle ne fait que filer la métaphore de l'agent. Cependant, il

[3] Voir J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 17-24.

convient de remarquer que l'agent n'interagit avec ses voisins que sur la base des valeurs de ses propriétés au moment du calcul. N'importe quel grain de sable de mêmes dimensions, de même masse, provoquerait le même type de réponse. L'historique des échanges avec ce *même* agent n'entre jamais en ligne de compte. Nous pouvons même durcir le trait : l'agent ne se voit lui-même qu'au travers des mêmes propriétés qu'il prend en considération pour interagir avec ses semblables. Il pourrait être un autre, son comportement n'en serait nullement affecté. Il s'ensuit que ni le critère énoncé pour l'identité psychologique ni celui pour l'identité sociale ne sont satisfaits ; l'agent est purement réactif. L'identité des particules, ontologiquement pourtant épaisses, ne joue aucun rôle ni dans le déroulement de la simulation, ni même après pour juger des résultats.

Tournons-nous maintenant vers un autre exemple, celui de RIVAGE[4], exemple tiré de l'hydrodynamique. La simulation doit permettre d'explorer le ruissellement, en l'absence de méthodes alternatives viables. Les agents principaux y sont des « paquets » d'eau. La différence d'avec GRANULAB est remarquable : là où, dans le monde observable, un grain de sable existe naturellement en isolat, il en va autrement dans le cas de l'eau. Les paquets d'eau ne bénéficient de l'individuation que sous l'effet conjugué de la métaphore et des contraintes des outils informatiques. En effet, la discrétisation du continu confère – dans le contexte très spécifique de l'informatique, grande pixélisatrice devant l'Éternel – aux paquets d'eau une unité d'agir.

RIVAGE ne s'arrête cependant pas là. La simulation réifie les regroupements des paquets d'eau en mares et ravines. Ces entités sont dotées des propriétés suivantes : un identifiant, une liste de paquets membres, l'historique d'évolution des membres, leur position. Mares et ravines peuvent se déplacer ou fusionner : il leur faut donc un critère de permanence, une règle de gestion des identités afin de pouvoir donner sens à la notion de progression spatiale, de déplacement. Certes, ces mares et ravines sont encore des unités d'agir : sinon, il n'y aurait tout simplement pas de sens à parler de *mouvement* : une chose qui se déplace (ou qui est déplacée) doit rester la même tout au long de son parcours. Nous pourrions même dire que ces mares et ravines ont une identité, vu qu'elles ont un identifiant et un certain type d'historique. Cependant, à y regarder de plus près, il appert que tant l'identifiant que l'historique ne jouent aucun rôle lors de la simulation. L'identifiant joue ici le rôle d'un nom de variable, qui permet de récolter des résultats en fin de simulation.

L'identifiant, en d'autres termes, renvoie à un type d'identité que nous n'avons pas encore rencontré au premier chapitre, une identité que nous pourrions qualifier de « macroscopique » : celle-ci est nécessaire à l'observateur *externe*. Cet étiquetage d'une unité d'agir n'est finalement rien d'autre qu'un *biais de l'observateur* qui – quoiqu'indispensable – est tout à fait extérieur au modèle étudié. Un tel identifiant peut cependant rapidement revêtir un autre sens, sans que son implémentation informatique change. Il suffit pour cela qu'il soit utilisé à l'intérieur même de la simulation par les agents eux-mêmes. L'identité est alors « microscopique » et implique la reconnaissance des agents entre eux.

Prenons l'exemple de DOMWORLD[5], dont le sujet d'étude sont les rapports de dominance entre membres d'une société de primates. Chaque animal y est modélisé par un identifiant, sa position,

[4] J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 24-32.
[5] *Ibid.*, pp. 57-63.

ainsi qu'un registre d'estimations de la valeur de dominance : ce registre est mis à jour chaque fois que l'animal affronte un autre membre du groupe. Nous ne nous arrêterons pas sur les règles, multiples et complexes, qui régissent les interactions d'affrontement. Pour notre propos, il suffit de souligner qu'un tel registre présuppose, chez l'animal, une capacité de reconnaissance mutuelle. Ceci recouvre ce que nous avons appelé l'identité sociale au premier chapitre. Pour accéder à l'identité qualifiée de psychologique, il suffirait que l'agent, lors de l'affrontement, se base non seulement sur une estimation agrégée de dominance, mais se souvienne plus particulièrement de ses propres affrontements passés avec tel opposant particulier. Même si DOMWORLD n'implémente pas ce mécanisme supplémentaire, il est tout à fait possible de dégager de cette possibilité un résultat important : rien, dans le paradigme multi-agents, ne s'oppose à la dissociation d'identités psychologisante et sociale.

### 2.1.2. L'agent intentionnel, ou la métaphore mentaliste

Nous avons vu comment une simulation à base d'agents peut fournir ce qu'il faut pour soutenir la métaphore d'un agent qui se perçoit, qui est perçu, comme ayant une identité. Ceci peut se faire de manières diverses – un identifiant, un registre… – ce qui compte, c'est le résultat d'interaction qui s'ensuit. Même si cette idée peut paraître contre-intuitive, il faut garder à l'esprit que l'être humain serait amené à traiter de façon nettement différenciée un robot qui se « souvient » de lui et d'un robot qui, à chaque rencontre, nous fera exactement le même discours, les mêmes gestes, évacuant par là même la possibilité d'une rencontre, fût-elle unilatérale, n'ayant lieu que dans le chef de l'être humain. L'identité se mesure donc « naturellement » aux effets produits dans le réel de l'échange.

Si nous représentons cependant la vie d'un agent comme un chemin à parcourir, l'identité ne nous renseigne que sur l'endroit d'où nous venons : elle parle origines, et non pas destinations. L'agent doit, en termes plus conceptuels, être doté d'informations relatives à sa *finalité*, aux objectifs qu'il poursuit, ou qu'il « doit » poursuivre. « Finalité », « chemin »… voilà des images bien parlantes ! Une métaphore, cependant, ne suffit pas à elle toute seule ; elle est germe en même temps que tuteur de l'innovation, elle n'en dit pas le tout pour autant. Dans le domaine de l'informatique, elle est amorce et prétexte de spécification. Dans le cas des agents mentalistes, l'intentionnalité est comprise comme un outil d'abstraction[6], permettant de traiter un agent intentionnel comme une boîte noire. L'idée à la base de ce traitement provient de Daniel Dennett, qui distingue trois niveaux d'explication : physique, de conception, intentionnel. Plus un système est compris, moins l'intentionnalité intervient pour l'expliquer.

Dans les travaux consacrés aux SMA, un type d'architecture est spécialement conçu pour donner corps à cette idée : l'architecture dite BDI (pour *Beliefs, Desires, Intentions*), librement inspirée des travaux du philosophe Michael Bratman. Le propos de Bratman[7] consiste notamment à défendre l'idée d'une intention non réductible à des désirs. Un agent peut avoir beaucoup de désirs,

[6] M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 31-34, 55, 83.

[7] Soulignons que notre présentation est basée sur un article relativement récent du philosophe (à savoir M. E. BRATMAN, *Intention, Practical Rationality, and Self-Governance*) alors que les travaux informatiques renvoient toujours vers un livre paru dans les années quatre-vingt du siècle dernier.

éventuellement incompatibles avec ses propres croyances, ou qu'il sait irréalisables à cause d'un manque de ressources. Une intention, en revanche, doit être vue comme un plan, que l'auteur conçoit comme un modèle cohérent d'actions en vue d'obtenir un effet que l'agent estime réalisable. Attentif aux exigences de la rationalité pratique, Bratman souligne que par « cohérence », dans ce contexte, il faut entendre deux choses : la compatibilité des intentions entre elles, d'une part, et la cohérence entre moyens et fins, d'autre part. La compatibilité des intentions d'abord : si nous avons l'intention de A et l'intention de B, nous devons croire que A et B peuvent advenir tous deux dans le monde sans se barrer mutuellement la route ; la cohérence des moyens et des fins ensuite : si nous avons l'intention de A et que nous croyons qu'il faut le moyen M pour réaliser A, cela implique que nous devons également avoir l'intention de M. La cohérence des intentions *entre elles* relève, selon Bratman, de la rationalité pratique, alors que la cohérence entre les croyances d'un agent est quant à elle l'apanage de la rationalité théorique.

Nulle contrainte ne pèse en revanche sur les désirs. Cependant, en tant que tels, ceux-ci ne mènent pas à l'action. Comme nous allons le voir dans le paragraphe suivant, les désirs ne jouent qu'un rôle subalterne dans les architectures BDI. Pour l'instant, nous nous limiterons à rappeler que la distinction entre désir et intention a sa place dans une discussion éthique, car le désir peut être conçu comme relevant d'un fait de valeur. En effet, ce n'est pas parce que nous éprouvons un désir que nous avons l'intention de le réaliser. De manière analogue, nous ne pouvons pas projeter toutes nos valeurs ensemble sur une même réalité, il faut choisir. Nous désirons tous fermement le bonheur, pourtant les plans que nous concevons s'arrêtent d'ordinaire à l'achat de notre prochaine voiture. En réalité, tout le sens de l'éthique se joue dans l'interstice entre le désir et le plan. C'est dire que la distinction entre intention et désir a tout sens dans un contexte éthique. Elle devrait être maintenue.

#### *2.1.2.1. L'agent raisonneur : PRS et ses descendants*

Or ce n'est pas toujours le cas, comme dans une implémentation très connue de l'architecture BDI, le système de raisonnement pratique PRS[8]. PRS étant un logiciel prototype, développé en LISP, l'industrie a surtout utilisé son successeur, dMARS, développé en C++ et dont la première mouture a vu le jour en 1995 à l'Institut australien d'intelligence artificielle (AAII). Comme PRS, l'architecture de dMARS est entièrement *réactive* : les agents répondent à des *évènements*, puisent dans une bibliothèque de *plans* des recettes pour y faire face. Un plan est activé sur base de patrons, faisant intervenir le type d'évènements, ainsi que certaines conditions contextuelles que les agents évaluent en fonction de leurs croyances. Si plusieurs patrons peuvent s'appliquer, le programmeur peut les départir en faisant appel au méta-raisonnement. Un plan entièrement instancié et actif est appelé une intention.

Nous constatons que dans PRS, la notion centrale est celle du plan, c'est-à-dire une suite d'actions à entreprendre accompagnée d'une condition d'invocation ; désirs et intentions sont ici quasi

[8] Les développements sur PRS, dMARS et leur descendance sont basés sur G. WEISS, *Multiagent Systems*, pp. 774-775, ainsi que les trois articles suivants : V. MASCARDI, D. DEMERGASSO et D. ANCONA, *Languages for Programming BDI-style Agents : an Overview* ; R. EVERTSZ, M. FLETCHER, R. JONES, J. JARVIS, J. BRUSEY et S. DANCE, *Implementing Industrial Multi-agent Systems Using JACK* ; M. D'INVERNO, M. LUCK et M. GEORGEFF, *The dMARS Architecture*.

confondus et ne correspondent qu'à un changement de statut du plan : si, à la suite d'une mise à jour des croyances de l'agent, la condition d'invocation du plan est satisfaite, celui-ci devient un désir. Tous les désirs se retrouvent dans une file d'attente. Le désir programmé pour exécution devient par ce fait même une intention.

À travers dMARS, PRS a connu une riche descendance. dMARS, en effet, a été appliqué dans un certain nombre de domaines, parmi lesquels l'aéronautique avec OASIS (logiciel gérant le trafic aérien entrant et sortant d'un aéroport) et SWARMM, application de gestion de combat aérien[9]. En outre, un héritier tout à fait direct de dMARS est JACK, produit commercial développé en Java, qui procède surtout d'une simplification de l'architecture héritée de PRS.

JACK distingue nettement deux types d'évènements : d'une part, les évènements dits « normaux », qui correspondent à des percepts environnementaux ; d'autre part, des évènements BDI, comme une mise à jour de croyance. Afin de faciliter l'intégration en entreprise, un élément de plan peut être non seulement un bloc de raisonnement BDI, comme dans dMARS, mais aussi tout simplement du code Java. Les blocs de raisonnements peuvent d'ailleurs être réutilisés, exactement comme du code Java, grâce aux « capabilités », sorte de modules réutilisables de fonctionnalité orientée-agent.

Étant un produit activement développé, JACK continue d'ailleurs de s'enrichir de nouvelles constructions. Parmi elles, mentionnons les équipes. Cette notion permet, en JACK, de modéliser des structures sociales, qui peuvent avoir leurs propres croyances et présenter une structure hiérarchique : ainsi une équipe hérite (de façon ascendante) des croyances de ses sous-équipes et propage, de façon descendante, ses croyances le long de la hiérarchie d'équipe, le tout avec possibilité de filtres.

Une autre branche héritière de PRS, cependant, a connu un succès encore plus important, car elle a fini par donner corps à l'idée d'une « programmation orientée agents ». Ici encore, l'intentionnalité telle qu'elle se donne à voir dans le BDI est avant tout une métaphore, comme l'a bien dit Yoav Shoham – le fondateur de la programmation orientée agents – lorsqu'il se sent obligé de s'expliquer sur son penchant pour une terminologie « pseudo-mentale » :

> *Intentional terms such as knowledge and belief are used in a curious sense in the formal AI community. On the one hand, the definitions come nowhere close to capturing the full linguistic meanings. On the other hand, the intuitions about these formal notions do indeed derive from the everyday, common sense meaning of the words*[10].

L'auteur poursuit, cependant, en arguant de *l'efficacité* du procédé :

---

[9] Voir à ce sujet la présentation de SWARMM dans G. TIDHAR, *Flying Together. Modelling Air Mission Teams*. dMARS y est utilisé pour le raisonnement et combiné à un composant en Fortran (PACAUS) pour modéliser les aspects physiques du combat. Notons que SWARMM, développé à l'origine en dMARS, est en passe d'être porté sur son successeur commercial JACK.

[10] Y. SHOHAM, *Agent Oriented Programming*, p. 299. Les références à Y. Shoham sont tirées d'un résumé, car l'article original de 1993 n'est disponible qu'au prix fort, même en passant par les abonnements de l'Université.

> *What is curious is that, despite the disparity, the everyday intuition has proven a good guide to employing the formal notions in some circumscribed applications. AOP [agent oriented programming] aims to strike a similar balance between computational utility and common sense*[11].

Des langages de programmation implémentant ce paradigme – ou des bibliothèques logicielles qui rendent ce paradigme accessible dans des langages de programmations classiques – prévoient alors des abstractions[12] qui enrichissent l'agent *individuel* : ses croyances (les informations à sa disposition), ses objectifs, ses plans, ses intentions. La dynamique d'un programme, dans ce paradigme, est basée sur la survenance d'évènements dans l'environnement : ce sont les changements perçus par les agents qui les amèneront à ajouter ou supprimer des croyances ou des objectifs, etc. En réaction à un évènement, l'agent va choisir parmi ces plans – des séquences d'actions qui, dans des circonstances particulières, pourrait aider l'agent à traiter un évènement particulier – celui qui selon sa connaissance du monde va lui permettre de se rapprocher d'un objectif jugé prioritaire, ou pertinent dans le contexte. Le plan retenu pour exécution en réponse à un évènement donné est appelé une intention. Le processus par lequel les désirs – source d'intentions futures – sont produits est la délibération[13], encore que le processus filtre autant qu'il produise ces désirs, sur la base des désirs et intentions courants. La planification, quant à elle, jette le pont entre les intentions et les ressources disponibles.

Parmi les variantes de programmation orientée agents inspirée de PRS, citons AgentSpeak(L). Ce langage est basé sur une abstraction de l'architecture PRS tout en formalisant la sémantique opérationnelle et en simplifiant l'héritage PRS. Ainsi, les plans en AgentSpeak(L) sont beaucoup plus simples qu'en dMARS. Parmi les aspects d'un état d'agent qui peuvent changer, nous trouvons ses croyances, ses intentions, et évènements à traiter. L'intention, en AgentSpeak(L), est réduite à une pile de plans, une séquence de plans instanciés, où tous les plans sauf celui tout au-dessus de la pile sont suspendus. Contrairement à dMARS, les buts et les conditions d'invocation peuvent faire appel à des formules temporelles, dont nous aurons à reparler au prochain paragraphe.

Parmi les continuations contemporaines d'AgentSpeak(L), citons AgentTalk et surtout Jason, dont nous fournirons un aperçu détaillé plus bas. Contentons-nous, pour l'heure, du constat suivant : alors que le moteur de raisonnement dMARS était surtout employé pour des systèmes physiques, la programmation orientée agents s'est surtout retrouvée dans des environnements virtuels, logiciels. Nous aurons l'occasion de montrer comment le pas d'un environnement virtuel « réel » à un environnement entièrement simulé est facilement franchi, et quelles en sont les implications.

#### *2.1.2.2. Les logiques BDI*

Avec PRS, nous avons vu un exemple de système applicatif se servant d'une approche multi-agents ; avec la programmation orientée agents nous avons trouvé un autre niveau d'abstraction, celui d'un

---

[11] *Ibid*.
[12] Voir R. H. BORDINI et J. DIX, *Programming Multiagent Systems*, dans G. WEISS, *Multiagent Systems*, pp. 587-595.
[13] M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 65-69.

paradigme de programmation. Tournons-nous maintenant vers un troisième domaine de l'informatique où la métaphore de l'agent mentaliste a su se montrer féconde, celui des outils logiques de spécification et de vérification formelles. En logique formelle, la recette mentaliste requiert des ingrédients divers[14] : il y faut une logique de la connaissance, une logique de la motivation, ainsi qu'une logique qui capte l'évolution. Cette dernière se présente comme ou bien une logique dynamique ou bien, pour tout le moins, temporelle. Avant de présenter ces différentes logiques, notons qu'elles ont en commun de faire appel, d'un point de vue sémantique, à des modèles dits « de Kripke ». Il sort du cadre du présent mémoire de s'attarder longuement sur les tenants et aboutissants d'un tel modèle. Retenons simplement le strict minimum nécessaire à présenter les logiques qui s'en réclament. Sémantiquement parlant donc, une telle logique se laisse caractériser par un ensemble d'états *S*, une relation d'accessibilité (pour un agent *i*) entre états $R_i$, et une fonction de valuation des atomes logiques *V*.

Les sources logiques modales permettent de modéliser les dimensions épistémique (la croyance) et subjective (la motivation) des agents[15]. Nous sommes alors armé pour exprimer des choses telles que : « Si l'agent i croit que l'agent j a l'intention de rendre p vrai à un moment futur donné, alors l'agent i croit que l'agent j désire rendre p vrai (à un moment futur donné) ». Formalisée, cette expression s'exprime comme suit :

$$(Bel\ i\ (Intend\ j\ A \Diamond p)) \rightarrow (Bel\ i\ (Des\ j\ A \Diamond p))$$

Où il faut bien remarquer les origines distinctes des diverses modalités : l'opérateur de croyance *Bel* provient des logiques épistémiques, les opérateurs de désir – *Des* – et d'intention – *Intend* – sont de nature subjective. Ce genre de raisonnements requiert à l'évidence, de la part des agents, des croyances pertinentes quant aux états mentaux de leurs interactants. Dans les travaux consacrés aux systèmes multi-agents, ces croyances ou connaissances sont dites personnelles ou relationnelles (*acquaintance knowledge*)[16]. Parmi les choses qu'un agent peut savoir d'un autre, nous retrouvons des informations comme son identifiant, les rôles qu'il peut tenir, ses objectifs, ses plans, ses compétences. Soulignons qu'il s'agit de connaissances non pas absolues, mais personnelles à un agent particulier. Elles peuvent donc être périmées ou incomplètes : la gestion de la complexité du monde réel est souvent à ce prix.

Un tel propos a cependant de quoi étonner, car la logique du premier ordre nous a habitué à une vision très carrée du monde : une proposition est soit vraie, soit fausse, et ce, *globalement* : aucun point de vue particulier n'est permis, ni même exprimable. Or tant la composante modale que la composante temporelle des logiques BDI connaissent le phénomène des contextes dits *référentiellement opaques.* Ces contextes opaques (ou encore « intensionnels ») sont absents de la logique classique du premier ordre ; en revanche, dans le langage naturel, ils sont plutôt la règle. Ainsi l'exemple suivant[17] : dans notre système solaire, les expressions « la planète la plus petite » et

[14] La présentation des formalismes logiques se fonde sur W. VAN DER HOECK et M. WOOLDRIDGE, *Logics for Multiagent Systems*, dans G. WEISS, *Multiagent Systems*, pp. 761-797.
[15] À l'attention des connaisseurs, précisons qu'une logique BDI puise typiquement dans une logique modale de type KD45 pour la motivation et dans une logique de type S5 pour les aspects liés à l'information.
[16] Cf. M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 170-173.
[17] A. E. OJEDA, *A Computational Introduction to Linguistics*, pp. 77-83.

«la planète la plus proche du soleil » ont la même *référence*, à savoir la planète Mercure. Si nous désignons ces expressions par les lettres a et b, la logique du premier ordre nous permet de conclure à leur équivalence (*a si et seulement si b*), voire à leur identité : $a = b$. Il nous est donc permis d'asserter les deux propositions suivantes :

*La planète la plus petite est la planète le plus proche du soleil.*

*La planète la plus petite est la planète la plus petite.*

La première proposition traduit $a = b$, alors que la deuxième n'est autre que la loi de l'identité, $a = a$. Or en langage naturel, les deux propositions – devenues affirmations – n'ont pas le même statut : alors que la première nous renseigne sur l'état du monde, la deuxième paraît incongrue[18]. Ce qui est en jeu est le principe de substitution ; le principe aussi de l'équivalence entre sens et référence. Partager la même référence n'est donc pas suffisant, en langage naturel, pour autoriser la substitution de « la planète la plus proche du soleil » par « la planète la plus petite », car ces deux affirmations n'ont pas le même *sens*. En termes informatiques, le sens est une *procédure de calcul* de la représentation de la référence :

*Référence :: planète_la_plus_petite(Référence)*

*Référence :: planète_la_plus_proche_du_soleil(Référence)*

Où l'opérateur « :: » est l'opérateur référentiel. Une formule telle que $A :: B$ sera vraie si $B$ est vrai. Étant donné une certaine base de connaissances, une certaine compréhension du réel, la variable « Référence » recevra dans les deux requêtes une même interprétation en guise de réponse, à savoir Mercure. Or, il est manifeste que dans les deux cas, la même réponse n'aura pas été obtenue de la même manière : le *sens* de nos procédures sémantiques ne coïncide donc pas. De telles perspectives pourraient ouvrir à la représentation du sens dans un cadre multi-agents[19].

Laissons là les aspects modaux pour nous tourner vers l'aspect temporel des logiques BDI. Dans les logiques temporelles, citons d'abord la logique LTL (pour *linear-time temporal logic*) : les opérateurs temporels sont au nombre de trois : ○ (« l'état suivant »), □ (« tous les états futurs ») et ✧ (« à quelque moment de l'avenir »). Le temps lui-même y est considéré comme modélisable discrètement, chaque « pas de temps » correspondant à un nombre naturel. L'ensemble d'états correspond donc à $\mathbb{N}$, la relation d'accessibilité correspond à « être le successeur de » : $\mathrm{R(n) = n + 1}$. Le temps, dans la logique LTL, n'est pas seulement linéaire, il correspond à un mouvement rectiligne, avançant toujours sans regarder en arrière. Pour reprendre la terminologie introduite dans le premier

[18] Ajoutons, pour être complet, que des affirmations telles « Un sou est un sou ! » exhibent également la loi de l'identité tout en étant recevables dans un échange. Leur étude sort toutefois du cadre du présent mémoire. Contentons-nous de dire qu'elles n'acquièrent un sens que dans une interprétation topique, renvoyant l'allocutaire à des valeurs générales de prudence et d'économie.

[19] Le conditionnel, ici, est de mise, car la pratique courante dans la recherche multi-agents est plutôt de faire abstraction de la question de la référence, et d'adopter des approches inspirées par des vues idéalistes, connues sous le nom d'*ontologies sémantiques.* Pour une présentation des ontologies, voir le chapitre *Understanding each other* dans M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 107-129.

chapitre, la logique LTL adopte une sémantique d'inspiration aristotélicienne : le temps y est, comme la distance, une *mesure* physique.

Dans la logique CTL (pour *computational tree logic*), le temps n'est pas conçu comme une droite, mais comme un arbre, où les nœuds représentent des moments de choix. Aux opérateurs temporels LTL s'ajoutent dès lors deux opérateurs de branchement : $\mathrm{E}$ (« il existe un chemin ») et $\mathrm{A}$ (« pour tous les chemins »). Ainsi, il est possible de créer en CTL des formules telles que $\mathrm{A}\Diamond\varphi$ : sur chaque chemin du modèle, il y aura un moment où $\varphi$ sera vrai. Ainsi, chaque nœud du modèle représente un nouveau point de départ, chaque branchement devient un *présent*. La perspective change : d'aristotélicienne, elle devient augustinienne. La notion de branchement dans CTL reste sans équivalent dans une conception physique du temps à la façon d'Aristote. Sachant que les architectures BDI préfèrent largement CTL à LTL, cela doit nous rendre attentif à une chose : l'interprétation d'un formalisme t, t + 1, …, t + n, n'y peut aucunement être compris comme une suite morne et implacable dont le tracé serait connu d'avance. La perspective du temps y est, pour le dire autrement, *interne* au monde des agents. Qui plus est, le temps y est non seulement interne, mais l'expressivité de la logique est telle qu'elle permet de rendre compte de l'irréversibilité du temps[20].

Il est également possible de représenter logiquement l'influence de nos choix non plus sur nous-mêmes mais sur autrui : le choix d'un agent peut, à un moment futur donné, avoir des conséquences pour un autre agent. Pour cela, il faut recourir à une logique « linéaire » temporelle[21], où il faut tout de suite préciser la portée de la notion de « linéaire ». La logique linéaire se propose d'attaquer la représentation de la « consommation » d'une ressource, ainsi que le formule – de façon imagée – Jean-Yves Girard :

> *[…] avec 10 F j'achète un paquet de Camels, avec 10 F j'achète aussi un paquet de Marlboro*, mais pas les deux*. Autrement dit, le principe A => A & A de la logique classique est* dynamiquement *faux (faire A n'est pas la même chose que faire A* ET *faire A).*[22]

La logique linéaire de base multiplie le nombre d'opérateurs. Mentionnons-en deux qui ont un intérêt particulier pour modéliser le vocabulaire du choix des agents : premièrement, la conjonction additive ($\mathrm{A} \,\&\, \mathrm{B}$), qui représente notre choix personnel : nous choisissons soit A soit B, mais non pas les deux. L'autre opérateur est la disjonction additive ($\mathrm{A} \oplus \mathrm{B}$), qui représente la possibilité de choisir A ou B, mais nous ne savons pas quel choix est fait. En d'autres termes, $\mathrm{A} \,\&\, \mathrm{B}$ modélise un choix interne de l'agent, alors que $\mathrm{A} \oplus \mathrm{B}$ modélise une possibilité indéterminée, ou un choix externe. À ces connecteurs provenant de la logique linéaire, la logique linéaire temporelle utilise les connecteurs temporels suivants : *next*, *anytime*, *sometime*, dont le sens est assez intuitif. *Next(A)* : A peut être utilisé exactement une fois, au prochain point t ; *anytime(A)* : A peut être utilisé exactement une fois à n'importe quel point futur ; *sometime(A)* : A peut être utilisé une fois dans l'avenir, mais le choix

[20] Cette observation est importante, dans la mesure où le formalisme a la réputation d'être conçu comme étant réversible, atemporel. Or tout le propos d'une logique telle que la logique linéaire de Jean-Yves Girard est de montrer l'inadéquation de cette abstraction, notamment dans un contexte de consommation de ressources.

[21] Le développement sur la logique linéaire temporelle – à ne pas confondre avec la logique LTL vue plus tôt – se fonde sur l'article de D. Q. PHAM, J. HARLAND et M. WINIKOFF, *Modeling Agents' Choices in Temporal Linear Logic.*

[22] J.-Y. GIRARD, *Le champ du signe ou la faillite du réductionnisme*, dans E. NAGEL, J. R. NEWMAN, K. GÖDEL et ID., *Le théorème de Gödel*, p. 171.

est externe à l'agent. Ainsi, une logique dotée d'un tel appareil formel peut modéliser une dépendance de l'agent non seulement sur ses propres choix, mais également des bifurcations temporelles sur le choix des autres à son égard.

Pour conclure ce paragraphe, posons-nous la question : quel sens donner à tous ces efforts de formalisation ? Ici plus qu'ailleurs, la question pèse de tout son poids, car « la » logique a partie liée avec une image de l'homme historiquement très déterminée. Si les mathématiques nous renseignent sur les relations et proportions dans le monde, la logique a pour vocation d'étudier les schèmes de raisonnement, les relations et proportions non du monde mais du raisonnement. Or en la matière, beaucoup reste à faire : alors que le raisonnement mathématique semble plus ou moins couvert, tout raisonnement, même rigoureux, mais qui, selon l'expression du logicien Jean-Yves Girard, « court-circuite » l'expression mathématique, reste traditionnellement hors de portée de la logique[23]. Par ailleurs, il convient ici de rappeler que des systèmes de raisonnement tels que PRS et dMARS se sont développés indépendamment de la formalisation, pour ne pas dire antérieurement à elle. Cependant, la formalisation nous permet d'accroître notre connaissance de ces systèmes.

Ce qu'elle nous permet de comprendre, ce qu'elle fait ressortir avec netteté des traitements symboliques auxquels se livrent ces systèmes, c'est un certain rapport au *temps*. Un agent qui raisonne dans une logique BDI aurait « conscience » non seulement de points de vue qui ne sont pas les siens, mais accèderait à une certaine conception du temps : selon les choix qu'il fait, choix toujours *liés* à ce qu'il anticipe d'autrui, des pans entiers du possible sont susceptibles de s'évanouir. Qui dit opacité dit, en somme, irréductibilité : la croyance de l'agent i lui appartient en propre. Aucun principe extérieur facile ne peut « calculer », à sa place, les inférences utiles. La formalisation arborescente du temps rend par ailleurs pleinement justice à la conception du temps qui, dans des systèmes comme PRS ou dMARS, reste implicite : comme nous l'avons vu dans ces systèmes, les raisonnements ne se déclenchent que grâce à des évènements, internes ou externes. Or la gestion des évènements – une fois formalisée – traduit des conceptions du temps auxquelles il convient d'être particulièrement attentif.

#### *2.1.2.3. Raisonnement BDI : limites et perspectives*

Finalement, afin de clore cette section consacrée au paradigme BDI, une question s'impose : quel sens pouvons-nous donner à cette technologie ? Quelle image de l'homme et de sa rationalité pratique dégage-t-elle, quelles sont ses forces ou ses faiblesses ? La question mérite d'être posée exactement dans ces termes, car le grand succès du BDI montre – si besoin en était – que le paradigme reflète bien une certaine façon dont l'homme *croit* penser[24].

---

[23] J.-Y. GIRARD, *Intelligence artificielle et logique naturelle*, dans A. TURING et ID., *La machine de Turing*, pp. 112-113. L'auteur fait également observer que l'éclatement contemporain des logiques pose la question de *l'unité du raisonnement*, voire de la Raison. L'esprit humain est-il « libre » de choisir la forme de raisonnement qui convient le mieux à la situation ? Ou ce choix répond-il lui-même à une « logique », formalisable à son tour et qui serait comme une logique première ?

[24] L'observation provient de l'article de R. EVERTSZ, M. FLETCHER, R. JONES, J. JARVIS, J. BRUSEY et S. DANCE, *Implementing Industrial Multi-agent Systems Using JACK*.

Un jugement hâtif du raisonnement BDI poserait l'équivalence entre celui-ci et une théorie de l'action sciemment pilotée par les objectifs : le sujet conçoit l'idée d'un état à atteindre dans le monde ; il calcule ensuite toutes les étapes intermédiaires pour atteindre cet état, avant de minutieusement exécuter toutes les petites actions planifiées. En forçant encore davantage le trait, nous pourrions dire que l'action, en somme, est un mal nécessaire, et que c'est dans les plans que se situe ce qu'il y a de plus vrai et de plus intéressant, car l'activité mentale préalable à l'action est le seul moment créateur de la démarche. Or un tel jugement serait faux : car s'il y a une chose que le paradigme BDI a bien saisi, c'est la conception réactive des buts : un but n'est « programmé », mis à l'agenda, qu'en réponse à un évènement. Un agent BDI ne recourt donc que peu à la *prévision*. En cela, la leçon systémique a été bien retenue, leçon qui veut que les systèmes adaptatifs se méfient de la prévision, car trop difficile, trop peu de données disponibles, etc. Lorsque de tels mécanismes sont présents, la prédiction est souvent même résolument déstabilisatrice : le comportement erratique des marchés financiers, lorsqu'ils sont sous l'emprise des spéculateurs, en est un exemple parlant[25]. L'agent doit donc avoir quelque flexibilité pour mettre en œuvre ses buts.

Toujours est-il que ses buts sont donnés à l'avance, et en nombre limité ; sa démarche reste ainsi sous l'emprise d'une « téléomachie » – si le lecteur veut bien nous pardonner ce néologisme – préalable à toute action. Notre question, notre problème pour cette conclusion, est donc la suivante : est-ce qu'une telle vue offre le meilleur de l'homme, illustre-t-il au mieux les exploits de son esprit comme de ses mains ? Il est permis d'en douter, pour plusieurs raisons ; la première est l'existence d'activités de planification, de conception de systèmes artificiels, de démarches de découverte et de recherche – toutes activités pleinement rationnelles – qui se passent d'objectifs ! Herbert Simon[26] nous rappelle que la recherche est parfaitement concevable avec pour seule référence une « heuristique générale de curiosité », sans d'autres objectifs finaux. Cette attitude de recherche s'étend à des secteurs d'activité très larges :

> *Il est généralement reconnu que si l'on veut acquérir de nouvelles sensibilités en musique, un bon conseil est d'en écouter davantage ; en peinture, de voir plus de tableaux ; en dégustation de vins, de boire de bons crus. La confrontation à de nouvelles expériences entraîne presque certainement le changement des critères de choix, et la plupart des êtres humains sont délibérément à la recherche de telles expériences. Une vision paradoxale, mais peut-être réaliste, des objectifs de conception est de considérer que leur fonction est de motiver l'activité qui à son tour engendrera de nouveaux objectifs.*[27]

L'objectif, ici, est *secondaire*, dérivé d'une *valeur* de curiosité. Par référence à un paradigme piloté par les objectifs, nous pourrions faire état d'un paradigme de la curiosité hédoniste : l'investissement dans la capacité de jouissances ultérieures, l'exploration et l'élaboration des possibilités de conception sont *en elles-mêmes* des expériences enrichissantes. Lorsque l'auteur en conclut que nos objectifs devraient se borner à déterminer les conditions initiales de la génération à venir, en offrant

[25] H. A. SIMON, *Les sciences de l'artificiel*, pp. 265-266.
[26] H. A. SIMON, *op. cit.*, pp. 288-293.
[27] *Ibid.*, p. 289. Nous voyons exposé ici, de façon solaire, le même mécanisme que Jacques Ellul présente sous un jour plus inquiétant, en tant qu'aveuglement de la dynamique technicienne à ce qui n'est pas elle.

autant d'alternatives que possibles aux futurs décideurs, nous touchons à des considérations qui ont une pertinence éthique certaine, et qui cadrent mal avec les conceptions BDI.

L'agent BDI nous montre l'homme comme engagé dans une course effrénée pour atteindre son but, alors que, parfois, le meilleur moyen de l'obtenir c'est justement d'oublier son but. Ce n'est pas le lieu ici de parcourir tous les paradoxes de l'action volontaire – tel l'insomniaque qui, plus il désire le sommeil plus il voit celui-ci s'éloigner, ou encore le joueur de tennis qui perd tous ces moyens après avoir reçu des félicitations un peu trop appuyées de la part d'un adversaire sur sa manière de jouer[28]. Le rappel des paradoxes de l'action volontaire permet cependant une deuxième mise en perspective du BDI, car il nous invite à changer légèrement notre regard sur l'intelligence de notre comportement. Rappelons-nous l'importance de la mémoire procédurale, détentrice de notre savoir-faire. Elle a son équivalent en BDI : c'est la bibliothèque des plans. Si nous changeons notre regard, nous devons constater qu'une bonne partie de l'intelligence des agents se trouve dans cette bibliothèque, qui est entièrement statique. En revanche, le raisonnement propre au BDI et qui, lui, est dynamique, n'est alors plus qu'un épiphénomène, dont bien peu d'adaptabilité peut être attendue.

Les formalismes logiques dans lesquels le BDI se laisse exprimer ont quant à eux un atout majeur : ils engagent à prendre en compte le *pourquoi* de l'action de l'agent, non pas seulement le quoi ni le comment[29]. Ils viennent cependant avec des inconvénients plus ou moins majeurs, dont nous venons de voir le premier : les formalismes logiques traditionnels, en informatique, conçus dans un esprit de grande rigueur, destinées à être validées statiquement, ne rendent pas compte du devenir, du changement. Non seulement l'apprentissage fait défaut, mais même la sélection des symboles contextuellement pertinents et ce, alors qu'une telle fonction s'explique élégamment à l'aide d'une boucle de rétroaction, au niveau sous-symbolique. Nous aurons l'occasion de revenir à cette limitation dans le paragraphe consacré à l'évolutivité des agents.

Un aspect plus technique de la limitation des formalismes logiques se révèle dans la représentation des croyances dans le BDI[30] : les croyances y sont le plus souvent rendues d'une manière excessivement simpliste, plus précisément comme une collection sans structure de formules closes[31]. Un raisonnement sur ces croyances n'est alors rien de plus que le mécanisme d'unification appliquée à des connaissances explicites. Une extension possible est la prise en compte d'ontologies, qui permettent des représentations plus riches des connaissances. Contrairement aux prédicats, les éléments des ontologies sont prévus pour comporter certaines métadonnées (comme l'origine : interne (*self*), un autre agent, un percept, ou encore un étiquetage, sous forme d'URL, de l'ontologie utilisée). Même s'il agit ici d'un enrichissement réel, nous avons vu aussi que la prise en compte de

[28] Pour une discussion des paradoxes de l'action volontaire, voir R. GRAZIANI, qui y a consacré un très beau livre, *L'Usage du vide*. Nous nous basons ici plus particulièrement sur les pp. 108-121.
[29] M. FISHER, R. H. BORDINI, B. HIRSCH et P. TORRONI, *Computational Logics and Agents*, p. 62.
[30] Cette perspective est soulevée dans l'article de R. VIEIRA, A. F. MOREIRA, R. H. BORDINI et J. HÜBNER, *An Agent-Oriented Programming Language for Computing in Context.*
[31] En anglais, *ground predicates* : il s'agit de formules ou prédicats sans occurrences de variables libres. Dans le contexte de l'article, il s'agit le plus souvent aussi de faits, c'est-à-dire des têtes de clause.

la référence est nécessaire au calcul du sens et que précisément la question de la référence est absente de l'ontologie.

### 2.1.3. L'agent autonome

Au premier chapitre (§ 1.7.2.3), nous avons vu que la notion d'autonomie pouvait être vue comme l'intersection entre trois tendances comportementales différentes. Une première peut être considérée comme étant interne, à savoir la création et la poursuite d'objectifs. Une deuxième se joue davantage à l'interface entre environnements interne et externe : il s'agit de la faculté de s'engager dans des interactions adaptées avec son environnement. Une troisième tendance comportementale est de nature collective, désignant la faculté d'un groupe d'agents de réguler leur vivre-ensemble au moyen de normes qu'ils se sont eux-mêmes imposées. Nous remettons la discussion de cette dernière tendance à plus tard, lorsque nous aborderons les normes, et nous nous en tiendrons aux deux premières pour les paragraphes qui vont suivre. La question qui devrait nous guider est celle qui consiste à savoir si le paradigme multi-agents permet de mettre en lumière l'autonomie des agents par rapport au paradigme classique, qui est celui de la *gestion centrale d'un système distribué*. L'autonomie est ici comprise comme un flux de contrôle distinct : un agent est un (sous-)système situé dans un environnement, capable d'action autonome en vue de réaliser ses objectifs[32], même s'il peut arriver – le plus souvent même – que le répertoire des comportements et les mécanismes de décision soient communs à tous les agents modélisés du même type.

#### *2.1.3.1. L'agent qui voulait atteindre ses cibles*

Dans l'architecture BDI que nous avons commentée au paragraphe précédent, les agents sont dits « autonomes » parce qu'ils sont dotés de leurs propres objectifs, et qu'ils doivent veiller à les atteindre. Certes, ceci correspond – nous l'avons vu au premier chapitre (§ 1.7.2.4) – à la définition fonctionnaliste minimale d'un comportement intentionnel, assimilée dès 1943 à un missile qui s'autoguide[33]. Or force est de constater que les buts, dans une architecture BDI, ont été *assignés* aux agents, ils leur ont été donnés *a priori* par le programmeur[34].

Hâtons-nous d'ajouter qu'un tel comportement est en fait dans bien des cas tout à fait voulu, recherché : dans un contexte industriel, une chaîne logistique par exemple, il serait tout à fait indésirable que des agents sortent du cadre relativement strict qui leur a été imposé[35]. En entreprise,

---

[32] M. WOOLDRIDGE, *op. cit*., pp. 21. L'idée qui sous-tend cette insistance sur l'autonomie des agents dans un environnement distribué est la même qui fait parler Nick BOSTROM de « fils téléologiques » (*teleological threads*) dans sa tentative de donner une caractérisation dynamique d'un individu dans un monde technique (*Superintelligence*, pp. 132-134) .

[33] Dans un article de la main d'Arturo Rosenblueth, Norbert Wiener et Julian Bigelow (cité dans D. ANDLER, A. FAGOT-LARGEAULT et B. SAINT-SERNIN, *Philosophie des sciences II*, pp. 1020-1021).

[34] Ce paragraphe est basé sur deux articles : premièrement, M. LUCK, N. GRIFFITHS et M. D'INVERNO, *From Agent Theory to Agent Construction* ; deuxièmement, T. J. NORMAN et D. LONG, *Goal Creation in Motivated Agents*.

[35] Cette réserve sur l'autonomie – cruciale pour comprendre la démarche de type ingénieur – est inspiré de S. KIRN, O. HERZOG, P. LOCKEMANN et O. SPANIOL, *Multiagent Engineering*, plus particulièrement les pages 2-4, 22-26, ainsi que tout le troisième chapitre, consacré à la flexibilité des agents.

un agent est un objet logiciel qui fournit un service, au cloisonnement informationnel rigoureux, dont les objectifs sont toujours donnés extérieurement. Si de tels agents sont dotés de moyens d'actions autonomes pour les poursuivre, c'est essentiellement pour des raisons pragmatiques : quand l'espace des problèmes et l'espace des solutions sont tels qu'ils ne peuvent pas être énumérés par des moyens conventionnels, l'approche agent devient pour ainsi dire inévitable. Ainsi, la complexité de l'environnement requiert une grande flexibilité au niveau des comportements individuels. Ici, la perte de contrôle induite par l'approche agent – il devient impossible de prédire analytiquement le comportement du système dans son ensemble à partir de la somme des comportements individuels – n'est pas considéré comme une qualité intrinsèque, mais bien plutôt comme un *prix à payer*, un *renoncement* à une planification centrale.

Dans d'autres contextes cependant, il peut s'avérer souhaitable de progresser sur la voie de l'autonomie en s'interrogeant sur la façon dont des buts peuvent être crées dynamiquement lorsque l'environnement se fait plus complexe encore. Afin de modéliser la création de buts, un mécanisme de *motivation explicite* doit être ajouté à un agent BDI. Les sources de motivation viennent en petit nombre, par exemple la faim, la peur ou la curiosité. La motivation constitue ainsi un mécanisme de contrôle supérieur, associé à une force et un taux de satisfaction. Si le taux de satisfaction descend en-dessous d'un certain seuil, l'agent se mettra à poursuivre des objectifs à même de ramener la satisfaction à un taux jugé acceptable. Concrètement, si le seuil d'activation est franchi, un évènement attentionnel (*attention-triggering event*) va être émis. Afin de garantir au mécanisme des performances raisonnables, le degré d'activation doit être facile à calculer, et le nombre de motivations devrait rester le plus petit possible afin de réduire le nombre d'évènements supplémentaires à gérer. La force de la motivation modélise quant à elle la sensibilité de l'agent à son égard, permettant ainsi de pondérer les motivations entre elles.

La création d'un but revient à instancier un patron (*template*) : dans les préconditions d'un tel patron se trouvent les motivations pertinentes qui, dès lors qu'elles réclament à être satisfaites, suffisent à déclencher l'instanciation du patron. Dans les postconditions sont renseignées deux choses : premièrement, quelles motivations le but du patron contribue à satisfaire, ainsi que dans quelle mesure (les postconditions positives) ; deuxièmement, les ressources consommées par le but (les postconditions négatives). L'agent choisira le patron qui satisfait au mieux la motivation activée au moindre coût. Le lecteur attentif se demandera comment un but peut consommer des ressources : n'est-ce pas bien plutôt les actions mises en œuvre pour atteindre le but qui conduisent, à proprement parler, à la consommation de ressources ? Nous sommes du même avis que notre lecteur et sommes enclin à y voir une méprise de niveau d'abstraction chez les auteurs consultés. Cette impression se confirme quand nous regardons de plus près l'exemple qu'ils citent : là, la motivation activée est de maintenir les réserves de marchandises à niveau, et les buts possibles sont des commandes à placer chez deux fournisseurs différents, aux caractéristiques différentes en matière de prix, de qualité du produit et de délais de livraisons. Or dans l'esprit de quiconque se veut fidèle au langage ordinaire, la motivation de l'exemple est plutôt un but, et ses buts simplement des moyens.

Or manifestement, cette méprise au niveau de l'abstraction ne doit pas occulter ce qui, aux yeux des auteurs, constitue l'enjeu véritable du mécanisme motivationnel : celui-ci réside en ce qu'il ouvre

une nouvelle possibilité, inconnue du BDI classique : la création *proactive* de buts. En effet, nous avons déjà vu (§ 2.1.2.3) que la sélection de buts dans le BDI est réactive : elle intervient à la réception d'un évènement externe ou après une mise à jour des croyances. Les buts proactifs, en revanche, se créent sur la base d'anticipations de la façon dont le domaine va évoluer. Le gain d'expressivité est considérable : il s'agit en effet d'une croyance par rapport à une croyance *future*. Pour rendre le même effet dans une architecture BDI, il faudrait créer un but, dont les préconditions devraient être continuellement réévaluées. La création motivationnelle de buts permet donc un mécanisme effectif de générations de buts proactifs.

Ce bref exposé de la motivation appelle plusieurs remarques. Premièrement, même s'il est aisé de voir que le mécanisme motivationnel constitue un réel pas en avant vers l'accomplissement d'un certain type d'autonomie, sa simplicité confine pourtant au simplisme. Prenons le cas où un enfant reçoit l'injonction parentale de « vaincre sa peur ». La motivation, la peur, devient ici l'objet d'un but. Ou, pire encore, que penser de certains psychothérapeutes qui nous enjoignent à vaincre « la peur de la peur » [36], quand il faut échapper à des cercles vicieux comportementaux ?

> […] *à la seule pensée que le symptôme se déclenche, se déclenche le symptôme. Ou plutôt, l'anticipation du symptôme suscite la peur, et cette peur se traduit par l'apparition du symptôme. Or la récurrence du symptôme justifie, aux yeux du patient, l'anxiété et la frayeur que ce dernier inspire.* […] *Frankl enjoignait ses patients à provoquer volontairement le symptôme tant redouté afin de les aider à surmonter leur anxiété. Cette méthode de « prescription du symptôme » repose sur l'hypothèse selon laquelle la chose à éliminer est avant tout la peur de sa propre peur, c'est-à-dire la peur panique à l'idée de se retrouver dans une situation qui, telle qu'on se la présente ordinairement, ne peut que provoquer angoisse et phobie.*[37]

Nous voyons là à l'œuvre ce qu'il faut bien appeler une *méta-motivation* : il semble qu'une architecture descriptive devrait donc complexifier la notion de motivation en la *hiérarchisant*.

Une deuxième remarque provient du constat de la proximité entre ce qui est appelé ici « motivation » et ce que nous avons appelé « valeur » dans l'introduction, ainsi que dans la section consacrée aux systèmes de valeurs (§ 1.8.1). La motivation attire l'attention de l'agent sur ce qui, pour lui, devrait importer : une source d'énergie quand il ressent la « faim », l'évitement d'accidents ou d'agresseurs potentiels quand il a « peur », etc. La motivation est, placée dans le registre de la volonté, ce qu'est la valeur dans le registre cognitif : elle procède d'une même démarche topique pour sélectionner ce qui compte, ce qui est important aux yeux de l'agent. Nous aurons l'occasion de revenir sur cette problématique lors du paragraphe consacré à la valeur en contexte multi-agents.

Pour notre propos, troisième remarque, il convient surtout de prendre bonne note de la difficulté que représente l'existence d'une intention « collective ». Certains pourraient objecter que c'est là un mauvais usage de la métaphore mentaliste ; qu'elle occulte plus qu'elle n'éclaire. Pourtant, « l'âme des foules » est un phénomène amplement décrit et attesté, tant chez les sociologues que chez les

[36] Voir R. GRAZIANI, *L'Usage du vide*, pp. 103-125.
[37] *Ibid.*, pp. 108-109.

littéraires. Par voie de conséquence et au moins pour une collectivité d'individus, l'extension de la métaphore est autorisée. Et l'exemple des auteurs, l'Université, nous ramène tout droit, à travers les âges, à l'étymologie du mot : *universitas*, soit une *communauté* de professeurs et d'étudiants. Donc encore dans le cas des institutions et organismes de tout genre, la métaphore nous aide à comprendre cette idée d'une « intention », une volonté, qui dépasse les bornes étroites de notre psychisme individuel.

Dans les architectures multi-agents, la notion d'intention collective s'est montrée féconde – efficace – dans le contexte du travail en équipe[38], où l'intention collective est appelée à expliquer les actions qui ne soient pas seulement coordonnées, mais aussi collaboratives, c'est-à-dire qu'un agent X non seulement se coordonne avec un agent Y pour atteindre un but Z, mais que les deux agents ont également l'intention que Z soit atteint ; une telle disposition d'esprit implique un sens du devoir (*responsibility*) à l'égard des autres membres du groupe. Dans le cas d'une telle intention collaborative, il n'est pas permis à l'agent de laisser tomber son intention de son propre gré lorsque des difficultés se présentent, sous peine d'être mis à l'écart du groupe. L'intention collective recouvre donc deux choses : d'une part, un engagement d'accomplir un but commun (*committment*) et d'autre part, une convention (*committment convention*) qui stipule les termes auxquels les agents individuels peuvent abandonner leur engagement, et le cas échéant, comment le faire. Nous voyons donc clairement ici qu'une intention collective est plus que la somme d'intentions individuelles ; cette question de l'intention collective déborde par ailleurs sur le sujet de la section suivante, qui aborde la question de savoir comment une *collection* d'agents peut faire *système* : sans trop vouloir investir de sens cette notion somme toute creuse, ne fait-elle pas pleinement droit à l'idée qu'un groupe d'individus est autre chose que leur juxtaposition aléatoire ?

Pour conclure ce paragraphe, rappelons-nous les questions que nous avons adressées à l'intentionnalité au premier chapitre : les SMA peuvent-ils se représenter la notion d'intention dans un sens fonctionnellement adéquat, à la manière d'un missile qui « veut » atteindre sa cible ? Il appert de ce qui précède que les SMA n'ont aucun mal ni à implémenter, ni à se représenter dans un formalisme logique de haut niveau, une telle notion d'intention en guise d'outil d'abstraction permettant de gérer la complexité du réel. Non seulement les agents peuvent être dotés d'intentions comprises comme des « objectifs exécutables », mais ils sont aussi capables de se représenter les objectifs de leurs congénères. Dans le procédé au cœur de la mécanique motivationnelle, nous retrouvons le modèle – le patron – dont nous avons vu avec Bratman l'importance quand il s'agit d'élaborer une série d'actions qui soit adéquate à l'égard de l'effet recherché. Même si nous avons vu que la mise en œuvre des implémentations laisse parfois à désirer, il se peut tout de même que nous touchions là à un universel de la rationalité pratique, qui doit être mis en application d'une façon ou d'une autre par toute forme d'organisation un tant soit peu élaborée.

Notons par ailleurs que les systèmes intentionnels ont été développés de façon largement indépendante des modélisations logiques. Nous mesurons ici la remarquable fécondité de la métaphore à la base de ces développements : non seulement elle les a inspirés, mais le sens qu'elle véhicule se communique même au formalisme dont nous pourrions croire qu'il contribuerait à la

[38] Voir la section qui y est consacrée dans M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 165-170.

rendre obsolète. Toutefois, nous avons vu que les SMA considèrent l'intentionnalité comme un moyen, un outil, servant à gérer la complexité des comportements réels. À ce titre, ils ne reculent pas devant l'extension de l'intentionnalité à des agents supra-individuels, ni à s'en séparer dans des cas où ils peuvent en faire l'économie.

#### *2.1.3.2. L'agent évolutif*

Les considérations qui précèdent ont peut-être donné une image fixiste de l'agent, une entité toujours égale à elle-même quoi qu'il arrive. À tort cependant, tout au long de son cycle de vie l'agent peut accueillir le changement, et cela à tous les niveaux. Le corps de l'agent est susceptible d'une évolution environnementale, extérieure, sous la forme d'un changement de position, un déplacement. Sur le plan de la différenciation des agents, celle-ci peut évidemment être donnée d'emblée (structurellement), mais elle peut également être l'objet d'une évolution fonctionnelle, d'une spécialisation, par individualisation progressive. Dans un exemple – sur lequel nous reviendrons plus tard (§ 2.3.4) – SIMANCHOIS, il s'agit d'étudier la régénération du stock d'anchois : l'agent anchois parcourt tout un cycle ontogénétique, allant du stade d'œuf à poisson en passant par le stade larvaire. Une question essentielle à cet égard se situe bien sûr sur le plan comportemental, à savoir la mesure dans laquelle l'agent est capable d'apprendre[39], c'est-à-dire – comme nous l'avons vu au premier chapitre (§ 1.7.2.10) – la mesure dans laquelle un agent est capable de prendre en compte les changements qui affectent son environnement.

Dans le contexte multi-agents, nous ne pouvons cependant pas nous contenter d'une vue purement individualiste de l'apprentissage ; nous pouvons discerner *grosso modo* trois types d'apprentissage dit « réparti »[40] : dans un premier type, dit *apprentissage multiplié*, l'apprentissage reste fondamentalement individuel mais un accent particulier est mis sur la *communication* des résultats de l'apprentissage entre agents. Dans un deuxième type, l'apprentissage *divisé*, la tâche d'apprentissage est divisée en sous-tâches, qui sont alors confiées aux agents. Enfin, le troisième type d'apprentissage est *interactif*, construit sur la base d'une recherche collaborative et négociée d'une solution à la tâche.

L'apprentissage peut donc être individuel dans le cas de l'apprentissage multiplié : nous y retrouvons alors les méthodes désormais classiques de l'apprentissage automatique, tel l'apprentissage par renforcement. Nous en voyons un exemple dans DAMASRESCUE[41], qui se veut une modélisation des actions de sauvetage après une catastrophe naturelle. Dans de telles situations, agir de façon coordonnée dans des délais courts est critique. L'espace dans lequel se déroule la simulation couvre environ 1,5 km², correspondant au centre de la ville de Kobé au Japon. Cette surface est peuplée de 5 refuges et entre 2 à 8 foyers d'incendie, de 800 routes (immédiatement praticables ou non), 700 immeubles (en proie aux flammes ou non), entre 70 et 90 civils à sauver, de 0 à 15 policiers et autant

[39] Voir la contribution de K. TUYLS et K. TUMER, *Multiagent Learning*, dans G. WEISS, *Multiagent systems*, pp. 423-468.
[40] La typologie provient de l'article de D. KAZAKOV et D. KUDENKO, *Machine Learning and Inductive Logic Programming for Multi-Agent Systems*.
[41] Exemple tiré de J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *op. cit.*, pp. 98-102 et 193-198.

de pompiers, de 0 à 8 ambulanciers. Chaque agent a une visibilité de 10 mètres, et la voix porte à 30 mètres. Pour choisir sa prochaine cible, un feu à éteindre ou une zone à atteindre, chaque pompier utilise un arbre de décision, qui fait appel à des critères comme le nombre d'agents nécessaires pour éteindre un feu, la proximité, etc. Cet arbre de décision est « appris » au moyen d'un algorithme de *Q-learning*. Cet algorithme consiste à mesurer la qualité de l'environnement avant et après une action particulière : si la qualité a augmenté, l'action s'en trouve renforcée, promue pour ainsi dire. Sinon, elle s'affaiblit. Ainsi, à partir d'un mécanisme d'apprentissage fixe[42], les dynamiques comportementales peuvent être fluctuantes.

L'apprentissage divisé explore quant à lui une notion qui sans le paradigme multi-agents serait impensable : nous voulons dire la *spécialisation* des agents dans un travail d'équipe. La spécialisation présuppose une certaine forme de conscience des autres agents. Cette conscience peut prendre différentes formes, plus ou moins explicites, dont l'impact sur l'apprentissage n'est pas à sous-estimer. Au niveau le plus fruste, nous trouvons les agents de degré 0 (*0-level agent*), qui n'ont conscience d'autres agents que par les changements que ceux-ci induisent dans l'environnement. Les agents de degré 1 reconnaissent quant à eux l'existence d'autres agents, mais sans avoir de connaissances sur leur comportement, ce qui revient à dire qu'autrui est modélisé comme un agent de degré 0. Viennent ensuite les agents de degré 2 : ceux-ci peuvent retenir leurs observations relatives à d'autres agents. Disposant d'une mémoire par rapport aux connaissances sur le comportement d'autres agents, ceux-ci sont donc vus comme des agents de degré 1. Et ainsi de suite, les relations sociales pouvant se compliquer à souhait. Les auteurs font cependant ici une observation importante : il n'y a aucun lien direct entre les capacités d'apprentissage individuelles d'un agent et le résultat d'ensemble : ainsi, un groupe d'agents qui sont tous de degré 1 aura, pour certaines tâches, des résultats significativement moins bons qu'un groupe où tous les agents sont de degré 0. Les meilleurs résultats, en revanche, sont obtenus par des groupes hétérogènes, c'est-à-dire constitués d'agents de *degrés de profondeur différente*[43].

L'apprentissage peut, dans un environnement multi-agents, aussi être interactif : alors que l'apprentissage par renforcement concerne des comportements complexes pour agents individuels, l'intelligence distribuée ou « en essaim » (*swarm intelligence*) donne naissance à des comportements complexes pour collectivités, au moyen d'interactions locales d'agents qui peuvent être cognitivement faibles. Parmi les techniques d'apprentissage en essaim, citons l'optimisation de

[42] Fixe et, ajoutons-le, assez rudimentaire. Pour une critique circonstanciée, nous renvoyons à l'article de D. KAZAKOV et D. KUDENKO, *Machine Learning and Inductive Logic Programming for Multi-Agent Systems*. Retenons-en ici les lignes de force : l'approche ne prend en compte que l'état courant et ne convient donc qu'à des agents réactifs ; trouver une fonction de récompense quantitative peut être une tâche ardue ; atteindre la convergence peut prendre beaucoup de temps ; finalement aucune explication relative à l'apprentissage n'est disponible.

[43] P. DANIELSON, dans son livre *Artificial Morality*, pp. 159-162, aboutit à des conclusions similaires. Partant du cadre théorique fourni par la théorie des jeux, il en fait la critique en montrant comment la prise en compte du coût de l'information – aspect que ce cadre théorique ignore largement du fait de l'hypothèse d'une information parfaite, complète et partagée entre tous les joueurs – change les résultats des « tournois », qui tendent alors de plus en plus à se stabiliser autour de populations mixtes. Notons, toutefois, qu'il n'est pas possible d'établir une correspondance un-à-un entre l'hétérogénéité de Kazakov et Kudenko – qui a trait au degré de conscience sociale – et la mixité de Danielson, qui est d'emblée comportementale.

colonies de fourmis, qui est le rejeton le plus répandu de la famille[44]. Dans les algorithmes d'optimisation de colonies de fourmis, l'idée de départ repose sur une métaphore tirée de l'observation naturelle : les fourmis communiquent entre elles non pas directement en s'échangeant des messages, mais indirectement via des traces de phéromone, sorte d'hormone, que les fourmis laissent sur leur chemin. Ces traces sont repérées par leurs congénères et, à condition que les traces soient suffisamment intenses, les autres fourmis se mettront à suivre le chemin qu'elles indiquent. Plus grand est le nombre de fourmis suivant une trace particulière, plus celle-ci devient attractive, car elle est renforcée au fur et à mesure. Nous imaginons en outre aisément que les chemins les plus courts seront favorisés, d'où l'effet d'optimisation. La famille d'algorithmes d'optimisation de colonies de fourmis se prête bien à des problèmes qui peuvent être représentés mathématiquement sous forme de recherche dans un graphe, des problèmes tels que celui du voyageur de commerce, des problèmes de routage ou d'ordonnancement. Or comme nous le verrons plus tard au cours de ce chapitre, la simulation multi-agents se conjugue sans effort avec un tel formalisme, eu égard à ses capacités étendues de représentation spatiale[45].

Parmi les formes d'apprentissage interactif, il est également possible de classer les algorithmes « neuro-évolutionnaires », soit la grande classe des méta-heuristiques constituée par les réseaux de neurones formels et les algorithmes dits « génétiques ». Même si ce type d'apprentissage semble moins utilisé dans le cas des agents pris individuellement ou collectivement, il est souvent utilisé pour calibrer les modèles de simulation, comme dans CUBES[46]. Dans cette étude portant sur le comportement des consommateurs, une population virtuelle de 10 000 agents a pu être générée grâce à un algorithme génétique. Les agents-consommateurs, en effet, se différencient entre eux par un nombre important de paramètres : l'âge, l'étendue et la qualité des relations sociales, le taux d'équipement, etc. Nous touchons ici à un thème qui nous reprendrons ultérieurement : le modèle, considéré à son tour comme un agent à travers la vie que lui soufflent ses exécutions au fil du temps, peut évoluer, s'adapter, apprendre.

Tous les apprentissages qui précèdent concernent des apprentissages en « boîte noire », particulièrement adaptés à des apprentissages de compétences, ou des connaissances procédurales[47]. Un apprentissage de nature discursive, symbolique, se laisse plus difficilement appréhender dans ce cadre. Pour les connaissances déclaratives, un apprentissage transparent (*white box*) semble approprié. Dans une situation multi-agents, et quel que soit le type d'apprentissage retenu, la communication du résultat d'un apprentissage devient en effet un enjeu !

---

[44] Il y en a d'autres, comme les algorithmes d'optimisation de colonies d'abeilles. Même si ces algorithmes s'appliquent grosso modo aux mêmes problèmes que les colonies de fourmis, l'idée qui en est à la base présente des différences notables en ce qui concerne le modèle de communication : les abeilles ne laissent pas de traces de phéromones mais communiquent directement entre elles au moyen d'une danse, qui fournit des renseignements précis sur la quantité de nourriture, ainsi que la distance et la direction du chemin à suivre.

[45] Il convient de se rappeler à ce propos la discussion au point « Négociation et langage » (§ 1.8.3.3) : un nombre important d'algorithmes se présentent comme des *parcours*, des recherches dans des *espaces* d'états. C'est dire que, en informatique, la notion d'espace désigne en réalité une *métaphore* extrêmement *féconde*.

[46] J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *op. cit.*, p. 81. Notons que plusieurs plateformes dédiées à la simulation multi-agents connaissent une fonctionnalité générique de calibrage des paramètres de simulation qui, elle aussi, recourt à un algorithme génétique.

[47] La section qui suit, consacrée à l'apprentissage transparent en contexte multi-agents, se fonde sur l'article déjà cité de D. KAZAKOV et D. KUDENKO, *Machine Learning and Inductive Logic Programming for Multi-Agent Systems*.

Qu'il s'appelle enseignement ou transfert de connaissances, l'impératif de la communicabilité change la donne ; elle pose ses propres exigences, notamment en favorisant l'apprentissage dans un format transparent, qu'il soit étiqueté symbolique ou discursif. Cette exigence va s'imposer jusque dans le cas des connaissances procédurales, où pourtant les méthodes dites de boîte noire sont *a priori* plus appropriées. Car si des agents peuvent s'échanger des (parties de) réseaux de neurones formels, ces échanges supposent une connaissance déclarative de ce qui est transmis.

La communicabilité des résultats entraîne aussi une préférence pour des formes avides (*eager*) d'apprentissage par rapport à un apprentissage casuistique ou paresseux (*lazy*) : dans cette dernière forme, l'agent enregistre chaque cas qu'il rencontre en mémoire ; lorsqu'il doit agir sur un cas inconnu, il essaie de le rapprocher d'un cas déjà connu. Une telle forme d'apprentissage est simple à mettre en œuvre et se montre très efficace dans plusieurs types de tâches. L'apprentissage paresseux passe cependant mal à l'échelle, voyant ses performances chuter lorsque le nombre de cas devient important, en raison de son pouvoir limité de généralisation. À l'inverse, l'apprentissage avide s'efforce de créer une théorie générale et de mettre en place une procédure de rappel ultérieur (*recall*) de ce qui est appris. Or cette théorie, étant indépendante des cas qui ont contribué à la construire, peut être transmise à d'autres agents et ce, en vertu même de son abstraction par rapport au contexte d'apprentissage.

## 2.2. L'agent et le système

Dans ce qui précède, nous avons fait la part belle au mentalisme, sondant le for intérieur de nos agents pour y dépister des états internes tels que des intentions, jaugeant leur autonomie… bref nous nous sommes intéressé au niveau « intra-agent ». Dans le paragraphe qui va suivre, nous proposons de déplacer le regard et de l'élever pour voir les agents à l'œuvre entre semblables. Ce niveau, inter-agents, est traditionnellement le principal centre d'intérêt des disciplines scientifiques qui pratiquent la simulation multi-agents ou, pour éviter la confusion dans les acronymes, la simulation à base d'agents (SBA). Le fonctionnement interne de ces agents est souvent très simple, au point de faire parler parfois de « simulation orientée objet » ; ce qui retient l'attention des scientifiques, ce sont les interactions de ces agents, et des configurations comportementales auxquelles celles-ci donnent lieu à un niveau supérieur. Pour user d'un terme à la mode, et sur lequel nous aurons à revenir, nous dirons que ce comportement global « émerge » des interactions simples.

Si la notion d'agent a déjà reçu toute notre attention au premier chapitre, celle de système ne semble pas avoir été thématisée en tant que telle dans l'éthique des machines. De même dans la notion de « systèmes » multi-agents, le système n'est que rarement thématisé. Voilà pourquoi il paraît nécessaire, dans un premier temps, de s'arrêter sur le concept : qu'est-ce qu'un système ? Que voulons-nous dire en affirmant que les agents se retrouvent, d'une manière ou d'une autre, en *système* ? Il reste à savoir si cette question est bien posée. Elle suscite, en tout cas, quelques réserves : n'est-il pas vain, en effet, de vouloir élaborer une théorie générale de ce que serait le système ? Nous voudrions cependant saisir l'opportunité qu'offre la notion pour articuler la tension, sensible dans le paradigme multi-agents, entre deux niveaux d'analyse qui semblent bien

irréductibles l'un à l'autre : les environnements *interne* et *externe* de l'agent : alors que le premier, son environnement interne, fait référence à son individualité propre – nous serions tenté de dire son intimité – l'environnement externe de l'agent peut lui aussi être étudié sous différents angles. C'est, en gros, la problématique de l'environnement externe qui sera étudiée dans les paragraphes qui suivent.

### 2.2.1. Le système en philosophie des techniques

Comme la notion de système n'a pas retenu notre attention au premier chapitre, commençons par une brève mise au point, avant de voir quel éclairage les SMA peuvent ici apporter. Notre guide sera Jacques Ellul, dont les développements sur le « système technicien »[48] peuvent servir de point de départ pertinent à notre propre analyse. Dans sa qualification du système technicien, il reprend à son compte le thème du moteur, cher à Simondon, pour en faire l'image d'une certaine conception de la vie en société. Cette conception se caractérise par son *dynamisme*, dans la mesure où le système est décrit en termes de *genèse* et de *comportements*. À la manière des différentes parties d'un moteur, le système en tant que tout a tendance à réagir de façon évolutive et innovatrice à un changement d'état d'une partie, grâce à des mécanismes de rétropropagation (*feedback*). En outre, les différentes parties font preuve d'une aptitude préférentielle à la combinaison – interaction – interne plutôt qu'externe.

Ce en quoi Ellul se montre d'ailleurs assez original, c'est de s'inspirer de la théorie de l'information pour donner une lecture sociologique de la concrétisation simondonienne. En effet, la théorie de l'information permet à Ellul de dire que le système technicien connaît une *intégration* de plus en plus poussée des composants, dans la mesure où entre les différentes parties du système, des rapports de plus en plus denses d'information se tissent. Chaque partie du système est ainsi conçue comme un demandeur – récepteur d'informations. Le rôle de l'ordinateur est alors celui d'un intégrateur de sous-systèmes qui sont déjà techniques ou que l'ordinateur va obliger à devenir tels. La technique, en effet, doit s'insérer dans un milieu préexistant, mais qu'elle vient toujours modifier. Reprenant les thèses de Simondon, Ellul affirme volontiers que le système se situe dans un milieu et le constitue en s'en nourrissant[49].

L'informatique, en d'autres termes, crée des liens entre réalités sociales d'un type nouveau. Là où, avant l'informatique, l'intégrateur principal était de type institutionnel, l'informatique procède à des mises en relation fonctionnelles. Selon Ellul, un ordinateur seul est une curiosité de foire, il doit être pensé avec la télécommunication, comme un ensemble corrélé, connecté. Pour comprendre l'ordinateur, il faut l'examiner dans ses relations au système technicien global, non pas par rapport à l'homme. Il crée une nouvelle réalité, qui est irréductible aux rapports humains qui lui ont donné naissance. L'informatique, en effet, comme la technique en général, constitue une *médiation* entre

[48] J. ELLUL, *Le Système technicien*, pp. 45-113, ainsi que les articles de M. TRICLOT, *« Milieu technique ». Généalogie d'un concept* et G. CHAZAL, *L'automobile comme interface technique*, tous deux parus dans l'ouvrage collectif de D. PARROCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*.

[49] Cf. G. SIMONDON, *Du mode d'existence des objets techniques* : l'auteur insiste, entre autres à la page 70, entre le conditionnement réciproque qu'exercent entre eux les individus techniques et leur milieu associé.

l'homme et le milieu naturel, elle ne saurait être réduite à un outil ou un instrument. En s'imposant, elle devient totale, au détriment d'autres médiations (poétique, magique, mythique, symbolique…). Elle nous fait entrer dans un univers de *moyens* ; elle tend à imposer un modèle de médiation unique.

Le vécu humain, dit encore Ellul, s'appuie sur le système technicien à la manière d'une interprétation de la lettre, ou du texte d'une loi. C'est dire qu'une interprétation du phénomène technique en termes d'opposition entre un sujet humain et des objets techniques serait inappropriée. Plutôt qu'un ensemble d'objets, la technique est une forme de vie. Même la notion d'objet est discutable : plutôt qu'objet, il faut concevoir la réalité technique comme une *interface*. L'interface technique doit s'entendre comme un intermédiaire entre l'homme et son milieu. Elle permet non seulement de transmettre, ou d'échanger, les informations, mais elle peut également agir sur l'information et sur les termes reliés par elle – d'où la possibilité récurrente d'une rétroaction sur l'homme et qui se présente souvent comme une confusion fin-moyens, comme si l'homme se soumettait à l'intermédiaire.

L'idée d'interface – sinon la notion – peut se montrer féconde en éthique. Nous en voulons pour preuve l'existence d'un courant philosophique « postphénoménologie »[50] : la postphénoménologie part de l'idée phénoménologique de décrire les relations entre sujet et monde, tout en reconnaissant que ces relations passent le plus souvent par une médiation technique. Pour notre propos, il est important de souligner que la technologie contribue à la constitution du sujet moral. Et Verbeek de donner l'exemple de l'échographie dont la pratique systématisée dans les maternités a largement contribué à « rendre homme » le fœtus, plutôt que comme un organisme faisant partie du corps de la mère. La technologie de l'ultrason fournit un nouveau type d'expérience sur la grossesse, sur le devenir-homme et par là, devient une pratique sur soi qui pour ainsi dire par ricochet se répercute sur la constitution du sujet humain. Il convient d'en tenir compte lorsqu'il s'agit d'évaluer la pertinence éthique de l'échographie, car la nouvelle expérience que donne l'imagerie à ultrasons est éthiquement tout aussi importante que le risque de santé induit par les ondes (et qui serait le point de vue éthique traditionnel sur ce phénomène). En définitive, l'éthique postphénoménologique n'est pas seulement une affaire humaine, mais un sujet d'associations multiples entre hommes et technologies.

Terminons ce bref aperçu par noter que ce qui, pour Ellul, est un système, correspond dans le vocabulaire de Simondon[51] à un milieu. Simondon, lui, utilise le terme de « système » dans un sens plus large que celui de milieu. Là où le milieu ne concerne que les relations entre les individus (techniques ou humains) et leur environnement, l'individu technique lui-même peut aussi être vu comme un système. Simondon utilise cependant la notion dans le contexte du processus de *concrétisation*, processus qui décrit une sorte de loi propre au devenir technique et qui veut que l'objet technique évolue en renforçant sa cohésion interne. Par là, l'objet technique développe une sorte de *nécessité* qui lui est propre ; nécessité et donc *sens*. Selon Simondon, dans une production de type artisanal, l'objet technique obéit à une cohérence externe, celle des besoins et d'utilisation,

[50] Nous nous fondons pour cette brève incursion dans la postphénoménologie – qui puise ses origines dans la philosophie des techniques américaine, avec un auteur comme Don Ihde – sur l'article de P. P. VERBEEK, *Obstetric Ultrasound and the Technological Mediation of Morality*.
[51] Voir G. SIMONDON, *Du mode d'existence des objets techniques*, pp. 21-102.

imposée par l'artisan. L'objet artisanal est essentiellement un système ouvert d'exigences, caractérisé par une organisation analytique mais d'une contingence intérieure. L'objet concret, lui, se présente davantage comme un système unifié, doté d'une cohérence interne, qui impose ses exigences dans une organisation synthétique. La concrétisation, en d'autres termes, se présente comme une forme de progrès, dans lequel l'assemblage de plusieurs sous-systèmes (ou ensembles) déjà donnés sont redéfinis par leur fonction complète et unique. Ainsi, l'objet technique s'émancipe de l'intention fabricatrice.

Le système est ainsi compris comme un ensemble de possibilités de médiation entre agents, entre agents et leur milieu, qui conduisent à un résultat de fonctionnement qui est plus que la somme des membres. Une telle situation peut se situer sur le plan de l'élément technique, porteur de technicité, qui en formant système devient individu ; les individus formant système font un ensemble technique, tel qu'une usine ou un chantier. À une échelle encore plus large, nous pouvons distinguer le réseau technique, faisant système, par exemple, sous forme d'une chaîne logistique.

### 2.2.2. Le système comme environnement

S'il y a une chose à retenir du paragraphe précédent, c'est qu'un système ne saurait se résumer à une simple *collection* d'agents : un système ne se constitue que dans un environnement qui agit sur les agents et sur lequel les agents agissent de mille manières. Pour pouvoir parler d'un système, une juxtaposition d'éléments épars n'est nullement suffisante : il faut que ces éléments fassent preuve de relations informationnelles privilégiées entre eux, tendent à une intégration de plus en plus poussée et par là même s'adaptent à un milieu tout en co-créant ce dernier. La « fonctionnalité » d'un système a dès lors partie liée avec son autonomie. En effet, au fur et à mesure qu'un système renforce sa cohésion interne, il accroît en même temps son autonomie fonctionnelle.

Au premier paragraphe de ce chapitre, nous n'avons défini l'environnement en SMA – premier candidat pour jouer le rôle de milieu technique – que comme un ensemble d'états, ce qui s'avère maintenant insuffisant. Il est donc temps de donner la définition complète de l'environnement *Env* :

$$Env = \langle E, e_0, \tau \rangle$$

C'est-à-dire que l'environnement ne comprend pas seulement l'ensemble d'états $E$, un état initial $e_0$, mais aussi la composante dynamique τ, qui se définit comme suit :

$$\tau : R^{AC} \rightarrow 2^E$$

Soit $R^{AC}$, le sous-ensemble des exécutions qui se terminent (momentanément) par le choix que l'agent doit faire. La notation $2^E$ désigne l'ensemble des parties (*powerset*) de l'ensemble E : tau est donc la fonction qui associe les exécutions des agents à l'ensemble dont les éléments sont tous les sous-ensembles possibles de l'ensemble E.

Notons que cette définition reste muette sur l'*état* en tant que tel. En effet, la notion d'état n'est pas caractérisable *a priori* : la définition de l'environnement présente ainsi un caractère plus ouvert que

celui de l'agent. Ceci ne devrait pas nous étonner, car la métaphore de l'agent doit permettre d'insérer celui-ci dans des environnements très divers, tant physiques que logiciels. Dans les environnements logiciels, il faut encore distinguer entre les environnements logiciels *opératoires* (pour éviter le mot « réel »), par opposition aux environnements *simulés*. Dans les environnements physiques, nous nous trouvons à l'intersection entre la robotique et le paradigme multi-agents. Ainsi, parmi les exemples de simulation que nous avons déjà cités, DAMASRESCUE est une simulation qui est passée du tout-virtuel à un module robotique.

La dynamique qui se dessine en filigrane de la définition donnée plus haut semble à première vue étroitement duale[52] : l'agent reçoit certaines entrées informationnelles de l'environnement – sensorielles dans le cas d'un environnement physique, logicielles dans le cas d'un environnement virtuel – et traite ces informations afin de choisir une action qui va à son tour influencer l'environnement. Or, la définition de l'environnement donnée ci-dessus est avant tout négative : relève de l'environnement toute information qui n'est pas portée par les agents en propre[53]. L'environnement en vient ainsi à englober une grande variété de fonctions. Ainsi, et en nous limitant ici à la simulation[54] : l'environnement est le dépositaire de l'ensemble des contraintes et des contrôles d'action, de l'ensemble des propriétés globales ; il a la charge de toute forme de calcul global ; il constitue le support de perception des agents ; finalement, c'est lui le porteur des informations ayant trait à l'espace. « Espace », ici, doit être compris dans un sens métaphorique assez large. L'espace est une topologie, exprimant toutes les relations d'accessibilité entre agents (dont celle de la proximité géographique, mais pas uniquement). Signalons qu'il est possible de modéliser l'environnement lui-même comme un agent d'un type un peu particulier. Même si cela peut paraître étonnant de prime abord, c'est une façon appropriée pour rendre compte des relations réciproques entretenues entre agents individuels et leur environnement[55].

Une telle réification de l'environnement se démarque fortement – il convient de le souligner[56] – de ce que nous avons vu se passer dans les technologies mentalistes : dans une technologie BDI, l'environnement se résume le plus souvent aux messages que l'agent échange avec le monde extérieur, un peu à la manière de bouteilles jetées à la mer. Comme nous venons de le voir, dans les travaux scientifiques consacrés au paradigme multi-agents, l'environnement des agents est souvent caractérisé de façon sommaire et abstraite. Or il est tout à fait intéressant de modéliser pour lui-

---

[52] M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 21-22.

[53] J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 112-118.

[54] Dans le monde physique, plus exactement « dans le cas de robots se déployant dans le monde réel, l'environnement n'a pas à être modélisé car il existe, tout simplement » (J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *op. cit.*, p. 112). Même si cette position a le mérite de souligner une différence de taille entre environnements physique et logiciel, rappelons qu'un robot, également, doit se former une carte cognitive de l'environnement, chose impossible sans un modèle plus ou moins explicite de cet environnement. Le travail de modélisation de l'environnement est certes un peu différent, mais la grande différence se situe à un niveau inférieur : alors qu'un agent opérant dans un monde virtuel ne reçoit que des entrées déjà symboliques, l'agent robot reçoit des entrées de ses capteurs, dont il doit extraire une réalité symbolique grâce – le plus souvent – à un réseau de neurones formels.

[55] Nous avons déjà rencontré cette idée d'interdépendance à plusieurs reprises, en philosophie des techniques avec Gilbert Simondon et Jacques Ellul, en épistémologie chez Isabelle Stengers. C'est aussi le propos de M. DUBOIS quand il relève que la notion de biosphère est justement une interdépendance entre l'environnement et les êtres vivants : ainsi notre environnement changerait du tout au tout si le vivant cessait de produire de l'oxygène (*La métaphore et l'improbable*, pp. 21-27).

[56] Nous nous basons ici sur l'article de H. DJERROUD et A. CHERIF, *Environment Engine for Situated MAS*.

même l'espace dans lequel les agents interagissent. Dans le cas des simulations que nous verrons plus loin, ceci est évidemment une nécessité immédiate : l'espace n'est pas donné, il doit être construit. Or même lorsque le paradigme multi-agents opère dans le monde extérieur, une modélisation de l'espace peut être précieuse : Djerroud et Cherif argumentent que le recours à un moteur de physique newtonienne est excessivement complexe et gourmand en ressources. Pour ces auteurs, la réification de l'environnement s'impose pour que celui-ci soit le garant des lois physiques inviolables, des contraintes fortes qui régissent l'espace. Qui plus est, sans un environnement doté d'informations globales, il n'est pas possible selon eux de donner un sens à l'idée qu'un agent (mentaliste) pourrait entretenir des croyances *fausses* vis-à-vis du monde.

Mentionnons le prototype que ces auteurs ont conçu : il a pour objet des agents (robotiques) qui doivent se frayer un chemin en déplaçant des obstacles. Pour ce faire, le robot se construit une représentation interne, un modèle, de l'espace sous forme d'un système multi-agents, avec environnement explicite. Le modèle va être enrichi avec tout ce que le robot observe du monde extérieur grâce à ses capteurs (*sensors*) : les obstacles physiques, avec leurs attributs de position, de poids, de forme, etc. Ce qui rend le modèle cependant particulièrement intéressant, c'est que le robot peut également l'enrichir avec les contraintes qu'il a découvertes par expérimentation ou tâtonnement : les zones d'accès difficile ou impossible, etc. Une fois construit, le modèle va permettre au robot d'exécuter des simulations afin de *tester* les solutions qu'il envisage avant d'agir dans le monde réel.

L'environnement a donc ceci de paradoxal qu'il est une représentation à la fois d'un agent individuel et de portée globale, portant sur l'espace entier : il « est » le monde dans lequel l'agent pense évoluer. Incarnation du monde et de son ordre, il représente tout ce sur quoi tout le monde ne peut être que d'accord, bref il s'agit d'une représentation que l'agent assume – à tort ou à raison – devoir être essentiellement commune. Or si l'environnement ainsi défini prétend au titre de connaissance partagée, l'agent peut cependant se tromper sur son compte, et ce de deux manières : premièrement, la représentation qu'il se fait du monde commun peut différer de celle des autres agents. En tant qu'elle est censée être commune, la représentation est alors en porte-à-faux ; elle est *inadéquate* pour décrire le monde. Deuxièmement, même si l'agent construit son environnement en se servant des mêmes présupposés ou théories que ses semblables, si sa représentation est donc vraiment commune, il peut toujours arriver que l'agent se méprenne perceptivement : en tant que l'environnement est support d'action, l'agent agira alors de manière *incohérente* par rapport à lui. Ceci reste vrai, insistons là-dessus, même si par ailleurs sa représentation est adéquate par rapport au monde.

Cependant, quand bien même cette caractérisation de l'environnement fait déjà la part belle à la problématique de la pluralité des regards, elle est toujours insuffisante, dans la mesure où elle passe à côté de l'essentiel de l'apport du paradigme multi-agents. En effet, jusqu'ici nous nous sommes surtout intéressé au niveau intra-agent : BDI, autonomie etc., tout cela rend compte de qui se passe, pour ainsi dire, dans le fonctionnement interne d'un agent individuel. À ce niveau de complexité s'ajoute celui qui est constitué de toutes les interactions, de toutes les communications et de toutes les négociations auxquelles les agents se livrent entre eux et que nous pouvons appeler le niveau *inter-agents*. Or le paradigme multi-agents ne se contente pas de ces deux couches de complexité,

intra et inter-agents. En réalité, son apport est à chercher autant – sinon davantage – dans la couche *supra-agents*, c'est-à-dire le niveau qui dépasse les agents individuels[57] : c'est là où se donnent à voir toutes les structures émergentes, permanentes ou non, réifiées ou non, ainsi que les mesures agrégées[58] (plus ou moins localement) sur la population des agents. Alors que l'IA classique a un intérêt presque exclusif pour des facultés individuelles et que le niveau supra-agents est traditionnellement la chasse gardée des statisticiens, le paradigme multi-agents met en relief leurs facultés sociales. *L'interaction* devient ainsi la clef de voûte de l'informatique, ou du moins des systèmes automatisés complexes. C'est là l'innovation majeure !

Pour nous en convaincre, tournons-nous vers une extension de la programmation orientée agents, JaCaMo[59]. JaCaMo est un mot-valise fusionnant Jason, Cartago et Moise. Jason[60] se présente comme un environnement de développement logiciel – écrit en Java – permettant de programmer des agents dans une architecture de type BDI. En d'autres termes, Jason est la composante prenant en charge l'aspect micro, intra-agent. Cartago, quant à lui, modélise l'environnement sous forme *d'artéfacts*. Un artéfact donne accès aux ressources disponibles dans une application déployée. Par ressources, il faut non seulement entendre des sources de données, que celles-ci se présentent sous la forme d'une base de données (relationnelle ou non), d'un système de fichiers ou de services web (de type REST ou non), mais aussi des objets techniques permettant des effets dans le monde réel allant de la simple imprimante à la commande d'une tour de contrôle. Cartago permet aux agents d'exploiter ces ressources de manière uniforme, en gérant les détails techniques de connexion et d'accès de bas niveau, mais aussi des problèmes moins triviaux comme la gestion du partage concurrent des ressources.

La composante qui nous retiendra le plus ici est cependant Moise. Moise se charge d'une forme avancée d'interaction entre agents, à savoir la formation d'*organisations*. Dans Moise, l'organisation revêt trois dimensions. Une dimension structurelle, d'abord : il est possible d'y modéliser des groupes d'individus, les liens entre eux, mais aussi les rôles institutionnels, apportant chacun son lot de contraintes et de droits. Une dimension fonctionnelle, ensuite : l'organisation est le plus souvent dotée d'un objectif global, d'une mission, nécessitant l'élaboration de plans non plus individuels mais sociaux. Une dimension normative enfin : Moise connaît l'existence de normes, entendues dans leur nature déontique comme des permissions et des obligations. Ces normes peuvent d'ailleurs varier selon des lignes structurelles (les rôles des individus) et fonctionnelles (dans le cadre d'une mission de l'organisation).

Moise peut être mis à profit même dans des contextes qui à première vue semblent relativement peu complexes, telle l'organisation d'une unité d'assemblage. Quoique l'unité ne comporte que deux robots autour d'une table rotative, deux gabarits de montage, et trois sources de pièces détachées, les problèmes de parallélisme et de synchronisation des mouvements viennent rapidement compliquer l'ordonnancement des opérations. Appliquer à cette situation une modélisation

[57] Pour des raisons que nous développerons plus loin, il est préférable de voir dans la SBA trois niveaux à l'œuvre plutôt que seulement deux, les niveaux désignés traditionnellement par « micro » et « macro ».
[58] Voir, à propos des mesures agrégées, les sections § 2.3.1 et 2.3.2.
[59] Voir G. WEISS, *Multiagent Systems*, pp. 590-624.
[60] Cf. également la présentation de R. H. BORDINI et J. F.HÜBNER, *An Overview of Jason*.

organisationnelle permet de résoudre ces problèmes de façon élégante, tout en gardant l'architecture interne des agents – les robots de l'exemple – très simple. C'est dire que l'intelligence, ici, réside dans l'organisation, ses rôles et ses plans, plutôt que dans les plans ou intentions des individus. À cet endroit, le rappel d'une thèse centrale de Simon semble opportun :

> *Les êtres humains, considérés comme des « systèmes comportementaux », sont relativement simples. L'apparente complexité de notre comportement, au fil du temps, est pour une grande part le reflet de la complexité de l'environnement dans lequel nous nous trouvons.*[61]

Et l'auteur de multiplier les exemples en faveur de cette thèse, à commencer par l'apprentissage collectif des fourmis que nous avons déjà eu l'occasion de commenter : un observateur fasciné par le tracé alambiqué du parcours d'une fourmi quelconque aurait tort d'y voir davantage que le relief complexe qui caractérise l'environnement de cette dernière. Un autre exemple est celui de l'ordinateur : en tant qu'objet de l'empirie, il précède la théorie : nous pouvons en faire l'histoire naturelle. Cependant, les caractéristiques communes d'organisation – qui donnent toute son importance à l'architecture d'un ordinateur – sont, dans une large mesure, indépendantes des détails du matériel. Enfin, l'intelligence de l'homme se réduit le plus souvent à ce qu'il a *appris* : Simon en tire la conclusion que la performance cognitive de l'homme est bien davantage révélatrice d'un certain environnement social de l'homme – de la qualité des stratégies qui lui ont été enseignées – que d'un système biologique ou psychologique particulier du traitement de l'information. Et l'auteur n'hésite pas à pousser cette logique à bout : l'information contenue dans notre mémoire fait, dans cette optique, partie de l'environnement auquel nous nous adaptons ; le langage nous est transmis et relève ainsi d'une construction sociale ; les deux échappent donc tout autant à notre « intimité », si nous pouvons risquer ce terme étranger au vocabulaire de l'auteur.

Ne nous appesantissons pas sur ces exemples : du primat de l'organisation, de l'intelligence collective sur l'intelligence individuelle, il n'y a qu'un pas vers le primat de l'organisation sur l'individu, tout court. Franchir ce pas entraînerait des conséquences éthiques considérables. Par ce petit détour, nous espérons simplement faire sentir à quel point la prise en compte de l'environnement, au-delà des agents individuels, non seulement est importante – ce serait assez trivial – mais aussi qu'elle est désormais envisageable dans un cadre non seulement *rigoureux* mais aussi, comme nous allons le voir plus loin, *opératoire*.

### 2.2.3. De l'environnement à la simulation

Quelle que soit la nature des liens qui unissent une technologie SMA particulière à l'environnement dans lequel les agents évoluent, l'existence même de ce lien rend le paradigme multi-agents particulièrement apte à la simulation. Ce paragraphe se propose tout d'abord d'illustrer cette affinité par quelques cas concrets. Nous essayerons ensuite de dégager comment cette affinité permet de mieux éclairer, en retour, la notion même d'environnement. Enfin, nous comptons introduire ainsi la

[61] H. A. SIMON, *Les sciences de l'artificiel*, p. 107.

problématique qui sera la nôtre par la suite et qui porte sur la connaissance que nous pouvons espérer recueillir du procédé particulier qu'est une simulation à base d'agents.

Pour illustrer les affinités entre paradigme multi-agents et la simulation, nous nous tournons vers l'ouvrage collectif *Multiagent engineering*[62] : non seulement il se prête particulièrement bien à notre propos, mais il présente en outre l'avantage appréciable de ne jamais se départir d'un point de vue d'ingénieur, ce qui le rend instructif sur la démarche constructive qui motive le recours au paradigme multi-agents. Deux scénarios industriels sont proposés dans l'ouvrage : une architecture multi-agents à l'œuvre dans un contexte hospitalier d'une part ; dans le pilotage d'une chaîne logistique très élaborée d'autre part.

Le scénario hospitalier, *Agent.Hospital*[63], a pour but de prendre en charge l'organisation d'un hôpital. Pour ce faire, l'application intègre des systèmes pré-existants : la gestion des agendas des salles, du personnel soignant et de ses horaires, ainsi que toutes les interdépendances entre ressources et compétences techniques pour les diverses interventions thérapeutiques (MedPAge) ; l'ordonnancement des interventions radiologiques (EMIKA) ; l'analyse, évaluation et planification des essais cliniques (ADAPT) ; la gestion des informations personnelles des patients (ASAinlog) ; la gestion des patients admis aux urgences (AGIL).

*Agent.Enterprise*[64] est un projet ambitieux visant l'intégration de solutions logistiques inter-entreprises avec des applications de planification de production intra-entreprises, où les deux objectifs principaux pour la solution multi-agents sont l'ordonnancement (*scheduling*) et la surveillance des opérations (*monitoring*). Pour atteindre ces objectifs, *Agent.Enterprise* se présente comme non pas un, mais jusqu'à cinq systèmes multi-agents interdépendants : DISPOWEB-MAS, ATT-MAS, IntaPS-MAS, SCC-MAS, ControMAS-MAS. Les agents de DISPOWEB créent le plan logistique initial où sont proposés les commandes, leurs prix et les dates de livraison. Ce plan initial passe par une phase de négociations entre les représentants des différentes organisations. Une fois approuvé, il permet aux systèmes multi-agents intra-organisationnels de planifier la production des pièces détachées (ControMAS) et l'assemblage (IntaPs). Les incidents mineurs (pannes, etc.) sont gérés par le système ATT. Si les incidents sont tels qu'ATT ne peut pas les résoudre, il se dessaisit de l'incident et passe la main aux agents de DISPOWEB, après quoi ceux-ci doivent renégocier entre eux le plan initial.

Dans les deux applications, *Agent.Hospital* et *Agent.Enterprise*, l'environnement est instable et changeant, au point de paraître non-déterministe. Que le non-déterminisme soit intrinsèque ou simplement dû à une information incomplète n'est pas ce qui importe à l'ingénieur, qui déplore ici avant tout les occasions manquées de contrôle de la qualité : peu ou pas de vérification formelle, un besoin important de cas de test. La simulation, dans cette optique, substitue un environnement virtuel à l'environnement « de production ». Elle permet dès lors d'étudier le comportement de

[62] St. KIRN, O. HERZOG, P. LOCKEMANN et O. SPANIOL, *Multiagent Engineering*.
[63] *Agent.Hospital* est présenté dans le chapitre de St. KIRN, Chr. ANHALT, H. KRCMAR et A. SCHWEIGER, *Agent.Hospital – Health Care Applications of Intelligent Agents*, pp. 199-220.
[64] Application présentée dans le chapitre de P.-O. WOELK, H. RUDZIO, R. ZIMMERMANN et J. NIMIS, *Agent.Enterprise in a Nutshell*, pp. 73-90.

l'application dans un environnement particulier, notamment en jouant sur les paramètres. En diffusant un flux réaliste d'évènements, la simulation peut servir de prototype et d'atelier de tests.

L'utilisation d'un programme multi-agents comme simulation apporte certes son lot d'exigences techniques : les agents ne devraient pas sentir la différence entre l'environnement virtuel et de production, c'est-à-dire qu'il ne doit pas y avoir de distinction technique explicite entre agents de simulation et d'application. Même s'il n'y a donc pas d'interfaces supplémentaires à développer, certaines fonctionnalités (comme la simulation du temps) doivent en revanche être soigneusement pensées à l'avance. Surtout, exigence cruciale, les environnements de simulation doivent pouvoir être interfacés avec des agents externes : c'est là où le protocole FIPA ACL s'est révélé précieux[65].

Nous ne pourrons pas faire droit, dans les quelques lignes qui vont suivre, au deuxième grand pilier (à côté du BDI) du génie logiciel en matière d'agents qu'est FIPA ACL. Il s'agit d'un protocole de communication entre agents au moyen de messages. La passation de messages suppose une distinction nette entre représentations interne et externe, garantissant un découplage maximum des agents à travers une épaisse couche d'abstraction des protocoles de communication, de découverte de services, du routage des messages, etc. Nous n'en dirons pas plus sur ces aspects, qui sont d'une grande technicité. Cependant, afin de pouvoir prendre la mesure de la pertinence d'un message FIPA ACL, il faut encore dire quelques mots sur la structuration de ses contenus. En simplifiant quelque peu, FIPA ACL spécifie le format d'une enveloppe de message[66] plus que la structure du message lui-même. Ainsi, parmi les métadonnées présentes sur l'enveloppe, nous trouvons des informations assez classiques comme un identifiant de conversation, un numéro de séquence, les destinataires, l'encodage, une indication sur le type de contenu (le plus souvent des ontologies, dont nous avons déjà parlé). En revanche, ce qui est un peu plus particulier, c'est l'indicateur performatif. FIPA ACL, en effet, appartient à une famille de protocoles dont les origines remontent à la théorie des actes du langage inaugurée par Austin et élaborée par Searle et qui introduit la notion de « force illocutoire » d'un message. Un message ACL inclut ainsi obligatoirement une mention de l'effet de discours voulu par le message : cela peut être une demande (*ask*), un refus (*refuse*), une communication purement informationnelle (*inform*), etc. L'implémentation la plus connue du protocole FIPA ACL est JADE, le *Java Agent DEvelopment Framework*[67].

Actuellement, il existe deux projets faisant le pont entre les agents mentalistes que nous avons déjà vus, et les agents s'échangeant des messages en JADE. Citons tout d'abord Jadex[68], qui est une extension BDI pour JADE écrite au sein de l'Université de Hambourg. Par ce mariage du mentalisme avec FIPA ACL, Jadex unit les deux chefs de file auxquels le paradigme multi-agents a donné naissance : l'aspect intra-agent avec son moteur de raisonnement BDI, l'aspect inter-agents avec les

[65] Au complet : *Foundation for Intelligent Physical Agents – Agent Communication Language 2000*. Le lecteur intéressé peut se référer aux spécifications disponibles sur le site de la fondation (voir notre bibliographie sous FIPA), ainsi que la présentation qu'en fait M. WOOLDRIDGE dans *MultiAgent Systems*, pp. 140-149.

[66] Aux lecteurs pour qui cette référence est utile, nous pourrions dire que c'est la même idée que celle qui préside aux messages SOAP, où là aussi, une spécification très complète des métadonnées du message sous forme d'enveloppe permet une grande abstraction par rapport aux couches de communication inférieures.

[67] Disponible sur la Toile à l'adresse suivante : https://jade.tilab.com/.

[68] Voir A. POKAHR, L. BRAUBACH et W. LAMERSDORF, *Jadex: Implementing a BDI-Infrastructure for JADE Agents.*

messages riches de FIPA ACL. Rappelons-nous ensuite JACK[69], cette implémentation du BDI que nous avons rencontrée au paragraphe consacré aux architectures mentalistes (§ 2.1.2.1) et qui a également développé une extension FIPA ACL.

Les deux outils, Jadex et JACK, ont par ailleurs des extensions toutes faites pour prendre en charge des simulations. Nous le verrons plus loin, une telle extension n'est certes pas triviale sur le plan technique : il faut gérer la simulation du temps (par évènements ou par « pas de temps ») ; il faut pouvoir déclarer des scénarios et fournir des moyens de visualisation. L'essentiel ici est de retenir qu'une fois ce supplément de machinerie disponible, JACK se transforme en véritable plate-forme de simulation, et la traduction d'un programme multi-agents en simulation devient pour ainsi dire *naturelle*.

Les deux applications par lesquelles nous avons ouvert la discussion, *Agent.Enterprise* et *Agent.Hospital*, ont de fait été enrichies d'un environnement de simulation. Dans le cas d'*Agent.Hospital*, un processus patient simplifié fut créé dans la plate-forme de simulation SeSAm[70], dont l'extension FIPA ACL a permis de communiquer avec les agents logiciels déjà en place. Aux dires des auteurs, l'exercice s'est montré relativement facile. Pour ce qui est d'*Agent.Enterprise*, les différents systèmes multi-agents sont accessibles pour simulation sur le banc d'essai (*testbed*) NetDemo, qui offre via une interface web un portail pour instrumenter les différentes composantes et pour visualiser leurs interactions. Là encore, FIPA ACL se trouve au centre de l'intégration. En définitive, ces deux applications illustrent assez bien l'intérêt, dans un contexte d'entreprise, du va-et-vient entre le SMA en production et en simulation, avec un environnement virtuel : les mêmes agents peuvent effectivement être réutilisés.

Jusqu'ici, nous avons tenté de relier environnements réels et virtuels dans un sens unidirectionnel. En effet, la simulation participe d'un mouvement mimétique qui va du réel vers le virtuel. Terminons en montrant qu'il est toutefois parfaitement concevable de pratiquer un chemin en sens inverse, en partant du virtuel pour imposer celui-ci au réel. C'est ce qui se passe notamment dans les jeux vidéo ; ce n'est d'ailleurs peut-être pas fortuit si Jadex est fort usité comme moteur de raisonnement pour développer des comportements plausibles pour les adversaires virtuels. En effet, le jeu[71] présente beaucoup de similarités avec une simulation : après tout, tous deux se déroulent dans un environnement virtuel ! Le jeu comporte cependant une dimension supplémentaire, car il en appelle au *joueur* : celui-ci doit se sentir *impliqué*. Il n'est pas, comme dans une simulation, simple *observateur*, qui à la limite change par-ci par-là l'un ou l'autre paramètre. Il participe au jeu, le jeu lui impose un monde dans lequel il doit « entrer ». Or, une fois qu'il y est entré, l'espace de jeu est centré sur lui et ce, sous deux aspects. Premièrement – notamment dans un jeu en trois dimensions – le point de vue est dynamique, très différent en cela d'une prise de vue de caméra, déjà constituée au moment où l'expérience cinéphile commence ; deuxièmement, l'espace autour du joueur ne prend

[69] Nous renvoyons derechef vers l'article de R. EVERTSZ, M. FLETCHER, R. JONES, J. JARVIS, J. BRUSEY et S. DANCE, *Implementing Industrial Multi-agent Systems Using JACK*.

[70] Voir le chapitre de R. HERRLER et F. KLÜGL, *Simulation*, notamment les pp. 581-593. La plate-forme elle-même est disponible à l'adresse suivante : http://www.simsesam.de/.

[71] Cette brève présentation du sens des jeux vidéo se fonde sur M. ROBERT, *L'acte vidéoludique à la lumière des caractéristiques du vivant*, dans D. PARROCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*, pp. 181-186.

sens que par rapport aux désirs qu'il projette en lui. C'est ce que l'auteur appelle l'immersion, le joueur est au centre d'un environnement qui acquiert son sens par rapport à lui :

> *Par ailleurs [dans le jeu Pac-Man], le joueur doit éviter des fantômes, qui ne cessent de se déplacer, sous peine de perdre « une vie ». Or la configuration de l'espace à partir de ce qui est recherché d'un côté, et ce qui suscite la fuite de l'autre, doit être pensée à partir de ce qui est seul capable de constituer la source de ces différentes valeurs ; elle doit être pensée à partir d'un vivant. Ce n'est que pour un vivant, en effet, que quelque chose est donné comme devant être poursuivi, ou, au contraire, évité. […] Si le vivant est au centre de l'environnement, c'est qu'il n'y a d'environnement que par lui. L'environnement, en tant que tel, a un relief qui tient à ce qu'il est traversé par du sens.*[72]

En cela, le joueur des jeux se distingue non seulement de l'observateur de la simulation, mais aussi de l'utilisateur des artéfacts que sont les chiens robotiques[73], où le monde du partenaire humain fait la loi, où la perception et l'action se font en principe toujours dans et sur le monde physique. Or dans le jeu, c'est le virtuel qui prime.

Ainsi nous voyons se dessiner trois regards vis-à-vis de l'environnement : celui de *l'observateur*, celui de *l'utilisateur*, celui enfin du *joueur* : l'observateur pose son regard de scientifique sur la simulation ; l'utilisateur, à l'affût des fonctionnalités à mettre en valeur et des interfaces à créer, adopte le point de vue de l'ingénieur ; le joueur, enfin, en s'investissant cognitivement et affectivement dans le jeu, en se plaçant en son centre, l'intègre dans son vécu : de ce fait même, le jeu acquiert une épaisseur nouvelle, car il se fait lieu de sens. Trois regards, trois façons de voir le monde du logiciel aussi, sur lesquelles nous aurons à revenir dans le prochain paragraphe.

Pour conclure, rappelons le cheminement de cette section : nous avons vu comment un système peut être vu comme une dynamique d'intégration, se fondant sur l'échange d'information. Or en SMA, ce rôle systémique est prioritairement dévolu à l'environnement. Nous avons relevé que la notion d'environnement, pour mouvante qu'elle soit, est au cœur des différences entre les diverses technologies qui se réclament du paradigme multi-agents. Elle est également décisive pour comprendre comment les agents se rapportent au monde et à autrui. En tant que l'environnement se laisse saisir comme l'ensemble de liens que les agents tissent entre eux, il sera au centre des préoccupations du troisième chapitre, qui aborde plus en profondeur les formes de vivre-ensemble qui se dégagent du paradigme multi-agents.

## 2.3. La simulation et la connaissance

Qu'est-ce que la SBA nous permet de connaître ? Une telle question bute sur le problème, bien plus général, de savoir ce que l'ordinateur nous permet de connaître ? Nous voudrions aborder cette question par un biais, rappelant d'abord qu'un ordinateur peut être vu comme un objet technique

[72] *Ibid*., p. 182.
[73] L'exemple provient de l'article de C. TESSIER, *Autonomie des robots*.

classique : en tant que tel, l'évolution concrétisante, dans le sens de Simondon, est manifeste. La matérialité de l'informatique s'étend sur des kilomètres de câbles, des tonnes de racks de disques, une quantité vertigineuse d'appareils en tout genre ! Et pourtant dans l'informatique, l'homme se pense en dualiste : l'homme, se tendant un miroir dans lequel il voit l'union fortuite d'un esprit et d'un corps, a créé quelque chose à cette image-là – dualiste – telle qu'il se voyait lui-même : la création permet maintenant de juger de la fécondité du miroir, soit du dualisme de départ. L'informatique pourrait même constituer un point de départ fertile à une réflexion sur *l'interface* (notion éminemment informatique) entre le corps et l'esprit, relation restée jusqu'ici toujours un peu mystérieuse dans la pensée dualiste. Si, néanmoins, il n'en est rien, c'est à cause de la tendance des pensées dualistes qui consiste à dévaloriser l'un des deux termes au profit de l'autre, et de vouloir tout expliquer par le terme valorisé (en l'occurrence, le logiciel). Même en admettant que l'informatique peut être caractérisée comme une idéalité agissante, cela ne nous dit en rien *si* et, le cas échéant, *ce que*, l'informatique pourrait signifier pour le monde hors d'elle-même.

Une première idée serait de dire que les réalisations informatiques, étant le produit de l'activité humaine, se réduisent entièrement à l'intention de celui qui les crée. Or de nombreux penseurs en philosophie des techniques, à commencer par Jacques Ellul et Gilbert Simondon, ont soutenu avec force la thèse selon laquelle l'intention du producteur ne suffit pas pour expliquer le phénomène technique. Ceci a pour conséquence que pour comprendre l'informatique, nous ne pouvons pas nous en tenir ni à la cause efficiente, ni à la cause finale, pour reprendre un vocabulaire aristotélicien[74]. Suivant Aristote, un même phénomène est toujours passible de plusieurs explications différentes. De fait, à l'intention du producteur s'opposent toute une série de contraintes purement techniques. Une autre explication, toujours en gardant le vocabulaire aristotélicien, se présente alors presque spontanément : celle de la cause dite formelle. En reprenant l'exemple bien connu de la question pourquoi une maison existe, la cause formelle nous éclaire non sur l'intention de l'architecte (cause efficiente), ni sur les briques ou le bois de la charpente (cause matérielle), ni sur la fonction de l'édifice dans l'espace urbain (cause finale), mais sur « le concept ou l'essence (λόγος) de maison »[75]. Si nous acceptons cette définition, la simulation nous livrera donc quelque chose de la définition, de l'essence, des phénomènes ou activités simulées.

Or une définition, en informatique, correspond à une double opération : d'une part, il faut abstraire un observable du monde sensible pour donner le coup d'envoi d'une nouvelle idéalité. Il s'agit là d'une démarche proprement informatique, qui pousse de raffinement en spécification vers l'implémentation algorithmique. D'autre part, il y a l'opération d'abstraction fonctionnelle, qui consiste à voir le monde au travers du prisme des effets produits sur lui : en quoi l'idéalité agit-elle ? Pour comprendre l'informatique, il faut donc la situer dans une double appartenance. Dans la sphère idéale d'abord, en tant qu'elle est créatrice d'une nouvelle réalité ; dans la sphère de la technique ensuite, en tant qu'ensemble de méthodes répétables, perfectibles, efficaces. L'abstraction ainsi pratiquée par l'informatique n'est pas sans ressemblance avec celle des mathématiques : celles-là également créent de nouvelles entités en s'inspirant du monde sensible. Et pourtant, l'abstraction

[74] Pour une introduction générale à la théorie des quatre causes chez Aristote, voir J. FOLLON, *Réflexions sur la théorie aristotélicienne des quatre causes*. Pour la pertinence de la cause formelle dans le contexte de l'informatique, voir G. CHAZAL, *Le miroir automate*, pp. 23-29.
[75] J. FOLLON, *loc. cit.*, p. 328.

pratiquée par l'informatique est différente : là où les mathématiques pratiquent l'abstraction pour nourrir notre connaissance du réel sous la forme d'une science des rapports et des proportions[76], l'informatique a d'emblée pour but de fonder une idéalité agissante, *efficace*. Or cette efficace n'est pas en premier lieu un savoir sur le monde, mais un moyen d'agir sur lui. Le savoir que pourrait constituer l'abstraction informatique sur le monde fait donc problème. Sans prétendre à mener une discussion de fond de ce problème, nous voudrions néanmoins, dans ce mémoire, nous placer sous l'éclairage dont la SBA pourrait le faire ressortir.

### 2.3.1. Interfaces entre théorie et expérience

Commençons par évoquer très brièvement la notion de théorie : élaborée ou très sommaire, la théorie précise quels observables peuvent être étudiés ; elle dit en général aussi comment cela peut être fait. Bref, la théorie nous dit quelles données peuvent être recueillies du système de référence qu'elle pose. La théorie est donc *a priori* : en prenant l'exemple de la psychologie cognitive, le scientifique ne peut collecter des données expérimentales – le temps de réponse moyen à certaines tâches, comme la reconnaissance lexicale – que grâce à une théorie sous-jacente, qui donne à « voir » la réalité psychique comme un traitement de l'information, un « processus » mental de calcul, qui en tant que tel prend du temps[77]. De même, parler d'un système de référence comme étant composé d'« agents » qui s'efforcent d'atteindre des objectifs, c'est déjà imposer à la réalité un cadre théorique lourd d'histoire et riche en conséquences. Sans théorie préalable, en gros, pas de données, ni expérimentation, ni à proprement parler système de référence[78].

Pour reprendre le même exemple, une fois que le psychologue cognitiviste se trouve en possession d'un ensemble de temps de réponses conséquent, il doit faire sens de ces données : comment se rapportent-elles les unes aux autres ? Afin de mettre de l'ordre dans ses données, le scientifique sera amené à élaborer un *modèle*, notion qui est définie comme la représentation (simplifiée) d'un système de référence afin de pouvoir répondre à une question le concernant. Selon la discipline scientifique, plusieurs techniques de modélisation sont possibles. Pour un système de référence, pour une question, une multitude de modèles sont donc envisageables. Aussi le système de référence est-il clairement antérieur au modèle : car, avant de disposer d'un modèle, le scientifique peut déjà mesurer la réalité, voire la soumettre à expérimentation.

Le modèle est donc clairement distinct de la théorie ; se pose alors la question du rapport entre eux. Prenons pour exemple de modèle une maquette d'architecte. Nous en convenons, il s'agit d'un cas limite, mais il devrait suffire pour faire ressortir une différence clef entre théorie et modèle : alors que la théorie, par définition, est de l'ordre du langage, constitue un discours sur le monde, le modèle est libre d'adopter une forme non-langagière. En d'autres termes, et même si tous les axiomes de la

[76] Nous empruntons cette définition de l'activité mathématique à R. LAVENDHOMME, *Introduction à la théorie des catégories*, pp. 493-501.
[77] D. ANDLER, *Processus cognitifs*, dans ID., A. FAGOT-LARGEAULT et B. SAINT-SERNIN, *Philosophie des sciences I*, pp. 271-296.
[78] Les termes du débat – système de référence, théorie, modèle – ont été empruntés à J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 1-15.

théorie doivent être valides dans le modèle, ce dernier est une *production originale* pour résoudre un problème. En tant que telle, il n'est pas réductible au langage[79].

Les rapports entre une théorie et ses modèles exhibent ainsi une tension, un enjeu épistémologique, et les tentatives de réduire le modèle au déjà-connu ne manquent pas, alors qu'il constitue pourtant une pratique scientifique nouvelle. Parmi ces tentatives, la plus répandue est sans doute celle connue sous le nom de *logicisme* : elle réduit le modèle à une formalisation mathématique. Selon Frank Varenne – l'auteur que nous suivons en la matière – des tentatives ultérieures qui visaient pourtant à liquider l'héritage logiciste n'ont pas réussi à mettre en cause le primat du langage ; il les qualifie de « linguisticistes ».

Afin de faire droit à la simulation – donc au modèle comme nous le verrons tout à l'heure – en tant que production nouvelle, Varenne la positionne en tant que troisième terme entre *théorie* et *expérience*. Se situant entre l'expérience réelle et la raison théorique, la simulation introduit un degré de liberté qui permet de penser, de nommer des pratiques nouvelles qui ne se laissent pas décrire par la dichotomie entre conceptions rationaliste et matérialiste des modèles. Si nous nommons « computationalisme »[80] la troisième attitude envers les modèles, deux pratiques nouvelles deviennent intelligibles. Ainsi un computationalisme tirant sur le rationalisme s'intéressera aux recherches de théories *calculables* ; un computationalisme se conjuguant au matérialisme se donnera pour but de constituer des *objets substitutifs*, d'illustrer la résistance des choses en réalisant des individualités virtuelles reproductibles sur ordinateur.

La simulation est donc davantage chose que signe, plus *idole qu'icône*, au sens de la sémiotique de Peirce. L'auteur est-il pour autant hostile à la (possibilité de) formalisation des savoirs modélisables par simulation ? Non pas, même lorsqu'il dénonce la formule galiléenne selon laquelle *la nature est un livre écrit en langage mathématique* comme « une profession de foi en faveur de tout type de mono-formalisation »[81]. Tout au mieux, la nature est une bibliothèque de livres : la connaissance que nous pouvons avoir d'elle doit être pluri-formalisable, c'est-à-dire ne pas fonctionner dans un seul registre axiomatique, faute de quoi la réduction de la simulation est immédiate.

Conçue comme une nouvelle source de connaissance, la simulation nous fait voir, après coup, beaucoup de modèles et théories mathématiques comme des mono-formalisations[82]. La grande force de la simulation, dès lors, est sa capacité d'*intégration* des différents formalismes. Or il convient ici, avant tout autre chose, de faire observer que la notion d'intégration peut recouvrir deux pratiques assez différentes : d'abord, dans un premier sens, l'intégration pluri-formaliste renvoie aux architectures dites hybrides. Les auteurs de l'article[83] sur lequel nous nous basons ici font explicitement référence ici aux architectures SOAR et ACT-R, mais l'architecture LIDA que nous avons

[79] En réalité, la différence entre modèle et théorie est confuse chez les auteurs précités. C'est pourquoi nous nous référons sur ce point à Fr. VARENNE, *Les notions de métaphore et d'analogie dans les épistémologies des modèles et des simulations*, pp. 26 et suivantes. Notons par ailleurs que l'équivalence entre théorie et langage remonte au moins à Henri Poincaré (cf. J. SCHLANGER, *La pensée inventive*, dans I. STENGERS et EAD., *Les conceptions scientifiques*, pp. 70-71).
[80] C'est l'orthographe retenue par l'auteur.
[81] Fr. VARENNE, *op. cit.*, p. 63.
[82] À commencer par la théorie des jeux, dont nous parlerons plus loin.
[83] Voir l'article de T. BOSSE, A. SHARPANSKYKH et J. TREUR, *Integrating Agent Models and Dynamical Systems*.

brièvement commentée au premier chapitre (§ 1.7.2.9) tombe dans la même catégorie d'approches hybrides. Dans ce type d'architecture, le sous-symbolique contrôle le symbolique, ou plutôt, le symbolique « émerge » du sous-symbolique, qui en donne les conditions de possibilité. En d'autres termes, le sous-symbolique mène la danse, dans la mesure où les processus symboliques sont contrôlés par des modules inférieurs, dont les traitements parallèles se laissent résumer à des équations mathématiques.

Cependant, l'intégration peut recevoir une interprétation plus forte. Pour y voir plus clair, comme d'habitude aidons-nous d'un formalisme, en l'occurrence LEADSTO, un langage logique temporel visant à modéliser des dynamiques multi-agents où il est possible d'utiliser, dans une même formule, des expressions quantitatives (telles des équations différentielles) et des expressions symboliques ou qualitatives. Ici comme ailleurs, il convient de se poser la question de ce que le formalisme fait ressortir, quel aspect de la réalité décrite est ainsi mis en relief. En l'occurrence, et contrairement à ce qu'une approche très symbolique comme le BDI pourrait faire croire, l'idée qui s'en dégage est qu'une approche agent permet d'intégrer des modélisations *continues*, dont en premier lieu – toutefois sans s'y limiter – le temps. Ces modélisations continues, en outre, ne doivent pas nécessairement être conçues comme étant plus fondamentales que les autres types de formalisations, qu'elles soient discrètes ou même seulement qualitatives. En définitive, dans le premier cas, l'intégration est verticale ou hiérarchique ; dans le deuxième, horizontale.

Nous aurons à revenir aux implications épistémologiques de la SBA pensée comme « objet » de connaissance. Arrêtons-nous cependant un moment sur le mode d'objectivité d'une telle simulation : comment pouvons-nous donner un sens à l'affirmation selon laquelle une idéalité est un objet ? Nous pouvons apporter un élément de réponse en rapprochant la simulation de l'opposition, bien connue en philosophie des techniques, entre outil et instrument. Alors que l'outil augmente la force ou l'agilité naturelle de son porteur, l'instrument augmente la perception de l'homme sur le monde au-delà de ce que ses sens peuvent lui apporter. La SBA serait ainsi un instrument qui fait voir le monde d'une façon épurée, selon les lignes claires des perspectives qu'elle intègre, ou plus exactement, un *outil générique* qui permet de *construire*, pour chaque situation qui s'y prête, un tel instrument *à la demande*[84].

Considérer la SBA comme un instrument la place dans une longue tradition d'échanges fertiles entre méthodes techniques et sciences. L'élaboration de modèles spécifiques est affaire de méthodes techniques, non de science, qui est ici consommatrice des instruments, des possibilités décuplées d'observation créées par la technique, au point de créer des champs entiers d'exploration scientifique. L'exemple des statistiques est à cet égard parlant[85]. À l'origine, il s'agit d'une technique utilisée en physique pour accroître la fidélité des mesures. Alphonse Quételet va reprendre cette technique à son compte pour l'appliquer aux phénomènes sociaux. Ainsi naît la sociologie scientifique moderne, avec des répercussions considérables. Elle permet d'envisager la société

[84] Puisque nous avons déjà eu l'occasion de parler – métaphoriquement au moins – des fourmis, rappelons que la myrmécologie scientifique n'est née qu'à partir du moment où l'on a eu l'idée de se servir de *nids artificiels*, rendant ainsi possible une observation sérieuse d'une colonie de fourmis (M. MAETERLINCK, *La vie des fourmis*, p. 8, 66) ; un environnement artificiel donc afin de mieux contrôler l'observation. La SBA fournit une informatisation de cette même idée, tout en la radicalisant, dans la mesure où les fourmis de l'exemple seraient, elles aussi, artificielles.

[85] Exemple dû à H. BERSINI, *Quételet, l'invention de l'homme moyen par un homme tout sauf moyen*.

comme une réalité physique, car celle-ci devient observable, exhibant ses propres statistiques, voire ses propres mensurations. Voir la SBA comme instrument, comme interface technique entre l'homme de science et l'empirie qu'il a prise pour objet d'étude, permet de généraliser le propos de Varenne. C'est ce que s'est notamment proposé de faire Éric Guichard[86], lorsqu'il prône l'abandon de la dichotomie entre théorie (ou conceptualisation) et expérience (ou pratique, ou encore empirie) et appelle de ses vœux une vision en triptyque, qui inclut à côté de la théorie et de l'expérience, la méthode (ou la technique).

Il va sans dire que la simulation est une méthode, une technique au sens fort de ce terme. En tant que technique, elle pose ses propres problèmes, dont l'ignorance peut avoir des conséquences néfastes sur les résultats. Le simulateur impose une contrainte structurelle forte : le simulateur exige que les modèles dynamiques à simuler se conforment à son *méta-modèle* : si, comme technique, la SBA fait preuve d'une grande souplesse, pouvant accommoder n'importe quelle représentation du monde, celle-ci doit pourtant s'organiser autour d'entités individualisables dont le potentiel dynamique se joue essentiellement sur leurs relations réciproques, que celles-ci soient quantitatives ou qualitatives, subjectives ou objectives. Le simulateur impose ensuite une contrainte d'accès, soit des paramètres d'entrée permettant de perturber le modèle, posant ainsi la question de la *calibration* du modèle. Faisant miroir aux flux d'entrée, le flux de sortie pose des questions et contraintes similaires.

La SBA, comme modèle, impose donc ses propres exigences, d'ordre technique, qui lui donnent son épaisseur. Cette épaisseur, en retour, permet à la SBA d'échapper à l'intention de ses créateurs, et de faire interface de façon inattendue entre l'homme et le monde. Signalons à ce propos que la simulation peut aussi être considérée comme une anthropotechnique, une technique d'organisation et ce, de plusieurs points de vue. La simulation peut d'abord être utilisée pour visualiser les effets de décisions individuelles à l'échelle collective. Jacques Ellul[87], qui s'exprime en termes certes généraux, restant neutre du point de vue des technologies utilisées, mais dont le raisonnement tient particulièrement bien pour la simulation à base d'agents, a rapproché la puissance de calcul de l'informatique à « l'usine du plan » du philosophe Cornelius Castoriadis. Il s'agit de calculer toutes les combinaisons des moyens possibles afin d'envisager toutes leurs conséquences. Selon Ellul, pour que l'informatique devienne outil d'émancipation, il faut procéder à une automatisation radicale de tous les secteurs d'activité, afin de libérer l'homme le plus possible du travail pénible et aliénant ; il faut aussi abandonner la dichotomie du comment et du quoi, la dichotomie entre moyens et fins, afin de penser en termes de possibles et de souhaitable.

Plus concrètement, la simulation multi-agents a déjà fait ses preuves comme anthropotechnique dans le cas de SELFCORMAS[88]. Cette simulation se voulait un outil d'aide à la décision dans le contexte de conflits villageois entre éleveurs et cultivateurs pour l'accès aux ressources rares – telles l'eau ou

---

[86] É. GUICHARD, *L'internet et l'informatique comme révélateurs*, dans D. PAROCCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*, p. 137.

[87] Voir l'article *Vers la fin du prolétariat ?* dans J. ELLUL, *Pour qui, pour quoi travaillons-nous ?*, plus particulièrement les pages 185 à 213.

[88] J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 82-90. Il s'agit ici d'un exercice de « modélisation d'accompagnement », dont nous étudierons les tenants et aboutissants au troisième chapitre, dans le cas pratique dédié à la prise de décision éclairée par les méthodes de simulation (§ 3.4.2.9).

les terres fertiles – au Sénégal. SELFCORMAS s'inscrivait dans une méthodologie participative, où les principaux intervenants des conflits étaient invités à modéliser eux-mêmes l'espace : dans un premier temps de discussion, les ressources étaient simplement dessinées sur une carte. Les participants détaillaient leurs usages des ressources, leurs critères de satisfaction, et aussi les heuristiques de décision utilisées. Dans un deuxième temps, les participants jouaient leur propre rôle au moyen de papillons autocollants (*post-it*©) symbolisant leurs activités sur une carte quadrillée. À chaque tour, un joueur pouvait placer un papillon d'activité sur une case libre, tout en veillant à la conformité de son « coup » par rapport aux règles précédemment définies. Pendant cette phase, les règles étaient encore sujettes à modification, suppression ou ajout par consensus. Dans un troisième temps seulement, les ressources et règles étaient traduites sur une plate-forme de simulation[89] : différents scénarios, simulant toujours une période d'une année complète, étaient alors testés et présentés aux participants. La méthode a été appliquée sur quatre sites au Sénégal, et a permis d'opérer des déblocages significatifs là où elle a été utilisée.

La simulation, ici, se déploie comme amplificateur d'un jeu, à la limite comme thérapie collective. L'exercice de formalisation du comportement, de mise en scène des acteurs et de leurs formes d'interaction, vaut plus par lui-même que pour son résultat scientifique. Si la simulation s'inscrit ici dans le prolongement d'un jeu, elle en accepte toutes les conditions : comme nous l'avons vu précédemment, le joueur accepte, au moins momentanément, le primat du virtuel sur le réel. Il accepte ses règles et ses contraintes pour faire partie du monde qui lui est proposé. Une telle adhésion au monde pacifié instauré par SELFCORMAS dépasse la pose de l'observateur tout comme celle de l'utilisateur, car elle s'accompagne d'un investissement personnel : le joueur ne s'expose-t-il pas toujours au risque de perdre sa mise ?

### 2.3.2. Émergence de sens ou simulacre ?

Ce qui rend la simulation fascinante, c'est qu'elle recouvre deux choses : une méthode, une technique de connaissance – nous l'avons amplement vu au paragraphe précédent – et un monde nouveau qui peut être observé. À l'instar de Frank Varenne[90], parlons de *simulation* pour le processus, la technique de connaissance, et de *simulat* pour son résultat, un état du monde produit par le processus. Dans le processus de simulation, il faut encore distinguer entre deux niveaux d'activité, deux étapes : d'une part, le déroulement ou l'opération de la simulation et qui donne naissance au simulat ; d'autre part, l'usage impose également la notion de simulation pour l'observation ou l'analyse du simulat. L'observation peut se faire soit par visualisation, si la plate-forme de simulation propose une interface graphique ; elle peut aussi se faire en appliquant des métriques, l'une ou l'autre forme de mesure.

L'idée d'observation est capitale : la simulation à base d'agents ne peut être connue qu'en observant ses résultats, qui constituent une connaissance nouvelle sur le monde créé par elle, sous l'influence

[89] En l'occurrence, la plate-forme Cormas, développée en SmallTalk. Le lecteur intéressé peut se référer au lien suivant pour plus de détails : http://cormas.cirad.fr/fr/demarch/sma.htm.
[90] Cf. Fr. VARENNE, *La reconstruction phénoménologique par simulation*, dans D. PAROCCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*, pp. 108-111.

respective des paramètres d'entrée et des perturbations du système ainsi engendrées. En d'autres termes, la connaissance véhiculée par la simulation porte non pas sur le monde empirique en soi, mais sur le modèle qui en a été extrait. Reprenons à ce propos la discussion sur le modèle comme interface entre science et monde. Soulignons d'abord que cette discussion n'est pas nouvelle, même si une technique comme la SBA permet de la reprendre à nouveaux frais. En effet, lorsque quelqu'un comme Éric Duyckaerts étend la notion d'expérience imaginaire (*Gedankenexperiment*) à l'intelligence artificielle, il la rapproche de l'expérimentation : l'IA est « comme » une technique d'expérience imaginaire à l'usage non plus des philosophes mais des scientifiques ; technique des sciences qui jette un pont entre les mathématiques et le réel, là où une « vraie » expérimentation est trop coûteuse ou impossible à pratiquer[91]. L'auteur, cependant, ne nous dit pas comment ce pont entre mathématiques et réel est jeté. C'est précisément le concept de modèle qui permet de le faire : la simulation est au modèle ce que l'expérimentation est au système « réel ». Notons, cependant, que ceci ne revient pas à dire, tant s'en faut, que la simulation est une expérimentation au sens propre.

Faut-il comprendre que la simulation est « comme » une expérimentation applicable à tout type de modèle ? Non pas, la simulation n'est expérimentation informatique que sur un modèle *dynamique*, c'est-à-dire un modèle qui fait intervenir le *temps*[92]. Hâtons-nous d'écarter une confusion possible : la simulation, en tant que procédé technique, prend un certain temps de calcul. Le système de référence, en tant que réalité du monde, se déploie également au cours d'un certain temps : c'est le temps réel du système cible. Il s'agit ici encore d'un troisième temps, celui interne à la simulation. Si la simulation est un mimétisme dynamique, il convient de comprendre qu'elle peut l'être de deux façons : elle peut être un mimétisme final *par* dynamique, c'est-à-dire qu'elle peut obtenir un résultat « fidèle » à l'égard du réel au terme d'une série de calculs intermédiaires qui, eux, ne le sont pas. La simulation peut cependant aussi être mimétisme *de* dynamique, c'est-à-dire se déployer au travers d'étapes intermédiaires qui, elles aussi, tendent à mimer les étapes intermédiaires du référent empirique.

Cette remarque peut même recevoir une signification élargie dans le cas de la SBA : le temps y sera en effet mimétique quant à l'évolution vers le résultat final agrégé, et mimétique encore au niveau des interactions entre agents nécessaires à l'émergence de l'état stable. Cependant, à une certaine échelle de temps, le mimétisme s'arrête. Si, par exemple, un agent « reconnaît » un semblable au moyen d'une entrée dans une table de hachage, il est évident que les opérations de calcul pour retrouver l'entrée grâce à sa clef n'auront rien de mimétique par rapport à celles mises en œuvre par un cerveau d'un agent vivant à partir des traits de visage d'un interlocuteur. Aussi l'opposition mimétisme de dynamique et mimétisme par dynamique doit-elle être comprise comme une affaire de degrés, et non de façon dichotomique.

Par cette distinction entre deux types de mimétismes, nous retrouvons un enjeu majeur de notre premier chapitre, à savoir la justification de l'acte éthique, qui doit être exprimable en langage humain. À tous les niveaux où le mimétisme *de* dynamique est en vigueur, la métaphore peut pour

[91] Voir É. DUYCKAERTS, *Expérience imaginaire et Intelligence Artificielle*, pp. 57 et suivantes.
[92] Cf. Fr. VARENNE, *La reconstruction phénoménologique par simulation*, dans D. PAROCCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*, pp. 109-110.

ainsi dire être *filée*, aboutir à une explication certes imagée mais recevable, dans la mesure où les comportements des individus simulés peuvent être interprétés en termes d'une prise de décision informée par une valeur. C'est là une piste capitale, qu'il faudra suivre le plus loin possible au troisième chapitre !

Cette distinction entre mimétismes ouvre cependant sur un problème difficile : la simulation n'est-elle, après tout, qu'un jeu de dupes ? C'est la critique énoncée par Daniel Parocchia en se basant sur Eliza, ce programme thérapeutique que nous avons déjà rencontré au premier chapitre (§ 1.7.2.7) :

> *[…] on a simulé un processus de façon assez ingénieuse, mais ce processus court-circuite le niveau du sens et est évidemment sans rapport avec les comportements effectifs et multiplement motivés qu'on observe dans la réalité, auxquels ces simulacres, en effet, extérieurement ressemblent.*[93]

L'argument doit être pris au sérieux : il soulève la question de la *validité externe* du modèle : quelles sont les propriétés qui peuvent être valablement étudiées à partir du modèle utilisé ? La question du simulacre ouvre sur une perspective plus large, celle de la question de l'articulation entre approche fonctionnaliste et approche par simulation (multi-agents ou autre) en éthique. En lui faisant imiter un comportement, nous demandons également à la SBA de fournir une justification. Or il est clair que les niveaux éthiquement pertinents ne sont pas les seuls à pouvoir être étudiés fonctionnellement, et que d'autres regards, d'autres savoirs ou disciplines, peuvent avoir d'autres boîtes noires, où le mimétisme de dynamique s'applique.

Reprenons l'exemple de l'identification des agents entre eux que nous avons développé lorsque nous avons filé la métaphore mentaliste : l'identification peut jouer un rôle important, sans que – hypothèse fonctionnelle – la façon d'identifier soit simulée de façon réaliste. Le mimétisme de dynamique peut être vu comme une boîte noire fonctionnelle ; elle n'est valide (fonctionnellement) que dans la mesure où elle ne constitue pas le niveau d'analyse qui nous intéresse. Ce qui nous intéresse, en revanche, ce sont les interactions et relations entre agents, qui dépassent les détails de l'implémentation. Le mécanisme d'identification devient alors une question d'implémentation dont l'intérêt n'est toutefois pas nié : nous ne pouvons en faire fi que dans la mesure où, par construction théorique, l'implémentation n'a pas d'incidence sur les propriétés étudiées du système. Sur ce point, les pratiques qui prévalent en SBA se distinguent nettement des discussions en éthique des machines, dont nous avons vu au premier chapitre qu'elle a tendance à se concentrer sur les implémentations psychologiques de phénomènes comportementaux, lesquels se définissent pourtant plus aisément de façon relationnelle qu'atomique.

Nous demandons donc à la métaphore de capter une couche de la complexité : seulement à ces conditions, les implémentations non mimétiques sont permises. En d'autres termes, pour protéger le statut épistémologique du mimétisme de dynamique face au mimétisme par dynamique, il faut prendre au sérieux l'idée fonctionnaliste de *modularité*[94]. La modularité repose sur l'idée d'une

[93] D. PAROCCHIA, *Le status épistémologique de la « vie artificielle »*, dans Fr. TINLAND, *Ordre biologique ordre technologique*, p. 182.
[94] La présentation qui suit se fonde sur H. A. SIMON, *Les sciences de l'artificiel*, pp. 32-48.

distinction entre environnements interne et externe. La caractérisation d'un système et son comportement peut s'appuyer sur l'interface entre les environnements. L'interface est source d'abstraction et de simplification : souvent, quelques hypothèses tout à fait minimales suffisent pour modéliser ses exigences à poser vis-à-vis de l'environnement interne. L'environnement externe est seul requis dans une explication de type fonctionnel, car c'est en lui que se manifestent les buts de l'environnement interne. La ressemblance entre l'artificiel et le naturel est rendue envisageable parce que les mêmes « buts » peuvent leur être assignés. Et Simon de citer à ce propos l'exemple d'une lune artificielle en orbite : elle doit obéir aux mêmes lois du mouvements qui régissent le mouvement des astres naturels.

La modularité repose sur une définition pour ainsi dire récursive de la distinction entre ces deux environnements : chaque système peut être décomposé en sous-système, et l'environnement interne à chacun peut à son tour être défini par la description des fonctions de chacun, exactement de la même manière que l'environnement interne du système englobant peut être défini en décrivant ses fonctions, sans spécification détaillée des mécanismes qui ressortissent des sous-systèmes. Cette façon de voir, reposant sur l'abstraction des détails des phénomènes, part évidemment de l'hypothèse que les systèmes sont décomposables dans des couches aux dépendances tenues entre elles. Cette exigence ne va pas de soi : ainsi dans la compréhension du langage naturel, il est bien connu qu'un locuteur « entend » mieux les phonèmes si ceux-sont placés dans des mots réels que dans des pseudo-mots, il comprend mieux les mots placés dans une phrase grammaticale que ceux placés dans une suite aléatoire… C'est dire que, dans ce traitement, les niveaux inférieurs (comme le traitement phonologique) sont directement influencés par les niveaux supérieurs (morphologique et syntaxique). Ces phénomènes reçoivent une formalisation élégante à l'aide de réseaux neuronaux avec rétropropagation. Il n'est cependant pas sûr que les différentes couches d'un tel réseau puissent être *interprétées* comme des modules correspondant à la hiérarchie des niveaux d'analyse linguistique.

Or ce n'est pas l'inconvénient majeur de la thèse de la modularité, car même un auteur profondément fonctionnaliste comme Herbert Simon doit bien reconnaitre qu'il y a *plusieurs façons de décomposition en modules possibles*[95] : comment dès lors choisir celle qui correspond le mieux à la réalité, éthique ou autre ? Sur cette question, force est de constater que l'approche fonctionnaliste reste muette et il faudra donc se tourner vers d'autres sources. La problématique que nous abordons ici a un nom : c'est celle de l'émergence. Étant d'abord une notion commune, l'émergence soulève la question de la nouveauté : les sciences qui font appel à elle sont celles qui ont intégré une composante historique, qui s'intéressent au devenir des choses. Elle est symétrique et inverse de la notion de réduction :

> […] *un « réductionniste » tiendra que, si l'on admet en première approximation que les phénomènes se distribuent en « plans » ou « niveaux » hiérarchisés de « réalité », de fait chaque niveau se « réduit » au niveau inférieur : le psychique se réduit au physiologique, le physiologique au physico-chimique. Pour peu qu'on poursuive le jeu de la réduction sur le ton d'un positivisme idéaliste : le physico-chimique se réduit à du quantique, le quantique se réduit*

[95] H. A. SIMON, *op. cit.*, pp. 230-231.

*à une représentation collective des physiciens, cette représentation collective (par la thèse de l'individualisme méthodologique) se réduit aux représentations individuelles des chercheurs en physique fondamentale, ces représentations individuelles se réduisent à des états cérébraux, et voilà bouclé le cercle de la réduction…*[96]

La citation qui précède donne le ton : le débat prend volontiers des accents polémiques. Avec la notion d'émergence, nous nous avançons en effet sur un terrain miné[97]. Notre seule ambition, dans ce contexte qui nous éloigne de notre problème central, est d'éclairer une unique question : qu'est-ce que la simulation informatique permet de connaître, que ce soit sur l'homme ou sur l'éthique ? Même si la notion est avant tout le lieu d'une polémique, il faut que nous en parlions, car la SBA se pose volontiers comme une « illustration » du concept.

Traditionnellement, la notion d'émergence a partie liée avec l'holisme. En sciences sociales, par exemple, Émile Durkheim était convaincu que la société (le tout) était plus que la somme de ses parties (les individus la composant) [98]. À l'inverse, un réductionniste tiendra que la société n'est rien d'autre que ces mêmes individus. Posée ainsi, de façon caricaturale certes, nous pouvons cependant déjà tirer un renseignement capital sur la manière dont l'émergence est souvent comprise : il s'agit, en effet, d'une notion qui cherche à dire quelque chose d'ordre ontologique : quelle est *l'essence* des phénomènes observés ? Or les deux points de vue – holisme et individualisme réductionniste – doivent être critiqués dans leur présupposition *statique* : ce que l'observateur observe, c'est un *comportement*. Quand on dit que le tout « est » plus que la somme de ses membres, il faudrait y substituer un regard plus dynamique : le tout « fait » plus – autre chose – que la somme de ses membres. La véritable nature d'un comportement étant dynamique, il faut – pour utiliser la terminologie de Stengers[99] – refuser sa réduction à un état.

À partir de là, beaucoup de possibilités restent cependant ouvertes. Hugues Bersini, dans le contexte de la SBA, défend quant à lui l'idée d'une émergence qualifiée de faible : l'émergence ne prend pas parti sur le plan ontologique mais se caractérise par son effet, *cognitif*, sur l'observateur humain, de *surprise*, surprise qui surgit par le changement de vue entre les niveaux micro et macro[100]. Pour cet auteur, en effet, l'émergence se lit à deux niveaux : le premier niveau est celui où s'observent les individus et leurs interactions. L'observateur macro, en revanche, adopte une vue hélicoptère : il y a émergence non à cause d'un ensemble de relations entre agents, mais en vertu de la différence de regard entre ces observateurs. L'un, l'observateur micro, connaît par le menu et par le plus infime détail un individu isolé, ainsi que les interactions de celui-ci. L'autre, l'observateur macro, a quant à lui besoin d'un *nouveau domaine sémantique* pour décrire ce qu'il voit. Ce domaine est tout à fait

---

[96] D. ANDLER, A. FAGOT-LARGEAULT et B. SAINT-SERNIN, *Philosophie des sciences II*, p. 947.

[97] Nous empruntons ces quelques mots de présentation de l'émergence au chapitre qu'y consacre Anne FAGOT-LARGEAULT, dans *op. cit.*, pp. 939-1048 ; au petit livre de Hugues BERSINI, *Qu'est-ce que l'émergence ?* ; finalement, au fascicule d'Isabelle STENGERS *La vie et l'artifice : visages de l'émergence* dans *Cosmopolitiques II*, pp. 193-284.

[98] Exemple dû à H. BERSINI, *Quételet, l'invention de l'homme moyen par un homme tout sauf moyen*.

[99] I. STENGERS, *op. cit.*, p. 201.

[100] Si nous tenons compte de la position de Bersini, qui est de dire que la statistique a permis au fait social de s'émanciper, d'être l'instrument de perception qui a permis au social de devenir un observable scientifique, *d'exister* en quelque sorte, nous pouvons aisément comprendre en quoi la SBA contribue ainsi, en sciences humaines, à autonomiser davantage encore le fait social. Dans la terminologie propre à la philosophie des techniques vue plus haut, la statistique et la SBA sont des instruments qui « augmentent » notre perception du réel.

indépendant du domaine sémantique micro. S'y ajoute cependant une dernière exigence : ce vocabulaire nouveau adopté par l'observateur macro doit être *nécessaire*, dans la mesure où celui-ci permet de décrire des fonctions que l'observateur micro n'est pas capable de comprendre, car il ne peut voir l'environnement qui en fait usage[101]. Qui dit domaine sémantique nouveau, peut dire aussi base axiomatique nouvelle : nous retrouvons ici l'idée de pluri-formalisation de Frank Varenne.

Toujours est-il que cette dichotomie entre niveaux micro et macro convient bien davantage à la statistique qu'à la SBA. Celle-ci, en effet, place l'interaction au centre de ses préoccupations. Du point de vue de la dichotomie micro-macro cependant, l'interaction reste au niveau micro. Ceci revient à dire que l'émergence définie en ces termes reste muette sur la spécificité de la SBA qui est de distinguer non pas deux mais trois niveaux : intra, inter et supra-agents. C'est là où le point de vue défendu par Stengers prend toute sa pertinence. Stengers rappelle que la question de la finalité – s'il y a organisation, c'est en vue d'une fin – a historiquement toujours été l'étendard de l'émergentisme. La finalité se trouve, dès lors, au cœur de l'émergence. Or les moyens mis en œuvre pour obtenir l'effet final comptent tout autant pour comprendre en quoi il y a émergence. C'est pourquoi elle transforme la vision classique dont Bersini reste tributaire et qui est duale : le tout (premier niveau) est organisé en vue d'une fin (deuxième niveau). Stengers y substitue une vision à trois niveaux : le niveau supérieur – où se situe la fin de l'organisation, le niveau central de l'organisation, enfin le niveau inférieur, qui est celui des moyens mis en œuvre pour obtenir l'organisation.

Le regard, ainsi, se déplace d'un intérêt exclusif pour la finalité de l'organisation vers la relation entre moyens et fins. Cette relation n'est pas donnée d'avance, elle est enjeu et problème. Une première forme de relation possible est celle qui fige la fin. C'est notamment ce qui se passe en sociobiologie : cette discipline, qui voit la main de la sélection naturelle partout et pour peu expliquerait par elle indifféremment les ailes des oiseaux, la qualité solvante de l'eau ou le régime alimentaire du panda, finit par inverser les rapports fins-moyens. L'émergence absolue, qui ramène tout à une seule fin, a pour conséquence d'abolir la finalité.

Une autre forme de relation est donnée par la démarche de l'ingénieur lorsqu'il maintient les artéfacts qu'il crée dans une double indétermination : une même organisation peut être construite selon des moyens fort différents les uns des autres ; une même organisation peut servir des fins différentes selon l'environnement dans lequel elle trouve à être exploitée. La seule obligation que se donne l'ingénieur est celle de l'efficacité et de la performance. Cet impératif mis à part, le domaine du concepteur est la *fiction*. Même si la radicalité du propos est discutable[102], nous retrouvons bien là l'aporie de l'approche fonctionnaliste signalée plus tôt.

---

[101] Le lecteur aura compris le souci pédagogique appuyé de Bersini. En réalité, expliquer l'émergence en termes de « surprise » soulève la question du *qui ? qui est surpris ?* (R. AXELROD, *The Complexity of Cooperation*, pp. 3-4), question qui ne peut être satisfaite par un renvoi à des observateurs abstraits. Poser le problème ainsi ne fait au final que le déplacer. Quant aux exigences liées au « domaine sémantique », elles s'en sortent certes un peu mieux mais prêtent le flanc à une critique postmoderniste, qui ne veut voir dans l'entreprise scientifique qu'une construction discursive sans reste.

[102] Le propos est discutable dans la mesure où la technique (ou la technoscience) produit des effets répétables dans le réel et pose donc ses propres exigences. Or tant Stengers que Bersini ne conçoivent l'artéfactuel que sur un mode artisanal, pour reprendre la terminologie de Simondon : l'artéfact est vu comme système ouvert d'exigences, dont le fonctionnement interne est de l'ordre du contingent.

La forme de relation suivante est celle qui retiendra le plus l'attention de la philosophe. C'est celle que cherchent à instaurer les sciences dites « du terrain » : lorsqu'un praticien d'une telle science (géologue, climatologue, spéléologue, etc.) cherche à expliquer la régularité d'un phénomène, il le fait primordialement sur le mode de la description : la régularité – exigence qu'il adresse à un phénomène donné – devient fin dans la description que le scientifique en donne. Les moyens de mise en œuvre sont ici non pas des « causes » – à la manière où les sciences de laboratoire entendent cette notion – mais des conditions de possibilité, des raisons nécessaires mais non-suffisantes. La cohérence de la démarche d'une science du terrain ne provient ainsi pas d'une démarche déductive[103], mais d'un agencement en narration, à la manière d'une *intrigue*, afin de raconter un récit où il n'est pas question de causes mais de contributions.

Il s'ensuit que l'émergence doit toujours être pensée comme un problème, comme une négociation, dont le modèle devra témoigner : sa pertinence se mesure à sa capacité de rendre intéressant le lien qui lie la situation modélisée à la question de l'universel, lien *a priori* indéterminé, à découvrir. Le modèle pose ainsi la question du comment de l'émergence :

> *Un modèle, tel qu'il fonctionne dans les sciences théorico-expérimentales, a un domaine de validité strictement limité car il exploite, dans ses définitions, des expédients simplificateurs dont la portée est explicitement relative à ce domaine. […] Dès qu'il est question de « sciences du terrain », en revanche, le modèle ne se définit plus par contraste avec une théorie. Le modèle ne se définit plus par ses simplifications, ou par des hypothèses* ad hoc*. Il ne correspond plus à une pratique dont l'enjeu est de « prouver » – puisque la validité d'une quelconque preuve ne vaudra de toute façon que pour « tel cas ». Il s'agit plutôt de mettre en tension problématique ce que requiert le modèle et ce qu'apprend le terrain. Un modèle, en désignant ses requisits, fait un pari et prend un risque : ce qu'il requiert de la réalité est nécessaire et suffisant pour « raconter » ce qu'il ambitionne de mettre en scène.*[104]

Et Stengers de donner l'exemple d'une colonie de fourmis. Quand le chercheur se déprend d'une vision robotique de la fourmi – vision qui préside aux approches économistes cherchant à y voir un optimum – et qu'il fait droit au comportement erratique des fourmis individuelles, il peut accéder à des questions nouvelles concernant le *comment* des interactions entre fourmis. En cherchant à comprendre les comportements individuels, modulés (non déterminés) par leurs interactions, il peut par exemple prendre acte de l'ensemble de stratégies adaptatives mises en œuvre par les différentes sous-espèces de fourmis et raconter l'histoire d'un ensemble de « choix », de stratégies collectives en réponse à des situations et des environnements réels. En définitive, lorsque nous contemplons l'exemple des fourmis, Stengers semble nous inviter à voir dans la SBA la possibilité d'une mise en intrigue des réquisits du modèle à simuler. La SBA permet alors d'explorer un problème de mise en rapport de moyens et fins au travers d'une organisation.

---

[103] Notons à ce propos que la SBA est parfois qualifiée de « troisième voie » entre déduction et induction (par exemple, dans R. AXELROD, *The Complexity of Cooperation*, pp. 3-4). Or s'il est clair que la SBA génère des données qu'il est possible d'étudier inductivement, rien dans cette génération n'oblige à une démarche déductive. Voir aussi la thèse défendue par Frank Varenne dont nous faisons état plus loin dans cette section.

[104] I. STENGERS, *Cosmopolitiques II*, pp. 255-256.

Stengers ne fait qu'effleurer le thème de la mise en intrigue. Cependant, il y a un auteur qui a longuement étudié le rapport entre cette dernière et l'activité scientifique, en l'occurrence l'historiographie : c'est Paul Ricœur. Même si sa théorie nous fait sortir allègrement du cadre du présent mémoire, rappelons en deux mots que dans l'historiographie, la compréhension de la mise en en intrigue – définie en termes aristotéliciens comme *l'un-par-l'autre* à travers *l'un-après-l'autre* – rencontre l'explication par des lois, par l'universel[105]. Cette rencontre prend la forme d'une *imputation causale singulière* : telle situation a pu se produire à la faveur de telle combinaisons de facteurs explicables en termes de régularités. Afin de tester le bien-fondé de ses hypothèses, l'historien peut faire appel à « la construction par l'imagination à des cours différents d'évènements, puis dans la pesée des conséquences probables de cet évènement irréel, enfin dans la comparaison de ces conséquences avec le cours réel des évènements »[106]. En d'autres termes, l'historiographie fonctionne comme un *intégrateur*, sous l'angle unificateur de l'espace et du temps, de savoirs venus d'ailleurs, de disciplines tierces. Dans les sciences du terrain, donc, le « terrain » doit être compris autour de ces deux axes, l'espace et le temps, comme une configuration réalisée au terme d'un devenir singulier.

La SBA doit donc, afin de tenir son rôle d'intégrateur, remplir plusieurs conditions. Elle doit être située sur un terrain, c'est-à-dire dans l'espace et le temps, pour que les divers mécanismes à intégrer puissent s'y superposer comme dans un cadre corporel, au moins sémantiquement. Elle doit poser les interactions de ses agents comme réquisit d'une organisation. Elle doit faire exister un milieu propice à l'organisation qu'elle décrit. La métaphore qu'elle véhicule doit avoir pour fonction première de préserver le finalisme qui fait problème pour le scientifique. Une bonne SBA ouvre le problème, elle met en question l'articulation entre niveaux de description, plutôt qu'une volonté de réduire un niveau à un autre. Cette articulation n'est pas figée mais doit être renégociable, sinon renégociée, à chaque nouvel effort de modélisation. Nous retrouvons d'ailleurs un rôle dévolu à la théorie, pour peu que nous nous en tenions à la définition assez large que nous en avons donnée en début de ce paragraphe, à savoir toute construction de nature essentiellement discursive qui précède ou accompagne le modèle.

Nous espérons avoir fait sentir en quoi la simulation a quelque chose qui lui appartient en propre, et qui ne se laisse qu'imparfaitement qualifier de quasi-expérimentation. Certes, la simulation partage avec cette dernière la caractéristique du contrôle : le réel est un mauvais modèle de lui-même, pour reprendre les mots de Frank Varenne[107], lorsque nous le considérons sous le rapport de l'expérience contrôlée : la simulation s'avère bien plus précise. Or la simulation se démarque fortement de l'expérimentation en ce qu'elle procède non pas d'un effort *d'abstraction* par rapport à la situation singulière, mais d'un effort *d'intégration* de savoirs afin de comprendre le devenir d'une situation

[105] Voir P. RICŒUR, *Temps et récit 1*, en particulier le chapitre *L'intentionnalité historique*. Faisons nôtre l'insistance de Ricœur sur le point suivant : parler de mise en intrigue dans le cas d'une explication scientifique – en l'occurrence, de l'historiographie – relève de l'analogie, d'où son utilisation des termes *quasi-intrigue*, *quasi-évènement*, *quasi-personnage*.

[106] *Ibid.*, p. 324. Rappelons-nous, à cet égard, qu'une simulation peut avoir plusieurs trajectoires ou « vies ». Or de telles trajectoires permettent de donner corps aux raisonnements contrefactuels de l'historien *en les calculant*.

[107] Fr. VARENNE, *Les notions de métaphore et d'analogie dans les épistémologies des modèles et des simulations*, pp. 68-69.

singulière au regard des exigences de l'universel. À ce titre, comme l'a souligné si pertinemment Hugues Bersini, la SBA nous surprend.

À l'aune de sa capacité de nous surprendre, il est vrai que la simulation à base d'agents dépasse de loin les méthodes statistiques traditionnelles. Cette puissance lui vient, notamment, de sa faculté d'intégrer des sous-modèles mathématiques sans les confondre dans un résultat d'emblée agrégé. L'intégration complexifie passablement le problème – esquissé plus haut – des rapports de correspondance entre le modèle et le système de référence[108]. Notamment les questions du rapport du comportement observé par rapport aux différents sous-modèles, ainsi que du rapport entre la manière dont la simulation désagrège les sous-modèles en comportements locaux, les fait interagir, et les réagrège dans la dynamique de son déroulement. Ceci revient à poser les questions du réalisme des sous-modèles, et du réalisme de l'imbrication des sous-modèles[109]. L'auteur conclut en remarquant que l'intégration, ici, doit surtout être comprise comme relevant de la technique : le simulat intègre en effet les sous-modèles de la même manière qu'un objet technique *concret* (au sens de Simondon) peut intégrer ses différents sous-composants, sans toutefois accéder au statut d'un objet symbolique doté d'une cohérence mathématique propre.

La conséquence directe de cette vue – la SBA est un intégrateur de formalismes plutôt qu'un formalisme en soi – permet de mieux comprendre en quoi elle peut être dite fournir des connaissances nouvelles. En effet, faire état de connaissance nouvelle s'accommode mal de l'idée qui voudrait que la simulation participe d'un mouvement déductif : les tenants de cette vue affirment qu'à partir de prémisses connues, la simulation fait voir des conséquences inconnues[110]. Or la déduction est une opération qui n'a de sens *qu'à l'intérieur d'un formalisme donné*. Par ce changement dans le niveau d'appréhension des symboles, l'observation de la simulation s'oppose donc à la déduction, qui reste toujours à un même niveau de symbole. En somme, les résultats qu'elle génère sont à traiter *inductivement*.

Arrêtons-nous un instant sur cette affirmation que la simulation n'a pas de cohérence mathématique qui lui appartiendrait en propre. Précisons-en tout de suite la portée, les implications, le sens. Il ne faudrait surtout pas en conclure que la simulation est rétive à la formalisation, qu'elle constituerait comme un point d'arrêt à la recherche, dans la nature comme dans les sociétés humaines, des relations et proportions – objet d'étude des mathématiques. Selon Frank Varenne, cela veut dire simplement que les mathématiques viennent soit avant la simulation, sous une forme équationnelle et essentiellement locale, soit après la simulation, sous une forme alors statistique.

Commençons par illustrer le point de la localité de la cohérence mathématique par l'exemple de mathématisation – donnée par Treuil et ses collègues[111] – d'un agent de type réactif très simple. À la base de leur mathématisation, nous retrouvons derechef une métaphore, physique cette fois : l'environnement y est interprété comme un ensemble de *champs de forces*, qui se déploie dans un

---

[108] Fr. VARENNE, *La reconstruction phénoménologique par simulation*, dans D. PAROCCHIA et V. TIRLONI, *Formes, systèmes et milieux techniques après Simondon*, pp. 113-114.
[109] Soulever le problème du réalisme des modèles nous fait sortir de notre problématique, pour en aborder une autre, qui est celle du rapport de vérité entre le monde et le modèle. Nous en toucherons un mot dans la section qui suit celle-ci.
[110] H. A. SIMON, *op. cit.*, p. 46.
[111] J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 159-168.

espace sur les points duquel sont situés les agents[112]. Sur chaque point, le champ exerce une influence potentielle, qui ne s'actualise que lorsqu'un agent est présent sur ce point. Si, en simplifiant encore plus que les auteurs l'exemple, nous nous limitons aux mouvements des agents dans l'environnement, l'environnement est constitué de trois champs : un champ de résistance au mouvement $\gamma(x)$, un champ de résistance aléatoire $\sigma(x)$, et enfin un champ de potentiel $V^t(x)$, seul champ à varier en fonction du temps. Les agents quant à eux sont dotés des attributs suivants : une position physique $x$, une vitesse $\upsilon$, une capacité d'influence sur le potentiel $\lambda$ et une sensibilité au potentiel $\alpha$. Avec l'ensemble de ces paramètres, il devient possible d'écrire, à chaque instant et pour chaque point de l'espace, les équations de mouvement de chaque agent.

Il serait donc parfaitement possible de progresser sur la voie de la formalisation à partir d'une simulation à base d'agents, pour arriver, finalement, à évacuer toute notion d'agents, de leurs messages comme de leurs interactions ; en d'autres mots, de *réduire* le modèle multi-agents. L'exemple de réduction ci-dessus appelle cependant plusieurs réserves : premièrement, les auteurs ne réduisent la métaphore de l'agent qu'au prix de l'introduction d'une autre métaphore, celle de la force, dont ils reconnaissent par ailleurs aussi la fécondité dans les sciences humaines en général (sous la forme de forces *vives*, ou *sociales*), car quand bien même la mise en équation fait disparaître la métaphore de l'agent, celle-ci reste nécessaire à la compréhension du sens de l'équation de la force :

> *Dans la vision mécanique, on considère que la bactérie se comporte comme une particule subissant une force la contraignant à se déplacer dans la direction de la plus grande variation de concentration. Il existe évidemment une autre manière de parler de la même équation : c'est de dire que la bactérie recherche une concentration plus favorable et se dirige en conséquence dans la direction qui lui apporte cette meilleure concentration au moindre coût. Prises au sérieux, les deux interprétations manifestent deux conceptions des déterminismes pouvant faire émerger le comportement exprimé par l'équation*[113].

Deuxième réserve, une telle réduction est de toute façon condamnée à demeurer une possibilité *de principe*, car quiconque tente de l'appliquer va au-devant de formidables difficultés, aptes à intimider le plus intrépide des modélisateurs ! Il faut ainsi séparer entre les échelles micro et macro pour pondérer l'influence de chacune, séparer les parts respectives du déterminisme et de l'aléatoire, combiner enfin les processus continus et discrets[114]. À ces difficultés techniques, recensées par les

---

[112] La notion d'environnement étant sous-déterminée, elle semble appeler une métaphore de second ordre pour être efficace, d'où une grande variété dans les soubassements mathématiques. Ici, certes, les modèles sont simples, de type réactif, par quoi il faut entendre qu'ils n'ont recours ni à des messages, ni à des évènements entre agents. Par ailleurs, même si le modèle comporte des messages, ceux-ci peuvent être traités comme des métaphores, voués à disparaître au niveau mathématique et reformulés en termes de changements d'état.

[113] *Ibid.*, p. 190.

[114] *Ibid.*, pp. 256-264. La possibilité d'aller au bout de la formalisation, d'évacuer complètement la notion d'agent, semble ainsi condamnée à rester surtout théorique. Nos auteurs semblent pourtant y tenir, un peu à la manière dont les mathématiciens considèrent la vérification de leurs théorèmes par la logique formelle : en général, ils se contentent de savoir que la possibilité existe, sans que celle-ci à aucun moment n'en vienne à faire partie de leur pratique scientifique courante : « A mathematician is happy to be reassured by a mathematical logician that what [he is] doing can be expressed in some foundation, but the mathematician does not care to work out precisely how » (F. KAMAREDDINE,

auteurs eux-mêmes, il faut ajouter le problème de fond abordé par Stengers : le sens d'une réduction doit être négocié au cas par cas, en conceptualisant dans une théorie quelles fins peuvent émerger à partir de quelle organisation. Troisième réserve, soulignons que le résultat ainsi obtenu reste toujours un modèle dynamique, c'est-à-dire… une simulation !

Quel que soit le mérite intrinsèque de ces efforts de mathématisation, leurs limites sont donc là. En effet, si nous voulons réduire la simulation en tant que telle, il faut resserrer davantage encore les mailles de la formalisation, formaliser davantage encore, étudier le résultat agrégé de la simulation – le simulat – au moyen de différentes *prises de moyenne*[115]. Avant d'en venir, cependant, aux moyennes à proprement parler, posons-nous la question comment décrire le système simulé à un moment *t* donné : nous pouvons nous intéresser aux seules structures spatiales, étudier comment les agents sont répartis dans l'espace ; nous pouvons prendre en compte les valeurs des attributs des agents et considérer ainsi la distribution des états du simulat ; ou finalement, nous pouvons également tenter de décrire l'évolution, la dynamique, de la simulation. S'intéresser aux structures spatiales peut se faire via le recours à des indicateurs scalaires, comme le nombre d'agents par unité de surface dans l'espace ou la moyenne des distances au plus proche voisin. Il est également possible de caractériser la classe de configurations, qui donne des renseignements sur les patrons décelables dans la répartition : ainsi, la répartition peut être, par exemple, totalement aléatoire (classe dite des processus de Cox) ou respecter une distance minimale entre agents (classe dite « hard core »). Lorsque nous nous intéressons à la dynamique du simulat, la « vie » d'un simulat est représentée comme une trajectoire parmi l'ensemble ou le *faisceau* de trajectoires rendues possibles par le modèle soumis à simulation. Le faisceau sera dit « stationnaire » lorsqu'il présente la tendance à un état stable après un certain bout de temps ; le faisceau sera dit « ergodique » lorsqu'il suffit de connaître une seule réalisation pour se faire une idée représentative des autres trajectoires.

Nous pourrions multiplier les exemples de propriétés. L'essentiel, ici, est de faire sentir qu'un simulat ainsi agrégé peut être vu comme un « macro-agent », qui symbolise le système entier, simulé lors d'une de ses vies possibles. C'est à cette occasion d'ailleurs que nous pouvons dire que le temps, à l'instar de l'espace, est un principe intégrateur dans la SBA : le simulat devient lui-même un sujet d'étude, dont les propriétés peuvent être analysées, les différentes réalisations mises en relation. Le simulat devient pour ainsi dire une empirie de second ordre, objet potentiel de mathématisation. En définitive, insistons sur une chose essentielle : cet « au-delà » de la simulation qu'est l'effort de réduction – que ce soit sous forme équationnelle ou probabiliste – ne saurait *remplacer* celle-ci : par rapport au modèle de départ, la perte de pouvoir explicatif, d'expressivité, serait par trop considérable. Il n'en demeure pas moins que l'étude mathématique du simulat est une nécessité, y compris du point de vue éthique : même si les comportements étudiés globalement par voie de moyennes relèvent davantage du niveau d'explication de l'automatisme social que de la prise de décision éthique délibérée, point d'arrêt radical de l'automatisme du comportement et de la pensée, il faut à tout prix surmonter les phénomènes de « boîte noire » observés dans beaucoup de cas

J. WELLS, Chr. ZENGLER et H. BARENDREGT, *Computerising Mathematical Text*, dans J. H. SIEKMANN, *Computational Logic*, p. 346).

[115] Le développement qui suit s'inspire – tout en simplifiant et résumant fortement – de J.-P. TREUIL, A. DROGOUL et J-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 201-224.

d'apprentissage automatique : plus la SBA sera connue mathématiquement, plus sa rigueur nous gardera de la fâcheuse surprise des *mauvais apprentissages* de l'intelligence artificielle.

### 2.3.3. Métaphore, formalisme et efficacité

Rappelons-nous la leçon importante du paragraphe qui précède : la simulation risque toujours de dégénérer en simulacre dès lors que l'expérimentateur court-circuite le sens. C'est là, cependant, que l'aspect métaphorique de la SBA peut avoir un rôle important à jouer, car même dans une démarche proprement scientifique, la pensée métaphorique préside à l'élaboration de nouvelles hypothèses, cherchant des réponses « dans le sens plutôt que dans le fonctionnement »[116]. La métaphore nous sert toujours pour communiquer les résultats de la science. La métaphore connaît ainsi deux moments privilégiés :

- Un moment gnoséologique : c'est le rôle dévolu à la métaphore visée par Stengers lorsqu'elle a qualifié les concepts scientifiques de métaphores « qui ont réussi ». La métaphore ainsi conçue est source d'idée, condition d'intelligibilité, opération de capture.
- Un moment éthique : rappelons-nous la discussion, au premier chapitre (§ 1.8.3.1), sur la négociation des valeurs. En opérant sur le sens, la métaphore opère sur notre subjectivité, qu'on considère ce travail comme « perte d'authenticité » ou non.

Dans les deux cas, la métaphore est vecteur de changement, l'opérateur qui le rend « humainement » possible. Toutefois, analogie n'est pas raison ; son champ d'action peut être limité ; doit l'être même, car si elle prompte à proposer du sens, souvent il est nécessaire de refuser la proposition :

> *La métaphore et l'analogie peuvent être utiles ici et trompeuses là. Tout dépend de la nature des similarités que capture la métaphore, selon qu'elles sont significatives ou superficielles.*[117]

La valeur de la métaphore dépend des similarités qu'elle capte, de leur adéquation comme de leur fécondité. Il faut donc pouvoir « caser » le discours métaphorique, en amont comme en aval du discours proprement scientifique. En définitive, nous retrouvons la métaphore au début (innovation) et à la fin (communication) du processus de la connaissance ; on a beau couper et ne réserver l'étiquette de sciences qu'à ce qui se trouve au milieu, une théorie de la connaissance complète doit en rendre compte, d'autant plus pour nous, qui interrogeons la possibilité de connaître l'homme et ses valeurs à partir d'une technique à base métaphorique. Voilà le problème du présent paragraphe.

Nous voudrions aborder le problème en le plaçant sous le signe des thèses de Jean Ladrière quant à la *dialectique du concept*, qui fait état d'une double réalité. En effet, la dialectique du concept se laisse saisir à la fois comme la manifestation, la révélation d'une réalité enfouie, et comme la constitution – dynamique – de cette même réalité, d'un mouvement. Le mouvement est celui qui part d'une intuition pour tendre vers l'objectivité formelle la plus rigoureuse. L'intuition, en tant que

[116] M. DUBOIS, *La métaphore et l'improbable*, p. 175. Rappelons-nous aussi les thèses de P. RICŒUR qui, dans *La métaphore vive*, voit en la métaphore un mécanisme de création de sens.
[117] H. A. SIMON, *Les sciences de l'artificiel*, p. 305.

donné, est opaque ; le concept en assure la compréhension, grâce à la médiation de l'esprit. Notons que pour Ladrière, le concept a sa réalité propre, distincte de celle l'esprit :

> *[…] si l'on veut bien considérer le concept selon sa réalité propre, dans son statut le plus radical d'objectivation, dans ce moment de son existence où il n'est plus le lieu d'échange entre l'esprit et la réalité et comme la frange le long de laquelle s'accomplit leur mutuelle pénétration, mais où il se constitue en objectivité et se dessine, entre l'esprit et la réalité physique, un espace neutre qui médiatise leur rencontre et devient réalité à son tour, mais réalité intentionnelle à travers laquelle s'annonce toujours l'horizon de l'expérience, duquel il a été puisé et auquel il ne cesse de renvoyer. La vie du concept, c'est ce perpétuel va-et-vient entre l'horizon sur lequel il s'arrache et qui lui prête son contenu – et cette suprême objectivation où il se coupe de son horizon et se vide de son contenu pour se constituer comme forme pure au regard de l'esprit.*[118]

Dans les mathématiques, la vie du concept donne lieu à une polarité entre une tendance formalisante et une tendance problématisante (ou réaliste). La tendance réaliste privilégie le mouvement qui va des problèmes du monde physique vers l'élaboration d'outils mathématiques appropriés : ainsi dans le monde antique, la géométrie a permis de comprendre les problèmes d'arpentage ; aux temps modernes, les questions posées par le mouvement physique se laissent aborder par l'analyse, la vie des collectivités par le calcul des probabilités, et les phénomènes de compétition par la théorie des jeux. La tendance formalisante s'intéresse à l'autre face du concept, la révélation d'une structure disant le vrai, l'universel : le formalisme progresse dans le sens de l'abstraction, par totalisation et par simplification.

La hiérarchie entre formalismes mathématiques, de par leur degré d'abstraction, leur niveau d'élaboration, introduit une dimension temporelle, un aspect de devenir à l'intérieur de chaque théorie. Ce point est capital, la progression des théories selon le mouvement de l'abstraction doit effectivement être comprise *temporellement* :

> *Entre ces différents niveaux [d'abstraction des théories mathématiques], il y a bien entendu, une hiérarchie, qui est d'ailleurs de nature dynamique : il ne s'agit pas simplement d'une superposition, mais d'un mouvement qui fait passer les théories à un état d'abstraction de plus en plus élevée. […] Mouvement qui introduit une dimension temporelle dans les rapports entre théories et un aspect de devenir à l'intérieur de chaque théorie, puisqu'une même théorie peut être reprise à des niveaux différents d'élaboration. […] en même temps qu'il échelonne les niveaux de formalisation selon une direction privilégiée, il maintient entre eux une solidarité efficace, qui empêche de sacrifier l'un quelconque d'entre eux aux prétentions totalitaires d'une rigueur trop géométrique. Car chaque niveau se nourrit de celui qui le précède et en même temps lui fournit comme l'image exemplaire d'une cohérence qui en règle déjà les démarches […]*[119]

---

[118] J. LADRIÈRE, *Mathématiques et formalisme*, p. 476. Notons que tout au long de son article, Ladrière utilise le terme « formalisation » dans un sens plus restreint que nous, qui avons tendance à l'utiliser comme simple synonyme de « mathématisation ».

[119] *Ibid.*, p. 457.

Le développement qu'Alain Badiou a consacré au concept de « modèle » peut nous fournir un exemple parlant de la thèse de Ladrière[120]. Cet auteur se base sur le concept de modèle tel qu'il est pratiqué en logique formelle : le formalisme de celle-ci comporte un ensemble de symboles, quelques formules initiales ou axiomes, puis des règles de formation et de déduction. Si une formule quelconque bien formée (c'est-à-dire obéissant aux règles de formation) peut être obtenue à partir des axiomes en ne se servant que des règles de déduction, cette formule est dite « démontrée », elle devient un théorème de la logique considérée. Or lorsque nous disposons ainsi d'un théorème, nous ne savons rien dire encore de son interprétation. Cette interprétation est obtenue en évaluant le théorème au moyen d'une fonction de correspondance $f$, qui réduit récursivement le théorème à une expression de type « vrai » ou « faux ». La fonction de correspondance repose sur la théorie des ensembles : elle projette les constantes prédicatives du langage formel sur des ensembles, et les constantes individuelles sur des éléments de ces ensembles. Ainsi une formule telle que $P(a)$ sera dite valide (*vraie*) si, et seulement si, $f(a) \in f(P)$. Une propriété importante d'un tel système formel est que les règles de déduction *conservent la vérité* : si nous pouvons établir la vérité des axiomes du système, nous avons la certitude que tous les théorèmes du système seront également vrais. Nous dirons que le domaine d'interprétation, composé de tous les ensembles considérés, est un *modèle* du système formel si tous les axiomes du système sont valides pour ce domaine[121].

Ce qu'il faut retenir de ce détour, c'est une mise en perspective de la façon habituelle de présenter le modèle comme relevant de la *sémantique*, et le formalisme en lui-même comme relevant de la *syntaxe*. Or, selon Badiou, la sémantique n'est rien d'autre qu'une théorie des ensembles intuitive, un métalangage, qui par rapport à la syntaxe se trouve à un stade antérieur de formalisation. Cette antériorité n'est pas à comprendre d'abord comme hiérarchique, même si le mouvement dont elle procède est orienté : elle est avant tout temporelle, *historique*. Insistons sur ce point : dire que la sémantique est antérieure à la syntaxe ne lui confère aucun statut fondationnel particulier, mais revient simplement à prendre acte de la dimension *diachronique* de la science considérée. La sémantique participe à l'histoire de la formalisation de la pensée, elle a permis de donner naissance au domaine d'interprétation en enveloppant mathématiquement – au niveau du métalangage – l'empirie. La logique formelle est elle-même pluri-formelle (au sens de Varenne).

Entre l'empirie et sa première formalisation la correspondance est celle d'un transfert métaphorique. L'histoire du concept de modèle fait donc entrevoir le rôle et le statut de l'efficace métaphorique à laquelle la SBA doit tant. Plutôt que de concevoir une discipline scientifique comme un édifice architecturé, érigé sur un *fondement*, changeons de regard en méditant l'exemple de la montagne, formée par dépôt de *sédiments* successifs, parmi lesquels nous pouvons accommoder l'efficace de la métaphore sous forme d'une couche ou d'une strate. La métaphore fait ainsi légitimement partie de l'élaboration du champ scientifique, sans toutefois que celui-ci soit réduit à celle-là. Une telle perspective accepte la diachronie : elle permet d'accepter que le « calcul mental » soit à l'origine métaphore, avant de devenir concept du champ scientifique concerné, sans que l'un puisse

[120] A. BADIOU, *Le concept de modèle*, pp. 81-137.

[121] En relisant ce paragraphe, nous nous rendons compte qu'il répond à une question laissée largement en chantier dans ce mémoire. En effet, nous avons longuement étudié la *validité externe* de la simulation : quelles propriétés du modèle est-il légitime d'étudier à partir des exécutions de la simulation ? Or la question que pose ici Badiou est celle de la *validité interne* du modèle : de quelle manière le modèle que la simulation exécute peut-il être dit « vrai » ?

prétendre remplacer l'autre : les deux couches de compréhension se superposent, donnent une épaisseur sédimentaire à la discipline qui ainsi n'a même pas à choisir entre sens métaphorique et rigueur scientifique[122].

La SBA, partant de la métaphore qu'elle met en contexte, n'opère pas sur un mode formaliste. En effet, toujours selon Ladrière, une tendance formaliste procède du général au particulier. La SBA, même si elle intègre des formalismes tiers, agit selon la tendance problématisante, pour qui le formalisme est toujours au service d'un champ d'application. Dans la simulation, le pouvoir de représentation prime sur les possibilités de déduction des formalismes. La formalisation doit être comprise comme un *après*, une strate de sédimentation qui se superpose à la simulation comme une couche d'abstraction supplémentaire, strate qui n'aurait pu voir le jour *avant* la modélisation et qui, une fois élaborée, n'a de sens que par rapport à ce qui lui précède. Faute de garder toujours à l'esprit ces couches plus profondes, la formalisation ne saurait pas constituer une avancée, mais perdrait inéluctablement de son intelligibilité[123].

À l'instar de la dialectique du concept dont nous sommes parti, l'image de la sédimentation nous interdit cependant de nous contenter de l'enveloppement initial, il faut progresser sur la voie de la formalisation, tout en reconnaissant que le « progrès », ici, contient nécessairement un aspect historique, temporel, de l'ordre de « l'un-*après*-l'autre », et que la trame qui ainsi se forme restera nécessaire à la compréhension du monde telle que la discipline scientifique considérée la construit[124]. Cela implique aussi que l'importance accordée à la métaphore dans ce paragraphe doit être rigoureusement cadrée. En éthique (comme par ailleurs en politique), l'argumentation raisonnable met en garde contre ses mirages. En science, l'expérimentation et la mesure viennent circonscrire son pouvoir. Dans le cadre très spécifique des disciplines formelles, dont l'informatique théorique et les mathématiques, la formalisation permet de « fixer » le sens, d'éviter les dérives sémantiques. Le formalisme comme tuteur : une fois qu'a germé la métaphore, il importe que sa croissance soit bien droite !

En technique, les questions d'efficacité quant au but recherché, de calculabilité en somme, viendront limiter sa prolifération. Donnons deux exemples afin d'appuyer ce dernier point. Le premier exemple est donné dans une application de simulation et de gestion du trafic aérien aux États-Unis, où il faut

---

[122] Ce paragraphe était déjà écrit quand nous avons pris connaissance des lignes que J. SCHLANGER a consacré au rôle de la métaphore dans le discours scientifique, dans le livre co-écrit avec Isabelle STENGERS (*Les concepts scientifiques*, pp. 83-98). Elle aussi insiste sur le lien entre métaphore et activité scientifique. En tant que la métaphore donne à voir, donne à dire, elle constitue la face verbale de la conceptualisation inventive. Comme le rôle de la métaphore est essentiellement heuristique et non cognitif, celle-ci ne doit pas être problématisée quant à sa valeur de vérité dans un langage considéré comme fond neutre et littéral, mais pose plutôt la question de la *pertinence* : la métaphore doit assurer la *cohésion* du langage intellectuel en ancrant le concept dans la *culture*.

[123] Cf. M. DUBOIS : « […] comprendre qu'il existe différents niveaux de réalités liés à des niveaux d'organisation, à des échelles quantitatives différentes, n'est possible que métaphoriquement » (*La métaphore et l'improbable*, pp. 208-209).

[124] Encore une fois, ce n'est qu'après la rédaction de ce chapitre que nous avons pris connaissance du livre d'I. STENGERS et J. SCHLANGER, *Les concepts scientifiques*. Stengers (pp. 178-183) y cherche à se frayer une voie entre empirisme humien – pour qui le fait est neutre, que le scientifique ne fait que prélever, cueillir, dans le réel – et intellectualisme kantien – pour qui le fait est activement construit, toujours déjà imprégné de théorie, et le scientifique s'érige en juge. Elle adopte une vue du fait comme *indice*, qui exige que le scientifique *déchiffre l'histoire qu'il raconte*. Nous retrouvons ici l'histoire, mais inscrite du côté du réel plutôt que du côté du concept.

gérer le nombre impressionnant de 40 000 vols par jour[125]. Une telle gestion doit pouvoir s'adapter de façon intelligente, notamment aux conditions climatiques, de manière à permettre des prévisions allant de 20 minutes à 8 heures, en répondant à des critères de sécurité stricts. Outre l'adaptabilité, la performance est cruciale ; or il s'est vite avéré que cette exigence ne saurait être maintenue en choisissant les avions eux-mêmes comme agents. La solution qui a été retenue est de prendre pour agents des secteurs, à qui il incombe de surveiller les plans de vol de tout avion survolant « son » territoire. Non seulement cette décision a permis de réduire le nombre d'agents de façon draconienne, mais aussi de dynamiser le modèle : en effet, en cas de survol intensif d'un secteur, le nombre d'informations qu'un secteur doit gérer pourrait saturer son algorithme d'apprentissage, auquel cas il peut être dédoublé. À l'inverse, quand le trafic est peu dense, des secteurs peuvent être désactivés.

Le deuxième exemple vient d'un article[126] présentant une application de gestion de données médicales dans le contexte d'un patient vis-à-vis d'un assureur. En l'occurrence, le respect de la vie privée est primordial. Le pari des auteurs a été de réifier l'information elle-même : elle devient un agent BDI, qui peut se reproduire, se diffuser, mais aussi accorder – ou refuser – l'accès aux données qu'elle véhicule. Les auteurs y voient plusieurs avantages : une grande cohérence dans la définition et l'implémentation des droits d'accès à l'information ; les conflits d'intérêt sont évités ; un gain de performance aussi. L'idée derrière la construction est que les éléments informationnels (*information elements*) se portent garants eux-mêmes de la confidentialité. L'exemple type d'un élément informationnel est un formulaire de demande pour une assurance vie. Le formulaire comporte trois sous-ensembles de données : données personnelles, données d'assurance et données médicales. Ces sous-ensembles, tout en étant interconnectés, ne peuvent pas indifféremment être consultés par tous les intervenants ; l'élément informationnel détermine qui a accès à quoi.

Certes, ces exemples ne nous disent en rien si les choix effectués au nom de l'efficacité technique peuvent nous apprendre quelque chose sur le réel, à la manière d'un rasoir d'Occam techniciste, ou si de telles considérations ne pourront être reçues autrement que comme *biais*. Mais après tout, peut-être une clef du succès de la SBA est-elle de transformer une métaphore – qui représente un niveau essentiel dans la compréhension *humaine* des phénomènes – en objet technique ?

### 2.3.4. La SBA dans le temps et l'espace

Eu égard aux caractéristiques de la SBA que sont l'introduction de l'aléatoire, un flux de contrôle décentralisé, la prise en compte explicite du temps et de l'espace, ses points forts semblent tenir au devenir, la mise en place d'une homéostasie ; les conditions de la continuation d'une telle homéostasie ; plus encore, l'étude de la pertinence des interactions locales. La SBA, en engendrant une reconstruction des phénomènes, donne ainsi une certaine épaisseur sémantique à des symboles, épaisseur qui lui vient d'une intégration de plusieurs niveaux de symboles, d'une mise en ordre hiérarchique de la complexité du référent. En cela, l'approche multi-agents se démarque fortement

[125] L'exemple est tiré de K. TUYLS et K. TUMER, *Multiagent Learning*, dans G. WEISS, *Multiagent Systems*, pp. 460-467.
[126] V. WIEGEL, M. J. VAN DEN HOVEN et G. J. C. LOKHORST, *Privacy, deontic epistemic action logic and software agents*.

d'approches plus traditionnelles en intelligence artificielle, où les mécanismes de décision et de contrôle sont toujours uniques, centraux.

Or les systèmes multi-agents n'ont pas le monopole de la gestion décentralisée : la théorie des jeux constitue un formalisme mathématique fondé sur l'idée que pour connaître l'optimum global, il faut tenir compte de l'intention des joueurs individuels d'optimiser leurs objectifs individuels mais en même temps interdépendants. Pour en revenir à l'article fondateur d'Axelrod et Hamilton[127], portant sur l'évolution de la coopération entre individus dans un cadre évolutionniste, darwinien, ces auteurs soulignaient déjà, dans leur recours à la théorie des jeux, l'interaction probabiliste entre individus, interactions inscrites dans le temps, afin d'expliquer l'évolution et la stabilisation d'un environnement coopératif. Traitement probabiliste, interaction entre individus, le tout pour expliquer le devenir d'une régularité, voici des éléments d'une recette dont le paradigme multi-agents n'a pas l'exclusivité.

Qu'est-ce qui distingue donc la SBA de la théorie des jeux ? De par la similarité des métaphores utilisées, la question arrive spontanément[128]. Au regard des observations que nous avons déjà faites sur la SBA, nous devons pourtant mettre en relief les différences, à notre sens profondes, entre les deux approches. La théorie des jeux est un calcul mathématique mono-formel : les « joueurs » – prenons la métaphore au sérieux un instant – ont tous la même stratégie et le même but, la même vision du monde, le même type d'intérêt. Si la métaphore présente une pertinence, c'est qu'elle attire l'attention sur ce que l'effort d'optimisation est *multi-focal* : le formalisme n'étudie pas un unique résultat global à optimiser, mais vise au contraire des cibles multiples, individuelles, résultant de *calculs interdépendants*. Partageant la même fin et les mêmes moyens, les joueurs ne se différencient que dans la mesure où le « jeu » considéré fait diverger leurs intérêts. Or contrairement à la théorie des jeux, la SBA n'est pas un outil mathématique. Elle peut servir de support à l'intégration de formalismes ; et même si elle est souhaitable, la formalisation n'est aucunement nécessaire. Nous voudrions éclairer les différences entre la théorie des jeux et la SBA sous le jour de quelques aspects qui lui donnent sa pertinence : la gestion répartie du flux de contrôle, d'abord, la gestion du temps et de l'espace ensuite.

Commençons par la gestion répartie du flux de contrôle. Dans un livre sur lequel nous aurons à revenir au troisième chapitre, Peter Danielson[129] étudie la prise en compte de la « moralité » – entendue comme l'astreinte personnelle de se conformer à certaines règles en dehors de contraintes institutionnelles – dans le cadre de la rationalité économique. Le sujet étant d'abord la rationalité économique, l'auteur ne se départit jamais d'une approche strictement quantitative des gains ; l'intérêt d'un tel ouvrage, dans l'optique qui est la nôtre dans ce mémoire, ne peut donc être qu'assez limité. Il innove cependant, dans la mesure où il cherche à évaluer l'utilisation, par ses joueurs, de stratégies *différentes* : par exemple, il y a des joueurs qui coopèrent toujours, d'autres qui font toujours défection. Et l'auteur de souligner qu'une telle entreprise n'est tout simplement pas possible en se servant de la théorie des jeux sur laquelle la théorie standard de la rationalité

[127] Est ici visé, l'article paru en 1981 de R. AXELROD et W. D. HAMILTON, *The Evolution of Cooperation*.
[128] Treuil et ses collègues pensent d'ailleurs qu'un rapprochement poussé est possible.
[129] Nous nous fondons ici sur P. DANIELSON, *Artificial Morality*. Pour une introduction à la théorie des jeux, le lecteur peut consulter également G. GIRAUD, *La théorie des jeux*.

économique prend appui[130]. L'auteur s'est donc vu obligé de recourir à Prolog afin d'implémenter ses stratégies et d'observer comment la population de joueurs évolue. Ainsi, la volonté de modéliser quelque chose d'aussi simple qu'une stratégie de jeu force l'auteur à abandonner le formalisme dont il cherche pourtant à s'éloigner le moins possible. Les « tournois » qu'il modélise ainsi en Prolog peuvent être vus comme un premier pas vers une modélisation multi-agents. Nous en sommes pourtant encore loin !

Pour le faire sentir, continuons par le deuxième aspect, la gestion du temps : celle-ci est très rudimentaire dans le cas de la théorie des jeux. Le jeu consiste en une séquence de coups, et chaque coup est joué simultanément par chaque joueur. La seule variante possible, ce sont les jeux dits « sous forme extensive » ou « séquentiels », où une différence est faite entre le joueur qui joue « en premier », etc. En revanche, la simulation à base d'agents se montre ici[131] nettement plus expressive. Comme la théorie des jeux, elle connaît la discrétisation du temps sous la forme de « pas de temps », l'équivalent des coups à jouer. Cependant à ceci s'ajoutent d'autres possibilités : selon les choix de modélisation, les agents peuvent « jouer » dans un ordre fixe (ce qui, le plus souvent, sera cependant considéré comme un biais) ou dans un ordre tout à fait aléatoire ; de façon synchrone ou asynchrone. À l'intérieur d'un même agent, la séquence des actions peut également varier. Toutes ces possibilités de séquençage confèrent à l'ordonnancement en SBA une dimension temporelle particulièrement riche.

Comme nous l'avons dit, la théorie des jeux cherche à modéliser comment des individus optimisent leurs gains au fil d'un certain nombre de tours. C'est-à-dire que le temps y a toujours le rôle d'un *explicandum*, facteur qui doit être expliqué, et non un facteur qui peut être soumis à enquête. Or la simulation à base d'agents peut faire intervenir le temps en tant qu'*explicans*, par exemple à l'occasion de recherches qui ont pour but d'expliquer l'influence de certains évènements historiques sur une configuration spatiale. Même si ce cas de figure ne semble pas fréquent, son intérêt mérite que nous prenions le temps d'un exemple, SIMPOP[132]. Dans SIMPOP, la question scientifique est, entre autres, de savoir quelle influence certains évènements peuvent avoir sur les « systèmes de villes », structures à temporalité lente, réputées pour leur résistance au changement. S'y étudient les changements dans les populations et la richesse des villes, ainsi que les réseaux de relations fonctionnelles qu'elles entretiennent. La perte (ou le gain) d'une fonction (administrative, commerciale, industrielle, culturelle…) peut résulter d'un changement dans les ressources en population ou en richesse de la ville, des changements fonctionnels affectant ses voisins, mais aussi – et c'est là que le temps prend manifestement un rôle *d'explicans* – de certains facteurs historiques, comme les innovations que sont la révolution industrielle, les chemins de fer ou les grandes manœuvres politiques comme la constitution de la zone Euro, etc.

C'est cependant dans la gestion de l'espace que les différences vont devenir éclatantes, car la théorie des jeux ignore l'espace, pour ainsi dire, absolument : elle ne connaît qu'une fonction de réponse aux stratégies des joueurs, un registre de gains et de pertes ; elle est dépourvue de topologie. En SBA, en

[130] P. DANIELSON, *op. cit.*, p. 93.
[131] Sur la gestion du temps dans les simulations à base d'agents, voir J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, *Modélisation et simulation à base d'agents*, pp. 129-134.
[132] *Ibid.*, pp. 91-97.

revanche, les possibilités topologiques foisonnent. Pour bien faire sentir ce point, le détour par quelques illustrations[133] nous semble opportun. La première est SIMANCHOIS, projet qui modélise un écosystème sous-marin pour répondre à la question de savoir quel rôle le vent et la température de l'eau jouent dans les évolutions de population des anchois. L'application de ce type de simulation est à chercher dans la règlementation de la pêche et la préservation de l'environnement. Les anchois évoluent dans un environnement à trois dimensions, à vingt niveaux de profondeur, où les vents, les températures et les courants sont simulés sur la base d'observations climatologiques sur dix ans. Nous n'irons pas dans le détail ni du comment ni du calcul, l'important est de se convaincre comment la simulation à base d'agents peut incorporer des données environnementales diverses et variées et qui se greffent sur une représentation spatiale.

Une deuxième illustration est fournie par le projet EOS, qui visait à modéliser des sociétés préhistoriques. Plus concrètement, le projet cherchait à expliquer la croissance de l'organisation sociale à l'époque paléolithique sur le sol de ce qui deviendrait bien plus tard la France. Cette croissance s'exprime par la multiplication des sites et objets archéologiques, par l'intensification des échanges commerciaux, par l'augmentation des richesses, peintures dans les grottes de Lascaux… Bref, les sociétés de l'époque connaissent une complexité grandissante. Le projet EOS a été mis sur pied pour corroborer la thèse selon laquelle cet essor est dû à la disponibilité prolongée de ressources alimentaires abondantes. Les agents sont des systèmes experts en Prolog[134], dont la connaissance du monde qui les entoure comprend des croyances sur les ressources disponibles dans l'environnement et des croyances sociales sur les autres agents. Le but des agents est de satisfaire leurs besoins en ressources (poisson, gibier, fruits…). Pour cela, ils sillonnent le monde connu – en l'occurrence une grille de 10 000 sur 10 000 cases – à leur recherche. Or n'importe quel agent ne peut pas valoriser n'importe quelle ressource : toutes les ressources sont associées à un profil de compétence : ainsi pour valoriser du gibier, il faut être chasseur. Les agents ont la possibilité pour collaborer afin d'obtenir des ressources convoitées au moyen d'une mise aux enchères[135] : lorsqu'un agent trouve une ressource pour laquelle il n'a pas la compétence requise, il peut en faire la publicité afin de trouver des partenaires. Les résultats principaux du projet EOS sont, d'une part, l'importance d'une perception la plus large possible de l'environnement et, d'autre part, l'influence primordiale de la complexité des ressources : des ressources nécessitant trop de compétences sont le plus souvent périmées avant de pouvoir être exploitées. Le projet met donc en lumière l'importance de la perception, ainsi que le traitement cognitif des ressources dans la constitution de groupes sociaux complexes, alors que la théorie des jeux ignore ces facteurs d'explication avec superbe[136].

---

[133] Les deux premières illustrations sont tirées de J.-P. TREUIL, A. DROGOUL et J.-D. ZUCKER, pp. 41-47 pour ce qui est de SIMANCHOIS et pp. 73-82 pour ce qui est de CUBES ; la dernière – EOS – provient quant à elle de M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 215-217. Notons que nous avons déjà eu l'occasion de parler de SIMANCHOIS dans la section consacrée à l'agent évolutif (§ 2.1.2.3).

[134] Pour une présentation de Prolog, voir B. KOWALSKI, *Logic Programming*, dans J. SIEKMANN, *Computational Logic*, pp. 523-569.

[135] Via le *Contract Net Protocol*. Nous ne saurions entrer ici dans les détails : voir E. DURFEE et Shl. ZILBERSTEIN, *Multiagent Planning, Control, and Exécution*, dans G. WEISS, *Multiagent Systems*, pp. 495-497.

[136] Notons que la théorie des jeux connaît les jeux « à information imparfaite » : il faut comprendre que les joueurs ne sont pas sûrs, dans ce type de jeu, de l'espérance de gain attachée à chaque coup. P. DANIELSON (*op. cit.*, pp. 148-162) note que la prise en compte de quelque chose aussi simple qu'est le *coût de l'information*, rendue possible par l'utilisation de Prolog, a une incidence importante sur le résultat de ses tournois.

En incorporant ainsi explicitement la dimension spatiale, la simulation à base d'agents parvient à apporter un éclairage nouveau sur certains résultats contre-intuitifs auxquels aboutit le recours à la théorie des jeux. Outre l'exemple d'EOS, que nous venons de citer, il est également possible, en appliquant la simulation à base d'agents au dilemme des prisonniers[137], de constater que face à l'intrusion d'individus tricheurs, la résistance s'organise sous forme d'îlots coopératifs, des regroupements d'individus en communautés d'entraide. La prise en compte de l'espace est donc un facteur essentiel, dont l'importance ne saurait être sous-estimée.

L'exemple qui précède pourrait donner à croire que la SBA privilégie une conception de l'espace *statique*, structure figée qui ne fait qu'accueillir les déplacements des agents qui s'y côtoient. En réalité, la plupart des plateformes de SBA définissent l'espace non comme un graphe, mais comme un automate cellulaire[138]. Pour prendre l'exemple de Netlogo – plateforme connue pour sa grande accessibilité aux non-informaticiens[139] – l'espace est modélisé par défaut en tant que « places », qui sont elles-mêmes des agents, immobiles certes, mais dotés d'attributs qui peuvent évoluer dans le temps. Ce type d'enrichissement, au demeurant très intuitif, permet de mesurer certains phénomènes non linéaires discrets : la topologie, ainsi, devient elle-même dynamique.

Le formalisme de l'automate cellulaire soulève cependant une interrogation quant à sa richesse d'expression spatiale : en effet, traditionnellement[140], la structure d'un automate cellulaire est assez rigide : il s'agit toujours d'un quadrillage, qui ne connaît que deux types de voisinages : dans le voisinage dit « de Von Neumann », une case a quatre voisins : ceux avec qui elle partage une frontière commune ; dans le voisinage dit « de Moore », une case a huit voisins : tous ceux avec qui elle partage un point commun. Que le modélisateur choisisse l'un ou l'autre, l'espace reste « une grille », foncièrement uniforme. Certes, ceci est utile pour modéliser des relations de voisinage entre unités géographiques qui se laissent déterminer selon des critères topologiques (notamment, la contiguïté) ou métrique (la distance euclidienne « à vol d'oiseau »). Ces hypothèses deviennent toutefois gênantes lorsqu'il s'agit d'étudier d'autres types de voisinage. Ainsi, en morphologie urbaine, les fonctions du bâti : les bâtiments, en effet, entretiennent entre eux non seulement des relations euclidiennes (voisinage visuel, diffusion du bruit ou d'un polluant), mais aussi des relations de type *réticulaire* : les cheminements des rues, les réseaux de transports, etc. Il faut donc relâcher la contrainte de régularité qui pèse sur l'automate cellulaire afin de pouvoir intégrer les deux types de relations, proximité aréale et proximité fonctionnelle.

Formellement, un automate cellulaire comprend deux composantes, une composante structurelle et une composante dynamique, comme suit :

$$AC = (\{U, V\}, \{E, F\})$$

[137] Exemple dû à H. BERSINI, *Quételet, l'invention de l'homme moyen par un homme tout sauf moyen*.
[138] Nous empruntons cette observation à P.-M. BOULANGER et Th. BRÉCHET, *Modélisation et aide à la décision pour un développement durable*, pp. 42-43, 100, 136-137.
[139] Le lecteur intéressé peut se référer à https://ccl.northwestern.edu/netlogo/ pour plus de détails ; le code source est disponible sur GitHub : https://github.com/NetLogo/NetLogo.
[140] La caractérisation de l'automate cellulaire, ainsi que l'identification de sa composante structurelle à un graphe, proviennent de l'article de D. BADARIOTTI, A. BANOS et D. MORENO, *Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine*.

Le binôme *{E, F}* est la composante dynamique, avec *E* l'ensemble des états possibles des cellules et *F* l'ensemble des fonctions de transition, telles qu'à chaque pas de temps *t*, il est possible de calculer l'état d'une cellule donnée en fonction de son état, ainsi que de l'état des cellules environnantes, au pas de temps précédent :

$$E_{i^{t+1}} = f\left(E_{i^t}; E_{V_{i^t}}\right)$$

La composante structurelle comprend *U*, l'ensemble fini des unités spatiales et *V*, l'ensemble des voisinages de ces unités.

Or il se fait que les deux types d'automates cellulaires classiques – voisinages de Von Neumann et de Moor - se traduisent sans problème en graphe de relations. Ainsi, le binôme structurel de l'automate *{U,V}* peut être interprété en termes d'un tel graphe *G(U,R)*, où *R* est l'ensemble des arcs exprimant les relations de voisinage entre nœuds *U*. Le recours aux graphes permet de fortement complexifier la représentation spatiale, tout en rendant disponible un large éventail d'outils mathématiques et informatiques développés pour analyser des graphes de grandes dimensions, où le nombre d'arcs et de nœuds se chiffrent par millions. Pour les besoins de notre exposé, il convient de noter deux choses : premièrement, que la simulation à base d'agents, capable à la fois de faire reposer l'espace sur des graphes ou des automates cellulaires, voire les deux en même temps (comme l'exemple de CUBES nous le montrera tout à l'heure), détient ainsi une base mathématique très expressive pour formaliser les relations spatiales ; deuxièmement, que l'espace ne doit pas être vu comme un « décor » statique, mais comme un ensemble d'agents sinon mouvants, du moins tout aussi dynamiques et dignes d'intérêt que les contreparties mobiles.

À la profusion de détails physiques de SIMANCHOIS et d'EOS répond le dépouillement de CUBES : il s'agit ici d'une modélisation cherchant à rendre compte de la pression de marques concurrentes proposant des produits similaires sur le comportement des consommateurs. La problématique de CUBES porte plus particulièrement sur les raisons qui poussent les consommateurs à l'achat ; la question de recherche est d'éprouver l'hypothèse suivante sur le facteur décisif qui guide la décision d'achat : plutôt que des considérations rationnelles sur le prix ou pragmatiques sur le lieu et l'instant d'achat, les interactions sociales du consommateur seraient déterminantes. Le modèle fait intervenir plusieurs niveaux de description, allant du niveau psychologique individuel au niveau économique du marché en passant par les interactions et dynamiques de groupe dont la sociologie se charge, qui font que le modèle se situe à un véritable carrefour d'apports disciplinaires. Cependant, malgré la complexité inhérente aux champs théoriques, et l'effort théorique considérable pour les articuler en un seul modèle, les agents et l'environnement de CUBES sont rendus de façon relativement simple.

Dans CUBES, chaque consommateur est placé sur une grille qui définit un voisinage de proximité immédiate, voisinage par lequel il faut entendre son cercle de fréquentations habituelles – groupe familial ou groupe d'amis. La grille a donc la particularité d'exprimer non pas une proximité aréale, mais – de façon qu'il faut qualifier de métaphorique – fonctionnelle[141]. Un deuxième type d'espace définit un voisinage social élargi, où les rapprochements expriment des liens d'affinité entre

[141] Selon les auteurs, ce choix a été fait pour simplifier le modèle.

consommateurs sur la base d'attributs communs (par exemple, un même niveau de revenus ou d'instruction). Ces liens sont valués : il est ainsi possible de calculer, pour chaque paire de consommateurs, une probabilité d'interaction. Pour notre propos, qui est celui de montrer la diversité des espaces, il est important de souligner que « l'espace », qualifié de « social », est ici entièrement métaphorique. Or cela n'empêche pas la formalisation de la notion de manquer de rigueur, car elle répond aux exigences de la théorie des graphes : que la proximité soit fonctionnelle ou sociale, elle se modélise dans tous les cas de figure par des arcs pondérés reliant des nœuds. La modélisation de l'espace devient ainsi un lieu, ou un niveau, de description mathématique à elle seule.

# Chapitre III. Les systèmes multi-agents et l'éthique

*« Quand tu atteins le sommet de la montagne,*

*Continue à monter », dit l'Androïde, gardien des cimes*

*électroniques…*

Éric BROGNIET

## 3.1. Éthique et moralité sous le signe de l'altérité

Dans les travaux que nous allons parcourir dans ce chapitre, il est le plus souvent question de *normes* : comment un agent peut-il acquérir, diffuser, apprendre, appliquer une norme, ou encore gérer des conflits entre elles ? La norme y est donc centrale. Or dans notre caractérisation de l'éthique, nous n'avons guère parlé de la norme ; dans l'introduction de ce mémoire, en effet, nous avons fait la part belle à la valeur qui – toujours dans notre conception de l'éthique – forme la charpente de l'édifice. Nous en tenir à une telle caractérisation nous ferait cependant courir le risque de passer à côté d'un aspect essentiel de l'éthique, qui est celui de *l'obligation* : l'éthique nous appelle, exige, impose des épreuves de nos valeurs comme de nos actions[1]. C'est pourquoi il faut articuler les deux notions : comment la valeur, centrale à notre conception de l'éthique et qui, avons-nous dit, implique une image de l'homme qui guide son action, se rapporte-t-elle à la norme ?

Notre définition de la valeur est très proche de ce que Paul Ricœur a appelé *l'estime de soi*. L'estime de soi est l'auto-interprétation, de type narratif, que chacun se donne de sa propre vie, à la lumière de l'idée qu'on se fait d'une « vie bonne ». Pour les besoins de la discussion qui suit, appelons *éthique* la *visée* d'une vie bonne, accomplie, et *moralité* ce qui s'impose comme *obligatoire*, ce qui articule la visée dans des *normes*. La visée de la vie bonne est reprise par Ricœur à Aristote, qui tient ici des vues qui l'opposent à la fois au déontologisme de Kant et au téléologisme des utilitaristes[2]. Contre Kant, qui lie vertu et bonheur – suivant en ceci la tradition épicurienne et stoïcienne –, il affirme l'irréductibilité du bonheur à la vertu. Contre les utilitaristes, il cherche le principe de la finalité d'abord dans la praxis. Les pratiques, en effet, disposent d'étalons d'excellence qui leur sont immanents : ainsi, pour être un « bon » nageur : les critères tiennent à la souplesse des mouvements, la vitesse obtenue, l'endurance à la nage, etc. La visée de la vie bonne s'ajoute au bien immanent de

[1] La distinction – et l'articulation – des aspects déontologiques et téléologiques de l'éthique est reprise à P. RICŒUR, *Soi-même comme un autre*, pp. 199-253.

[2] Précisons que Ricœur ne parle que très incidemment de l'utilitarisme, lorsqu'il critique John Rawls d'avoir confondu la téléologie avec l'une de ses formes, l'utilitarisme justement (pp. 230-231, 267-268). Quant à nous, nous sommes bien obligé d'expliciter davantage l'opposition, étant donné l'importance de l'utilitarisme en éthique des machines, vue au premier chapitre.

la pratique comme « idée limite », comme finalité qui se rattache à un niveau praxéologique supérieur, finalité propre au « plan de vie » : ainsi le nageur trouve dans l'attention portée au corps une finalité de niveau hiérarchique plus élevé, qui cependant reste toujours *intérieur à l'agir humain*.

La conception aristotélicienne du *telos* d'une pratique a cet avantage tout à fait capital de ménager un espace pour les « préceptes » et, par là même, à la norme. Déontologie et téléologie peuvent, par suite, se conjuguer de façon un peu particulière : la téléologie enveloppe la déontologie, en tant qu'elle est à la source des normes et qu'elle préside à la résolution de conflits entre elles, lorsqu'elles en viennent à se contredire. La résolution des conflits est due, en effet, à la délibération (*phronèsis*), qui préside aux choix entre plusieurs cours d'action, c'est-à-dire des choix entre autant de fins qui trouvent leur ancrage dans l'idéal de vie. Il est ainsi possible de remonter la hiérarchie des fins, tout en nous rappelant que la question du bonheur, fin dernière, met un point d'arrêt à l'interrogation délibérative. Néanmoins, la délibération éthique ne peut faire fi des normes : celles-ci sont le lieu où s'applique la rigueur du formalisme déontique. Un formalisme à vocation universelle dont Kant donne l'exemple constitue ainsi l'épreuve à laquelle il faut soumettre la visée éthique. Nous avons donc, au final, trois lieux de rencontre entre téléologie et déontologie : la visée de la vie bonne préside à la création des normes, ainsi qu'à la résolution de conflits entre elles dans une situation concrète[3]. Dans un mouvement inverse, la norme constitue l'épreuve à laquelle la visée doit se soumettre.

Rappelons-nous, ces points de rencontre où Ricœur voit un rôle pour la téléologie ne sont pas sans évoquer les champs de recherche que la SBA éthique cherche à occuper : l'acquisition et la diffusion des normes nous parlent de l'émergence de la norme ; la façon dont la SBA envisage la gestion des conflits entre normes – nous le verrons – investit la fonction de la norme comme épreuve d'universalisme. Ricœur prend cependant soin d'insister[4] que l'exigence d'universalisme ne doit pas aboutir à un formalisme procédural vide qui condamnerait toute prise en considération sérieuse du contexte. L'exigence d'universalisme doit se conjuguer aux conditions de mise en contexte ; elle doit prendre acte qu'elle ne peut agir que dans un environnement constitué, en dialogue avec une tradition.

Dans ce qui précède, nous avons toujours réduit la « visée éthique » à la vie bonne. Or la vie bonne n'en constitue qu'un premier niveau, le niveau personnel. Il faut dire, en effet, que si Ricœur distingue entre déontologie et téléologie, il le fait à trois niveaux : les niveaux personnel, interpersonnel et impersonnel. Au niveau personnel, le pendant « moral » de l'estime de soi est le respect de soi, au sens kantien : plus nous intériorisons la norme, plus nous devenons autonomes. Dans l'interaction interpersonnelle avec autrui, des relations de type « je-tu », la composante de la visée éthique est la sollicitude : comment vivre avec et pour l'autre ? La sollicitude pour autrui exprime la valeur de l'autre en tant qu'être irremplaçable, insubstituable. La sollicitude pour autrui tire sa force de *sa similitude avec l'estime de soi* :

[3] Insistons sur le caractère toujours particulier de la délibération dans ce contexte : elle porte toujours sur la question de savoir ce qui est bon dans une *situation concrète*. À ce titre, elle est réflexion sur la contingence, à l'opposé d'une réflexion de type scientifique, qui porte sur le nécessaire (A. COMTE-SPONVILLE, *Petit traité des grandes vertus*, pp. 48-51).
[4] Voir la critique de John Rawls et Jürgen Habermas à la page 333.

*Les agents et les patients d'une action sont pris dans des relations d'échange qui comme le langage conjuguent réversibilité des rôles [du « je » et du « tu »] et insubstituabilité des personnes. Ce que la* sollicitude *ajoute, c'est la dimension de valeur qui fait que chaque personne est* irremplaçable *dans notre affection et dans notre estime. […] C'est d'abord pour l'autre que je suis irremplaçable. En ce sens, la sollicitude répond à l'estime de l'autre pour moi-même. […]*

*Au-dessus enfin des idées de réversibilité des rôles et d'insubstituabilité des personnes – cette dernière idée élevée jusqu'à celle d'irremplaçabilité – je placerai la* similitude*, [qui] est le fruit de l'échange entre estime de soi et sollicitude pour autrui. Cet échange autorise à dire que je ne puis m'estimer moi-même sans estimer autrui* comme *moi-même. Comme moi-même signifie : toi* aussi *tu es capable de commencer quelque chose dans le monde, d'agir pour des raisons, de hiérarchiser tes préférences, d'estimer les buts de ton action et, ce faisant, de t'estimer toi-même comme je m'estime moi-même*[5]*.*

Notons encore que la contrepartie morale de la sollicitude est le respect absolu d'autrui, tel que Kant l'a thématisé.

Enfin, l'éthique s'applique au niveau impersonnel, c'est-à-dire en tant qu'elle vise le rapport aux autres avec qui le soi n'aura jamais une relation interpersonnelle ou dialogale. Ce troisième niveau est celui de la pluralité : il renvoie à un tiers inclus, mais anonyme, sans visage. Il constitue le terrain du pouvoir, de l'agir-ensemble, de l'exercice politique en tant qu'il fait appel à l'éthique, c'est-à-dire, à *un savoir de soi* qui ne se révèle que dans le cadre d'une institution ou d'un État. À ce niveau, la visée éthique prend la forme de la justice distributive, l'exigence d'égalité. La justice comme « valeur », *à chacun son dû*, présuppose une vie sociale politiquement et socialement organisée. Elle ne s'applique pas à l'état naturel, mais à l'homme en société. Par « société », il faut comprendre ici un système de contrôle de répartition de rôles, un ensemble d'institutions qui sont autant de mises en relation. L'égalité est à la vie dans les institutions ce qu'est la sollicitude aux relations interpersonnelles. Sa contrepartie au plan moral est la justice réparatrice.

Ricœur fait ainsi droit aux dimensions interactive et collective de l'éthique. Tenant compte de ce que nous avons vu tout au long du deuxième chapitre, nous devinons que les systèmes multi-agents seront particulièrement propices à explorer les niveaux interpersonnel et institutionnel. Nous avons vu en effet comment ces systèmes s'organisent comme des réseaux de relations, des espaces réticulés – spatialement ou relationnellement ; par là même, ils donnent « corps » à la notion même de *distance*, à la possibilité de configurations qui dépassent les individus. Ainsi, ils permettent d'exprimer la différence entre les sphères interpersonnelles et impersonnelles. Au niveau interpersonnel, rappelons-nous que les SMA ont fait de l'interaction le ressort principal de leur efficace ; nous y verrons à l'œuvre le jugement éthique, mais aussi l'acquisition et la diffusion des normes, comme leur mise en pratique dans la coopération et la négociation. Au niveau institutionnel, nous avons vu la facilité avec laquelle de tels systèmes s'élargissent à un enrichissement collectif sous la forme de rôles, d'équipes et d'institutions. Nous verrons maintenant comment ces ajouts peuvent contribuer à la moralité des agents, entendue comme l'emprise de la norme.

[5] P. RICŒUR, *op. cit.*, p. 226.

## 3.2. La téléologie en SMA

### 3.2.1. La question de la motivation

Dire que les systèmes multi-agents brillent aux plans interpersonnel et impersonnel ne signifie pas pour autant que nous pouvons faire entièrement abstraction du niveau personnel. Certes, nous avons vu comment le fonctionnement interne d'un agent échappe le plus souvent au rapport que la métaphore du système – de l'organisation – cherche à capter. Il faut donc que la métaphore « vise juste », que le choix des individus qu'elle implique, cible ces deux niveaux supérieurs. En outre, les agents doivent être des candidats sérieux au titre d'unités d'agir ; rappelons-nous qu'il s'agit d'individus ou « références identifiantes » qui produisent des effets répétés – pour tout le moins répétables – dans le réel. Or sur le plan personnel, qu'est-ce qui peut pousser l'agent à agir conformément à l'éthique, plus précisément à *une* éthique, *son* éthique ? C'est tout le problème de la motivation des agents. Comme l'explique Michel Meyer[6], toute théorie éthique, qu'elle soit utilitariste ou contractualiste, qui cherche à fonder le souci des autres dans l'individu, se doit de *postuler* une affinité naturelle, spontanée, subjective, de l'individu à faire le bien ; or cette affinité présuppose déjà ce qu'il s'agit de fonder. Il y a donc cercle.

Le problème n'est pas que conceptuel ; les psychologues[7] se penchent également sur le problème de la motivation éthique. Traditionnellement, leurs recherches se sont intéressées au rôle de la *cognition* morale (le raisonnement) dans la motivation. Plus tard, sous l'influence des travaux d'Antonio Damasio, *l'émotion* morale y a été jointe. Or même considérées ensemble, cognition et émotion s'avèrent ne jouer qu'un rôle modéré dans l'action morale. C'est pourquoi les psychologues se tournent vers *l'identité morale* comme source de motivation :

> *[…] when morality is important and central to one's sense of self and identity, it heightens one's sense of obligation and responsibility to live consistent with one's moral concerns*[8].

Dans cette conceptualisation, l'identité morale comprend deux versants : un versant objectif, fait de contenus et de représentations, et un versant subjectif, l'expérience de soi. Il s'agit ici d'un sentiment d'agentivité, de maîtrise ou de contrôle de soi. Plus ce sentiment est développé, plus l'individu se sent la responsabilité de protéger sa propre identité. En d'autres termes, plus l'identité subjective est développée, plus l'individu désire être consistent avec lui-même, et donc vivre selon les contenus – moraux – de soi[9]. Même si les liens entre les trois « fonctions » psychologiques – émotion, identité et cognition – ne sont à ce jour pas tirés au clair, l'identité implique un changement d'accent, plus « agentif » (*agentic*) de la moralité. Les individus, plutôt que d'être des réceptacles passifs de facultés

[6] M. MEYER, *Principia Moralia*, pp. 15-25.
[7] Voir l'article de S. A. HARDY et G. CARLO, *Identity as a Source of Moral Motivation*.
[8] *Ibid.*, p. 234.
[9] Ricœur n'aurait probablement pas goûté cette façon dualiste de présenter de l'identité, lui pour qui même le désir est une catégorie mixte de *sens* (dans le registre de la justification) et de *force* d'agir (*Soi-même comme un autre*, p. 83).

à connotation éthique (*morally-relevant capacities*), procèdent à une intégration active de contenus moraux choisis, plus ou moins délibérément, au fur et à mesure qu'ils se construisent :

> *[…] a more agentic picture of morality, where individual differences in moral desires, rather than differences in morally-relevant capacities (e.g., empathy, moral reasoning, or moral schemas), are the root of individual differences in moral behavior*[10].

Dans le domaine multi-agents qui nous occupe, disons-le d'emblée, ce genre de considérations est peu exploré. Cependant, la question de la motivation n'est pas totalement absente des travaux dont nous avons pu prendre connaissance. Citons pour exemple la plate-forme multi-agents York[11]. Il s'agit d'un environnement où les agents ont un « corps » en Java et un « esprit » Progol. La force de Progol est d'intégrer deux aspects : d'une part, un aspect d'apprentissage par programmation logique inductive, que nous avons déjà rencontré au premier chapitre (§ 1.4.2) chez les époux Anderson ; d'autre part, Progol inclut un interpréteur complet Prolog, c'est-à-dire qu'il peut aussi servir comme base de données de clauses apprises. En l'absence d'apprentissage, le comportement par défaut est basé sur les pulsions primaires (*drives*) que sont la faim, la soif, la peur et le désir sexuel. Nous retrouvons ici les motivations comme moteur du comportement : la faim et la soif introduisent à la gestion des ressources dans l'environnement ; l'appétit sexuel permet d'introduire une forme rudimentaire de sélection naturelle ; la peur, pour finir, donne lieu à des comportements d'évitement de dangers et d'obstacles. À chaque tour, l'agent satisfait alors la pulsion à l'intensité la plus élevée. Nous y voyons le rôle des structures de récompense dans l'apprentissage qui, par défaut sur cette plate-forme, sont purement individuelles. Il est cependant possible d'introduire des récompenses autres, notamment dans le contexte d'un travail d'équipe.

L'aspect motivationnel, ici, n'est certes pas ressenti comme participant d'une valeur[12]. Cependant, il convient de noter que même les pulsions les plus élémentaires – et spécialement celles-là – peuvent faire l'objet d'un travail sur soi éthique considérable. Prenons ainsi l'exemple de la faim ; celle-ci a tout à voir avec la façon dont nous voyons et considérons notre corps et ses besoins. La faim, dès lors, peut être thématisée comme source de tempérance ; ce qui explique que dans certaines philosophies (écologiste, juive, épicurienne, orientale…), le « bon usage » de la faim devient un lieu éthiquement investi, dont l'automatisme est banni. Simple affaire d'éthos, comme nous l'avons soutenu précédemment (§ 1.8.1) ? Nous inspirant de l'apport de Ricœur, qui nous apprend à conjuguer contextualisme et universalisme, il est possible d'aller un peu plus loin. En effet, si beaucoup de pratiques culturelles proposent des stylisations des motivations primaires, nous pouvons dire que ces rituels – portant sur le contenu de ce qui peut être mangé, sur les façons de les préparer, sur leur inscription dans le temps calendaire, etc. – sont éthiques en tant qu'ils impliquent une image de l'homme comme « maître de maison » : l'homme domine son corps et ses pulsions.

---

[10] S. A. HARDY et G. CARLO, *loc. cit.*, p. 237.

[11] D. KAZAKOV et D. KUDENKO, *Machine Learning and Inductive Logic Programming for Multi-Agent Systems*.

[12] Dans un article dont nous avons pris connaissance tardivement, le lien entre motivation et valeurs dites « de base » sont en revanche au cœur des préoccupations (Sh. H. SCHWARTZ, *Les valeur de base de la personne*) : l'auteur y soutient qu'une dizaine de valeur de base, se retrouvant dans toutes les cultures actuelles, permettent de répondre à trois nécessités élémentaires : la satisfaction des besoins biologiques des individus, l'interaction sociale, le bon fonctionnement et la survie des groupes (nous y retrouvons sans peine les trois niveaux de distance personnelle de Ricœur).

Nous ne voudrions pas terminer cette section sans rapporter le cas d'un projet de recherche[13] qui prend pour objet de simulation le scénario des incendies de forêt qui embrasent le continent australien. Le problème des auteurs est de simuler de la façon la plus réaliste possible les comportements des civils confrontés à la menace de perdre leur maison, la vie, ou leurs proches. Le souci de réalisme a été poussé très loin : les auteurs ont ainsi intégré des apports venus des sciences sociales et psychologiques afin d'enrichir leurs agents d'aspects émotionnels et sociaux importants.

L'architecture proposée se présente, de prime abord, comme une extension BDI pour plateforme de simulation, telle qu'il en existe déjà pour NetLogo ou Matsim. Nous retrouvons donc des prédicats de croyances, de désirs et des plans, avec une originalité toutefois : une distinction est faite entre croyances certaines ou incertaines – les auteurs parlent aussi *d'attentes* pour cette dernière sorte de croyance. Qui plus est, cette base BDI ne dit cependant pas le tout de l'état « mental » d'un agent, il n'en exprime que le volet cognitif. Deux autres volets viennent compléter l'état mental : un volet émotionnel et un volet social.

Le volet émotionnel repose sur le modèle dit « OCC » (pour *Ortony, Clore et Collins*, d'après les chercheurs en intelligence artificielle qui l'ont porté sur les fonts baptismaux). Cette architecture propose une typologie formelle des émotions, dans laquelle une vingtaine d'émotions sont réparties sur trois groupes, en fonction des « faits » auxquels elles se rapportent : les émotions liées aux conséquences des évènements (joie, tristesse, espoir, peur, satisfaction, déception, soulagement et peur confirmée), les émotions qui se rapportent aux autres agents (contentement pour autrui, compassion, ressentiment et jubilation), les émotions en lien avec les actions, actions de soi (fierté, honte, gratification et remords) ou actions d'autrui (admiration, reproche, gratitude et colère)[14]. Une émotion est ainsi caractérisée par son type, sa base cognitive – c'est-à-dire le prédicat par rapport auquel l'émotion est ressentie – l'agent causant l'émotion, ainsi que l'intensité et la valeur de décroissance de son intensité dans le temps.

Le volet social, quant à lui, modélise les relations sociales. Une relation sociale d'un agent *i* pour un autre agent *j* est modélisé en fonction des quatre facteurs suivants : l'appréciation, la dominance, la solidarité et la familiarité (c'est-à-dire le partage d'informations personnelles). Ces facteurs sont pondérés sur une échelle allant de -1 à 1 pour l'appréciation et la dominance et de 0 à 1 pour la solidarité et la familiarité.

La simulation se déroule, à chaque pas de temps, en exécutant un cycle d'évaluation. Les agents commencent par percevoir leur environnement, après quoi ils mettent à jour les croyances. Également en début de cycle, des phénomènes de *contagion* et de *création d'émotions* ont lieu. Par création, il faut entendre l'application de quelques règles précises, en fonction de la mise à jour des croyances : l'émotion de peur peut se muer en peur confirmée ou en soulagement, l'émotion d'espoir

[13] M. Bourgais, P. Taillandier et L. Vercouter, *Cognition, émotions et relations sociales pour la simulation multi-agent*. Dans ce travail, nous avons eu le plaisir de retrouver la plate-forme de simulation GAMA, que nous avions utilisée dans le cadre de notre stage et de notre travail de fin d'études du bachelier. Les auteurs ont contribué leur architecture à la plate-forme, voir https://gama-platform.github.io/wiki/Using-BEN-simple-bdi.

[14] Pour être tout à fait complet, ajoutons que le modèle OCC comporte également des émotions liées aux objets (ou artéfacts) – cette catégorie ne comporte cependant que deux émotions et n'a pas été reprise par nos auteurs.

se transforme en satisfaction ou déception, à leur tour satisfaction et soulagement peuvent donner lieu à l'émotion de joie ou déception et la peur confirmée à la tristesse.

Ensuite, après tous les changements émotionnels, le moteur social va mettre à jour les relations sociales en se basant sur les émotions. Les émotions à valence positive (la joie, etc.) vont avoir une influence positive sur le facteur d'appréciation, et inversement pour les émotions à valence négative. Certaines émotions spécifiques, telle la peur, vont influencer le facteur de dominance. La solidarité se met à jour en fonction du degré de similarité entre les agents, similarité qui a pour entrées l'ensemble des états cognitifs : désirs, croyances et incertitudes. Le degré de similarité subit ensuite une correction à la baisse si des émotions à valence négative sont apparues entre les agents concernés. La familiarité, à ce stade, n'est impactée que par le facteur d'appréciation : l'idée étant que des agents qui s'apprécient auront plus tendance à échanger des potins. Après la mise à jour des relations sociales vient le noyau du traitement BDI : un désir est sélectionné ; par ce fait même, il devient intention courante. L'intention courante détermine alors le plan d'action. Après la sélection du plan, une dernière phase du cycle consiste à dégrader l'intensité des émotions – un apaisement a lieu, concrètement le moteur soustrait la valeur de décroissance de l'intensité des émotions ressenties – ainsi qu'à dégrader les états cognitifs : l'oubli, en effet, s'installe.

Nous voilà donc en présence d'agents dotés d'un environnement interne déjà très riche. Or, il convient d'être attentif aux *effets* de cette richesse invisible à la surface : dans l'exemple de l'article, les liens sociaux ont une incidence directe sur les comportements d'entraide ou de grégarisme : ainsi un agent à la recherche d'un abri anti-feu mais qui ignore où il pourrait en trouver un, se mettra à suivre un agent avec qui il a un bon contact. Le facteur de solidarité contribue fortement aux comportements d'entraide, en dépit de ce que dicterait un rationalisme égoïste en la circonstance. Il en résulte que les relations sociales et les émotions sociales deviennent des *sources de motivation* de premier plan, au même titre que les états proprement cognitifs des agents. Par suite, nous pouvons conclure que l'agent qui cherche à protéger son identité aura « naturellement » tendance à protéger les liens sociaux et affectifs à autrui qui en font intimement partie.

### 3.2.2. Le jugement éthique

Nous voudrions commencer notre exploration du niveau interpersonnel par le jugement éthique. La raison de ce choix est simple : il s'agit du seul domaine ou nous avons trouvé une tentative d'introduire une dimension téléologique. En SMA[15], le jugement éthique est compris comme un jugement *d'adéquation* entre l'éthique d'autrui et l'éthique de soi, notamment avant de s'engager dans des rapports de coopération. Le degré de compatibilité entre éthiques sera un élément pour décider si autrui est assez *fiable* pour établir un cadre coopératif avec lui.

Le jugement est ainsi une évaluation de l'éthique *des autres* ; il dépasse donc nécessairement un cadre mono-agent. L'évaluation prenant la forme d'une comparaison, il faut aussi que chaque agent

[15] Nous nous référons, pour cette section, à deux articles de N. COINTE, Gr. BONNET et O. BOISSIER, le premier étant *Jugement éthique dans le processus de décision d'un agent BDI*, 2017 ; le deuxième *Jugement éthique dans les systèmes multi-agents*, 2016.

ait une représentation explicite non pas seulement de sa propre éthique, mais aussi de l'éthique des autres (ce qui, en sciences cognitives, est appelé une *théorie de l'esprit*). Le processus de jugement éthique (PJE) est modélisé comme suit :

$$PJE = <RS, PEv, PM, PA, O>$$

où RS est *la reconnaissance des situations* ; PEv *le processus d'évaluation* ; PM *le processus moral* ; PA *le processus d'acceptabilité éthique* et O *l'Ontologie des valeurs morales*. Parcourons maintenant ces différentes composantes.

La reconnaissance des situations (RS) établit le lien avec l'architecture BDI de l'agent (rappelons-nous du deuxième chapitre, § 2.1.2 : ses croyances, ses désirs, ses intentions) en générant l'ensemble de croyances qui décrivent l'état du monde, ainsi que les désirs qui décrivent les buts de l'agent. Formellement, cela donne :

$$RS = <\mathcal{B}, \mathcal{D}, ES>$$

où $\mathcal{B}$ est l'ensemble des croyances que l'agent possède sur l'état courant du monde (W), $\mathcal{D}$ l'ensemble des désirs de l'agent, qui peuvent porter soit sur une action particulière, soit sur un état du monde. ES est une fonction d'évaluation de situations qui met à jour les croyances et les désirs de l'agent à partir de l'état du monde[16] :

$$ES : W \rightarrow 2^{\mathcal{B} \cup \mathcal{D}}$$

Le processus d'évaluation (PEv) établit les actions désirables d'une part, et les actions réalisables dans l'état courant du monde, d'autre part. Formellement, cela donne :

$$PEv = <A, \mathcal{A}_d, \mathcal{A}_r, ED, ER>$$

où A est l'ensemble des actions disponibles (où chaque action est assortie de connaissances de ses conditions d'effectuation et de ses conséquences sur l'état du monde) ; $\mathcal{A}_\mathrm{d}$ le sous-ensemble des actions désirables, produit par la fonction d'évaluation de désirabilité ED ; $\mathcal{A}_\mathrm{r}$ le sous-ensemble des actions réalisables, produit par la fonction d'évaluation de réalisabilité ER.

La fonction de désirabilité se définit comme suit :

$$ED : 2^{\mathcal{D}} \times 2^{A} \rightarrow 2^{\mathcal{A}_d}$$

Une implémentation simple de cette fonction peut être de ne retenir que les actions désirées telles quelles ou, si le désir considéré porte sur un état, de ne retenir que les actions qui sont connues pour comporter cet état dans leurs conséquences. Quant à la fonction de réalisabilité, elle ne retient que

[16] Rappelons-nous le sens de la notation $2^X$ : il s'agit de l'ensemble des parties de l'ensemble X, c'est-à-dire l'ensemble de tous les sous-ensembles de X. Pour reprendre l'exemple de Wikipédia, si l'ensemble E = {a,b,c}, l'ensemble $2^E$ inclut l'ensemble vide (puisque l'ensemble vide est inclus dans tout ensemble), les trois singletons {a}, {b}, {c}, ainsi que les paires {a,b}, {a,c} et {b,c}.

les actions dont les conditions sont compatibles avec ce que l'agent croit être vrai dans le monde. Formellement :

$$ER : 2^{\mathcal{B}} \times 2^{A} \rightarrow 2^{\mathcal{A}_r}$$

Parmi les actions A, le processus moral (PM) établit les actions morales $\mathcal{A}_m$ à partir des supports de valeur (SV), des règles morales (RM), ainsi que des croyances et désirs produits par la reconnaissance des situations RS. Le processus moral s'organise autour de la notion pivot de « règle morale », règles qui sont justifiées et soutenues par des valeurs. Formellement :

$$PM = \langle SV, RM, \mathcal{A}_m, EM \rangle$$

où EM est la fonction d'évaluation morale :

$$EM : 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^{A} \times 2^{SV} \times 2^{RM} \rightarrow 2^{\mathcal{A}_m}$$

Notons bien que les actions morales ne sont pas forcément désirables, ni même réalisables, d'où la présence de $2^A$ dans la fonction. SV est la base de données de tous les supports de valeur, définis comme des couples <s,v> où v est une valeur (élément de $O_v$, l'ensemble des valeurs morales) et s son support, tel que

$$s = \langle a, w \rangle$$

où a est une action et w un sous-ensemble des désirs et croyances ($w \subset B \cup D$). Ainsi la valeur de la générosité pourrait être défini, pour un agent x, comme l'action de donner à un agent y si l'agent x a la croyance que l'agent y est pauvre.

Une règle morale, quant à elle, est un triplet <w, o, m> où w, comme avant, décrit l'état courant du monde en termes de désirs et de croyances ; o est l'objet de la règle et peut être soit une action a, soit une valeur ($v \in O_v$) ; m est une valuation morale, à prendre dans une liste telle que {immoral, moral, amoral} ou {bien, mal} ou encore un coefficient entre 0 et 1. Nous pouvons ainsi formuler des règles telles que « il est immoral de tuer un être humain » ou « l'honnêteté vis-à-vis d'un menteur est morale ». Selon l'approche éthique considérée (utilitariste, déontologique, etc.) le degré de spécificité des règles peut bien sûr considérablement varier. Notons par ailleurs que la règle morale est définie en toute indépendance de ce qui serait de l'ordre de la punition ou de tout autre mécanisme visant à la faire respecter.

Parmi toutes les actions possibles, il faut en choisir une, l'action dite « acceptable éthiquement ». Ce choix est dévolu au processus d'acceptabilité éthique (PA), vers lequel nous nous tournons maintenant. Il est défini comme suit :

$$PA = \langle P, \succ_e, \mathcal{A}_a, EE, J \rangle$$

Le processus d'acceptabilité détermine, dans une situation concrète, l'action satisfaisant au mieux les règles morales, les désirs et les croyances de l'agent. Le processus peut juger « acceptables » des actions qui ne sont pas morales, par exemple un vol peut être jugé acceptable s'il s'agit d'un prénommé Jean qui vole un pain dans une boulangerie pour subvenir aux besoins de sa sœur et ses

sept enfants. Un tel processus se base sur un principe éthique p : il s'agit d'une fonction, basée sur une théorie du juste, qui représente une théorie éthique et qui détermine l'acceptabilité (oui/non) d'une action, prenant en compte les croyances, désirs, règles morales et valeurs dans une situation donnée. Formellement :

$$p : 2^{A} \times 2^{B} \times 2^{\mathcal{D}} \times 2^{RM} \times 2^{O_v} \rightarrow \{\top \mid \bot\}$$

Il arrive cependant qu'un principe éthique n'arrive pas à trancher. Si l'hésitation entre principes est telle que l'agent ne peut faire un choix, il est confronté à un *dilemme* : pour le résoudre, il doit faire appel à un autre principe éthique ; d'où dans la définition de PA, la présence de P, qui est une base de connaissances de principes éthiques, et $\succ_e$ qui représente une relation de préférence entre principes éthiques. Concrètement, l'agent commencera par son principe préféré, puis par ordre décroissant essayera tous les principes connus tant qu'il ne peut pas prendre un choix acceptable. L'application de la base de connaissance de principes se fait par la fonction d'évaluation éthique (EE), formalisée comme suit :

$$EE : 2^{\mathcal{A}_d} \times 2^{\mathcal{A}_r} \times 2^{\mathcal{A}_m} \times 2^{P} \rightarrow 2^{\mathcal{E}}$$

où $\mathcal{E} = \mathrm{A} \times \mathrm{P} \times \{\top \mid \bot\}$. Le parcours des principes dans l'ordre de leur préférence est le fait de la fonction de jugement J. Formellement :

$$J : 2^{\mathcal{E}} \times 2^{\succ_e} \rightarrow 2^{\mathcal{A}_a}$$

Nous voilà arrivé au bout de la définition du processus de jugement éthique (PJE). Il est temps, désormais, d'en faire usage : la première application du jugement se fait, pour ainsi dire, à la première personne : l'agent s'en sert dans un problème de choix social comme procédure de décision. Cependant, le jugement intervient aussi pour évaluer le caractère éthique du comportement des agents avec lesquels l'agent évaluateur interagit. Un tel jugement peut être aveugle, partiellement informé, ou encore pleinement informé, selon les informations dont l'agent évaluateur dispose.

Dans le cas d'un jugement pleinement informé, l'agent évaluateur a accès à l'ensemble des états mentaux de l'agent évalué : il s'agit en somme de ses désirs ($\mathcal{D}$), ses croyances ($\mathcal{B}$), ses actions (A, rappelons-nous qu'une action contient aussi ses conditions et ses conséquences), sa théorie du bien (RM et SV) et sa théorie du juste (P, $\succ_e$)[17]. Ce type de jugement, selon les auteurs, peut être pertinent lorsqu'il s'agit de vérifier la conformité d'un comportement à l'égard d'une éthique publiquement déclarée. À l'opposé, un jugement aveugle se base uniquement sur un comportement observé : l'évaluateur, dans ce cas, substitue la totalité de ses propres états mentaux à celui de l'agent évalué pour aboutir à un jugement. Au milieu, nous trouvons le jugement partiellement informé : l'agent évaluateur utilise alors les connaissances de l'agent évalué dont il dispose (quitte à ce que celles-ci soient partielles ou périmées) et supplée aux lacunes en y substituant ses propres états.

[17] Dans le modèle présenté, les éléments ontologiques, à savoir les valuations et les valeurs morales (respectivement $O_m$ et $O_v$), sont supposés communs à tous les agents et ne peuvent donc pas être échangés.

Que pouvons-nous retenir de cette cascade de définitions, au sujet desquelles il convient de se rappeler qu'elles sont une étape indispensable sur la voie de l'opérationnalisation et l'efficacité informatiques ? Analysons, pour commencer, la portée de l'élément téléologique du jugement dans ce modèle – c'est-à-dire la valeur – qui se trouve très exactement là où nous l'attendions : comme noyau de la règle morale d'une part, comme principe de résolution de conflits dans le principe éthique, d'autre part. Toutefois, il faut bien reconnaître que les prototypes fournis dans les deux articles implémentent ces aspects de façon très simpliste : ainsi pour la résolution de conflit, le premier principe éthique est celui de « l'action parfaite » : s'il y a une action qui est tout à la fois désirable, réalisable et morale, elle sera choisie. Ensuite, si un tel cas de figure n'est pas présent, tel agent peut préférer de suivre son devoir (action possible et morale, mais non désirable) ou plutôt son désir (action possible et désirable, mais amorale) ; certains agents admettront des actions désirables et immorales, d'autres non… En tant qu'inspiration des normes, la valeur se réduit à une sorte d'explicitation. Ainsi, la valeur de générosité sera dite morale si, et seulement si, elle s'exerce en faveur d'un agent pauvre :

*moral(robin_hood,A,X,B):-generous(A,X,B), poor(B), action(X).*

S'il y a cependant un point où les auteurs s'écartent sensiblement de Ricœur – dont ils se réclament pourtant – c'est bien dans le maintien dualiste d'une sphère normative, d'une part, et d'une sphère descriptive, d'autre part : vérité et fausseté sont prédicables des croyances qui, seules, nous informent de l'état du monde. Ricœur refuse résolument un tel partage et fait intervenir la téléologie jusque dans la description que l'agent a de soi, plus exactement encore dans l'interprétation de celle-ci, comprise comme une mise en histoire de sa vie. Or une histoire échappe à la dichotomie du fait et de la valeur, en tant qu'elle fait intervenir des opérations pragmatiques telles que le point de vue, l'agencement de l'intrigue, etc.

Ce qui est encore plus frappant, c'est que la démarche est strictement descendante : le jugement des actions procède unidirectionnellement à partir des règles, principes, etc., définis au préalable. Cette impression se renforce encore lorsqu'on observe que dans les prototypes, le processus de reconnaissance des situations – en charge de générer les désirs et les croyances de l'agent – n'est simplement pas prévu : les états mentaux sont *donnés d'emblée* et se mettent seulement à jour grâce à l'observation ou au moyen de la communication. À partir de là, le processus peut se dérouler de façon strictement déductive. Nous pouvons donc dire que, contrairement à celle que nous avons esquissée dans l'introduction, cette démarche-ci n'est pas topique. Elle ne saurait d'ailleurs pas l'être, car une approche topique n'est pas possible en recourant exclusivement à des méthodes symboliques : en termes informatiques, nous pourrions dire que cela supposerait une boucle de rétroaction entre la valeur et les croyances, qui renforcerait le relief ou la pertinence des faits les plus significatifs au regard de la valeur retenue.

Notons pour finir l'absence totale de la dimension émotionnelle. La démarche est donc strictement rationaliste ; mais là encore, il serait difficile dans une démarche strictement symbolique de faire autrement. Il n'en demeure pas moins que, malgré les quelques réserves que nous ayons pu émettre, nous mesurons ici toute la distance parcourue depuis les travaux de Danielson, dont la moralité artificielle comprenait elle aussi un jugement sur l'action d'autrui, mais de la façon la plus

élémentaire qui soit : l'agent évaluateur se contentait d'exécuter, pour son propre compte, la procédure de décision de l'agent évalué et de baser son choix sur le résultat de cette exécution[18].

### 3.2.3. Réputation et confiance

Une fois que nous nous sommes doté d'une procédure de jugement, nous sommes à même de nous former une *image* de l'éthique d'autrui. Dans les travaux en SMA, cela passe par l'idée de la *réputation* des agents[19]. La réputation y est vue comme une réalité sociale, qui peut affecter non seulement un agent mais aussi une organisation. Elle implique la *communication* entre agents de leurs images. Sur base de la réputation, l'agent peut alors décider d'accorder sa *confiance*.

Commençons cependant par la construction des images. Dans le contexte BDI que l'article examiné propose[20], une image d'autrui est modélisée comme une croyance. Pour se construire une telle image, un agent peut se fonder sur l'interaction directe – et alors il peut faire jouer sa procédure de jugement – ou il peut se baser sur une image construite – et communiquée – par un tiers qui a déjà gagné notre confiance ; car c'est bien toujours de confiance qu'il s'agit.

Les auteurs distinguent résolument entre deux types de confiance : d'une part, la confiance qu'un agent peut accorder à un autre de se conformer à un ensemble de règles morales ; d'autre part, la confiance peut porter sur l'acceptabilité éthique des actions de l'autre agent. Nous retrouvons donc les mêmes catégories que dans le jugement éthique. La confiance morale est toujours relative à un ensemble de règles R. Ainsi, un agent $a_i$ peut estimer qu'un agent $a_j$ est conforme à l'ensemble des règles qui touchent à la valeur de l'honnêteté, mais non pas à celles qui régissent un comportement généreux. D'où la formalisation suivante de la confiance morale :

$$confiance_morale(a_j, a_i, R, sc).$$

où la valeur *sc* est le *seuil de confiance* : comme la conformité à une règle morale fait l'objet d'une valuation plurivalente (par exemple : moral, immoral, amoral), le seuil de moralité indique la valuation minimale à obtenir pour être qualifié de conforme. La confiance morale pour un ensemble de règles R se base sur l'image de moralité pour cet ensemble R :

$$image_moralité(a_j, a_i, cv, R, sc, t_0, t).$$

[18] P. DANIELSON, *Artificial Morality*, pp. 74-78. Il y a, certes, quelques complications : ainsi un agent peut refuser l'accès à sa procédure de décision (mais dans les faits, ceci revient à signaler de mauvaises intentions par rapport à un interactant éventuel). Pour finir, il importe de signaler que les agents de Danielson étaient écrits en Prolog, alors que ceux de Cointe et ses collègues le sont en ASP, langage sur lequel nous aurons à revenir en abordant l'épreuve que constitue la norme (§ 3.3.2).

[19] Voir le chapitre de J. SABATER-MIR et L. VERCOUTER, *Trust and Reputation in Multiagent Systems*, dans G. WEISS, *Multiagent Systems*, pp. 381-419.

[20] Pour la discussion des images, nous nous référons à l'article de N. COINTE, Gr. BONNET et O. BOISSIER, *Coopération entre agents autonomes fondée sur l'éthique*.

où cv est le niveau de conformité. Notons qu'une image de moralité est définie sur l'intervalle de temps allant de $t_0$ à t.

La confiance éthique, en revanche, est unique d'un agent à un autre :

$$confiance_éthique(a_j, a_i).$$

Elle se base sur l'image correspondante, exprimée formellement comme suit :

$$image_éthique(a_j, a_i, cv, t_0, t).$$

Le prototype que les auteurs proposent, simule l'activité boursière d'achat et revente d'actions. Dans ses opérations d'achat, un agent peut être guidé par des valeurs : par exemple, la valeur environnement avec ses sous-valeurs que sont *la transparence dans le rapportage environnemental*, *la lutte contre le réchauffement climatique* ou encore *l'énergie renouvelable*. L'idée qui préside au choix de cet exemple en particulier est que des dilemmes peuvent surgir dès lors que l'agent cherche à appliquer ses valeurs. Ainsi un fournisseur d'énergie qui produit de l'énergie nucléaire pourrait enfreindre les règles liées à l'énergie renouvelable, mais être tout à fait en règle en matière de rapportage et de réchauffement climatique. Les agents doivent alors faire appel à leurs principes éthiques pour résoudre de tels dilemmes. L'exemple s'avère dès lors doté d'une grande force discriminatoire entre différentes combinaisons de règles morales et principes éthiques.

Le prototype est implémenté dans une technologie que nous avons déjà eu l'occasion de présenter au deuxième chapitre (§ 2.2.2), la plate-forme de programmation orientée agent Jason, qui permet de modéliser des agents BDI en Java. La construction d'images se fait ici grâce à un *plan* de Jason, composé des actions suivantes : *évaluer la conformité des comportements*, *construire une image morale*, *mettre à jour la confiance*. Ainsi pour un ensemble de règles morales, l'agent peut évaluer la conformité de chaque échange d'actifs en faisant appel à ses connaissances factuelles (par exemple telle société a une certification énergétique) et sur ses supports de valeur, qui associent des valeurs à des actions dans l'état courant du monde : « échanger des actifs d'une société certifiée EMAS répond à la valeur de rapportage environnemental ». Ensuite, il est possible de construire l'image morale sur la base des actions ainsi jugées, éventuellement assortie d'une pondération différenciée. Finalement, si l'image morale satisfait le seuil concerné, l'agent peut mettre à jour sa confiance par rapport aux agents impliqués dans les actions.

Les auteurs ont raison d'insister sur le fait que, malgré un usage intensif d'états internes, la construction de la confiance se limite à prendre en compte des comportements effectivement observés, sans tenir compte d'éventuelles intentions. En cela – et malgré le cadre BDI – l'approche reste prudemment fonctionnelle. Toutefois, la même remarque que *supra* prévaut : toutes les valeurs, toutes les règles morales, sont déjà données. En outre, l'article laisse dans l'ombre comment des images pourraient être *communiquées*.

La communication est, en revanche, primordiale dans le cas de la réputation à proprement parler, en tant que réalité sociale. Nous allons voir cela de plus près dans un article qui y est explicitement

consacré[21], où la réputation est modélisée comme un *artéfact* : nous avons déjà abordé cette notion lorsque nous avons parlé de la modélisation de l'environnement à l'aide de Cartago (§ 2.2.2). Pour rappel, Cartago est une extension de Jason qui permet de donner – toujours à l'intérieur de la programmation orientée agents – une consistance à l'environnement en réifiant certaines fonctionnalités sous la forme d'artéfacts. Ici, les auteurs mobilisent ORA4MAS – une plate-forme elle-même basée sur Cartago – dont les artéfacts exposent deux *interfaces* : une interface d'usage (*usage interface*), qui renseigne les agents sur les opérations qu'ils peuvent déclencher sur l'artéfact. Cette interface expose également une série de propriétés consultables par les agents. La deuxième interface est *l'interface de liage* (*link interface*) qui fournit des opérations exploitables par d'autres artéfacts de manière à offrir des fonctionnalités composées.

Dans le prototype que les auteurs proposent, il y a trois types d'artéfacts : des artéfacts de groupe, des artéfacts de schème et un unique artéfact de réputation. Commençons par le type le moins problématique : l'artéfact de groupe. Chaque artéfact de ce type permet de consulter les membres du groupe que l'artéfact « représente », de demander une adhésion à ce groupe, etc. Puis, les artéfacts de schème représentent des missions : ils exposent un arbre détaillant les étapes des missions (les buts), les agents affectés à la mission, etc. L'artéfact de réputation, pour finir, est relié à tous les autres artéfacts, et peut être consulté à tout moment par l'ensemble des agents. En simplifiant quelque peu, nous pouvons dire que les artéfacts de groupe l'informent sur des violations de normes obligatoires (liées aux rôles que les agents occupent dans leur groupe), et les artéfacts de schème l'informent sur l'accomplissement de résultats de missions. L'artéfact de réputation quantifie alors ces informations afin d'évaluer chaque agent.

Notons le caractère global de l'évaluation : elle présuppose une connaissance parfaite ; elle n'est située ni dans le temps, ni dans l'espace. Il s'agit, pour ainsi dire, d'un bulletin. Il faut également dire que ce bulletin, n'est pas, aux yeux des auteurs, la réputation elle-même :

> *It is important to notice that the evaluation is not the reputation of the agent […] reputation is a shared voice circulating in a group of agents. The reputation artifact is indeed an instrument to influence the reputation of the agent*[22]*.*

La réputation, en effet, se construit en articulant une (ou des) évaluation(s) comme décrites dans l'article, et les images personnelles de l'éthique d'autrui que les agents se communiquent entre eux. La dimension communicative n'est toutefois pas abordée dans l'article ; elle reste d'ailleurs une question de recherche ouverte[23]. Beaucoup reste donc à faire avant qu'une telle réification puisse espérer éclairer utilement la dynamique réputationnelle ; des éléments de réponse pourraient cependant être obtenus en regardant du côté d'un champ connexe, celui de la négociation entre agents.

---

[21] Nous nous basons ici sur l'article de J. HÜBNER, L. VERCOUTER et O. BOISSIER, *Instrumenting Multi-Agent Organisations with Reputation Artifacts*.
[22] *Ibid*., p. 22.
[23] J. SABATER-MIR et L. VERCOUTER, *loc. cit.*, pp. 398-402.

### 3.2.4. SMA et négociation

Nous avons déjà rencontré la notion de négociation à plusieurs reprises au cours de ce mémoire, notamment en abordant la question de la négociation autour d'un système de valeurs (§ 1.8.3). C'est le même enjeu – la négociation de la valeur – qui nous retiendra ici : ses implications lors de la catégorisation de certains actes, ses limites d'applicabilité, son rôle dans la justification d'une action entreprise. Or force nous est de constater que l'approche de la négociation qui domine aujourd'hui les systèmes multi-agents et celle de la théorie des jeux. Après un bref rappel des limites de cette approche, nous nous attacherons dans ce paragraphe à explorer des alternatives.

La théorie des jeux, lorsqu'elle s'applique à la négociation, se limite au marchandage de ressources (*bargaining*) et aux enchères (*auctions*). Même dans ce cadre circonscrit, les technologies basées sur la théorie des jeux ont certaines insuffisances, plus particulièrement les présuppositions d'une connaissance parfaite et d'une disponibilité illimitée de capacité de calcul. Pour la problématique qui nous retient ici, la négociation des valeurs, son inconvénient principal est de partir du principe que l'optimum à atteindre est *connu d'avance*, ce qui le rend inapte à traiter des cas où l'agenda optimal de l'agent n'est pas fixé d'emblée[24]. L'exemple donné est le cas – relativement simple – d'un agent qui veut acheter un certain nombre de voitures *g*. Si les enchères proposent m voitures (où *m* est plus grand que *g*), l'agent doit d'abord se faire une idée de quelles voitures acheter avant d'entamer les négociations à proprement parler. Eu égard aux enjeux qui sont les nôtres ici et qui regardent au-delà de la gestion des ressources, un problème supplémentaire se fait jour : la *valeur* est toujours déjà *donnée* et supposée *homogène* : qu'il s'agisse de vies ou d'années de prison, c'est toujours le même paradigme d'un même optimum à atteindre, optimum *connu d'avance*.

Il faut donc partir à la recherche d'une autre base pour fonder la négociation entre agents. Le premier candidat qui se présente est alors la négociation argumentative (*argumentation-based negociation*) : celle-ci a pour point de départ un système d'argumentation abstrait de Dung[25]. Un tel système d'argumentation $\mathcal{S}$ peut être modélisé par un graphe orienté où les nœuds sont des arguments et les arcs sont des relations d'attaque. Ainsi l'arc *(q,p)* signifie que q *attaque* p, c'est-à-dire que le fait d'accepter l'argument *q* revient à rejeter l'argument *p*.

Une fois un tel système posé, il s'agit alors de déterminer un *ensemble d'arguments* ou une *position* P qui peut être dite *raisonnable*. Une première méthode est alors celle dite des extensions préférées. Pour prétendre à la qualification d'extension préférée, une position doit satisfaire plusieurs critères. Premièrement, elle doit être dépourvue de conflits internes, c'est-à-dire qu'un argument de P ne peut en attaquer un autre. Deuxièmement, la position doit être défendable : chaque argument de $\mathcal{S}$ qui attaque un argument de P doit lui-même être attaqué par un autre argument élément de P. Une fois que nous disposons d'un ensemble d'arguments qui répond à ces deux critères, nous pouvons dire que notre position est *admissible*. Pour être une extension préférée, il faut encore qu'il soit impossible d'ajouter aucun argument à P sans que P perde son admissibilité.

[24] Voir le chapitre de Sh. FATIMA et I. RAHWAN, *Negotiation and Bargaining*, dans G. WEISS, *op. cit.*, pp. 143-176.
[25] Nous nous référons ici au chapitre *Arguing* de M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 337-354.

Cette méthode de déterminer une position raisonnable a cependant l'inconvénient d'admettre une multiplicité de positions. Une définition plus stricte alors est celle dite de l'extension fondée (*grounded extension*) : contrairement aux extensions préférées, elle est unique. Elle regroupe en son sein l'ensemble des arguments inattaquables du système d'argumentation. Voilà comment l'extension fondée est calculée :

1. d'abord, ajouter à P tous les arguments qui n'ont aucun attaquant dans $\mathcal{S}$ ;
2. ensuite, éliminer de $\mathcal{S}$ tous les arguments attaqués par les arguments de P.

Après la deuxième étape, il y aura de nouveau des arguments sans attaquants dans $\mathcal{S}$. Il suffit dès lors de boucler sur les étapes 1 et 2 jusqu'à ce que la position devienne stable.

Tout ce qui précède fait fi du *poids relatif* des arguments entre eux. Or il est possible de *hiérarchiser* les arguments. Il est alors impossible pour des arguments plus bas dans la hiérarchie d'attaquer des éléments plus élevés. Cette hiérarchie, cependant, est globale et statique. Il est également possible d'enrichir le système d'argumentation en donnant un poids relatif à une audience donnée. À chaque argument sera associée une valeur $\eta$ qui indique le poids relatif de l'argument pour l'audience A. Nous pouvons ainsi écrire :

$$\eta(b) \rhd_A \eta(a)$$

Ce qui exprime simplement que pour l'audience A, l'argument b a plus de valeur, plus de poids, que l'argument a.

Un tel système d'argumentation – dont nous n'avons donné qu'un exemple très simple – peut servir de base à des *dialogues* ou des négociations argumentatives. Dans l'article qui nous inspire ici[26], le système est en effet d'une bien plus grande complexité, nous n'en ferons donc pas le tour. Nous nous bornerons simplement à indiquer que la négociation se base sur des arguments, des connaissances, des buts (ou des préférences) des agents, ainsi que sur une série *d'actes de négociation* : l'acte de base pour un agent est la *proposition d'une offre*. Les autres agents peuvent y répondre en *l'acceptant*, en la *refusant* ou en la *défiant*. Si une offre est refusée, elle est retirée de l'ensemble de solutions possibles. Si, en revanche, elle est défiée, l'agent qui l'a émise est sommé de s'expliquer en fournissant des arguments en faveur de son offre. Les arguments qu'un agent avance, peuvent inciter ses interlocuteurs à modifier leurs croyances, les amenant de la sorte à réviser l'acceptabilité de l'offre émise. Nous n'en dirons pas plus sur cet article, qui va assez loin dans le raffinement : les connaissances sont assorties d'un degré de certitude, les buts des agents assortis d'un degré de priorité, le tout se combinant pour calculer la force des arguments avancés pour ou contre leurs buts, etc. Soulignons que les agents peuvent, dans une mesure limitée certes, changer d'avis : des mises à jour de croyance sont en effet possibles.

Cependant, les buts sont le seul mécanisme qui sous-tend la négociation : les agents ne peuvent en changer. Le mécanisme des valeurs que nous avons vu précédemment peut certes supplier à mieux

[26] Cf. L. AMGOUD, S. BELABBES et H. PRADE, *Towards a Formal Framework for the Search of a Consensus between Autonomous Agents*.

fonder les buts. Cependant, nous sommes toujours confronté à une insuffisance dudit mécanisme : le formalisme interdit l'intercommunicabilité des valeurs ! Or la valeur, en argumentation, est l'objectivation d'une passion ; c'est elle qui rend une discussion sur les préférences possibles et par là même argumentables[27]. Une piste – que nous ne pouvons ici qu'esquisser très rapidement – pourrait être le travail[28] fait sur des préférences qualitatives à attributs multiples (*qualitative multi-attribute preferences*). Les préférences, ici, sont comme des critères qui déterminent si un but a été atteint ou non. Ces critères sont pris en compte par ordre lexicographique, c'est-à-dire qu'ils sont parcourus jusqu'à ce qu'un niveau de priorité donne une solution ; les critères de priorité inférieurs ne sont alors même plus considérés. En d'autres termes, les buts à atteindre peuvent être motivés par des valeurs elles-mêmes comparables entre elles. Puisque ces attributs-là sont qualitatifs, ils se laissent facilement décrire dans un formalisme logique, et par là l'échange et la comparaison de ces préférences deviendraient envisageables. Mais beaucoup, ici encore, reste à faire, à commencer par l'articulation de ce type de préférences avec les valeurs qui sont censées les justifier.

### 3.2.5. Les organisations et leur éthique

Terminons en invoquant le programme de recherche sur les éthiques collectives esquissé dans un article[29] qui fait suite à celui sur le jugement moral (exposé dans § 3.2.2). Rappelons en deux mots la théorie dont il s'agit : la prise en compte de l'éthique d'autrui commence dès la reconnaissance de situation. Que ce soit par induction sur l'observation du comportement des autres agents, ou via la communication de règles, de croyances, etc., l'agent doit pouvoir reconstruire le modèle de raisonnement de son entourage. Une fois l'éthique d'un autre agent acquise, elle peut être utilisée dans la décision, que ce soit pour estimer la similarité entre deux éthiques ou pour jauger la conformité d'un comportement par rapport à une éthique déclarée.

Une éthique collective doit être comprise comme l'éthique – un ensemble de règles morales et de principes éthiques, comme nous l'avons exposé plus haut – dont est pourvue une structure éphémère ou (plus ou moins) pérenne ; si elle est éphémère, les auteurs parlent d'une *coalition*, sinon une structure qui s'inscrit dans la durée sera dite une *organisation*. L'éthique collective peut être d'emblée explicitement définie ou être le résultat d'une interaction entre agents.

Dans le cas où l'éthique collective est le résultat d'une interaction, les auteurs distinguent encore entre des mécanismes de constructions *explicite* et *implicite* (ou *émergent*). Une éthique collective implicite n'existe que par la similarité des éthiques individuelles des membres du collectif. En revanche, si elle est explicite, elle sera le plus souvent *imposée* aux membres, quitte à ce qu'elle soit le fruit d'un processus d'agrégation, par exemple au moyen d'une argumentation – nous retrouvons ici le thème de la négociation vu dans la dernière section. Dans tous les cas de figure, le résultat peut

[27] M. MEYER, *Principia Rhetorica*, p. 195.
[28] Nous nous fondons ici sur l'article de W. VISSER, K. HINDRIKS, C. JONKER *Argumentation-Based Preference Modelling with Incomplete Information*.
[29] N. COINTE, Gr. BONNET et O. BOISSIER, *Éthique collective dans les systèmes multi-agents*.

réserver des surprises importantes, car il est tout à fait possible que l'éthique collective construite soit contradictoire par rapport à une ou plusieurs éthiques individuelles qui l'ont pourtant inspirée.

Sur ces bases, le programme de recherche que les auteurs esquissent est vaste. Ainsi les agents peuvent avoir des *appartenances multiples*, comme c'est le cas dans des coalitions *recouvrantes*. Un agent, interne ou externe au collectif, doit pouvoir identifier l'éthique du collectif ; par exemple, par observation du comportement des membres du collectif. Un agent interne au collectif doit donc avoir des moyens pour résoudre des conflits entre éthique individuelle et collective, ou entre ces différentes éthiques collectives. Dans son jugement de l'éthique d'autrui, un agent peut tenir compte des éthiques auxquelles souscrit l'agent avec qui il interagit. Des agents observant une organisation dont l'éthique ne leur convient pas peuvent faire sécession et créer leur propre organisation dont l'éthique sera dérivée de la première mais différente toutefois sur des points jugés importants. Finalement, l'exploration des moyens dont dispose un collectif pour faire respecter son éthique est aussi une piste intéressante. Tout cela, il est vrai, reste entièrement au stade programmatique. Il n'en demeure pas moins que les auteurs proposent un programme de recherche dans lequel le processus de jugement éthique – dûment formalisé – revêt un caractère *structurant* dans les échanges et les interactions entre agents. L'éthique – dans une mouture certes rationalisée – est ainsi placée au cœur de la vie des institutions et de leurs membres.

## 3.3. La déontologie en SMA

Dans cette section, nous abordons la norme, là où le règne de la logique déontique est souverain. Comme nous l'avons fait pour la section téléologique, nous procéderons ici en respectant, autant que faire se peut, la présentation de Ricœur, remontant la hiérarchie des normes en partant du respect de soi, où doit s'exercer l'autonomie ou l'universalité du vouloir. Ensuite, sur le plan interpersonnel, où le respect d'autrui exige de considérer la personne comme une fin en soi, nous aborderons toutes les normes que des agents en interaction peuvent – ou non – observer. Enfin, sur le plan impersonnel s'impose le principe de justice réparatrice : nous y serons amené à examiner comment une organisation, en SMA, peut s'employer à faire respecter les normes émises pour protéger son pendant téléologique, la justice distributrice.

Avant de continuer notre exposé, insistons sur le fait que cette partie restera entièrement à l'intérieur de l'épreuve de cohérence qu'impose la norme ; nulle part nous ne ferons appel à un principe téléologique ni ne sortons du formalisme. Ainsi, pour citer l'exemple que nous traiterons plus loin en quelque détail, celui du *mensonge opportun* ou du *droit de mentir*, une résolution du conflit pourrait s'envisager en inscrivant l'exigence de la vérité dans son effet sur le bonheur de la personne à qui nous mentons ou à qui nous disons vrai. Telle est, en tout cas, la leçon de Ricœur[30]. Elle ne sera cependant pas à l'ordre du jour, pour la simple raison que cette idée ne semble pas effleurer les auteurs consultés.

[30] Voir P. RICŒUR, *Soi-comme comme un autre*, pp. 307-314.

### 3.3.1. L'internalisation de la norme

L'expression la plus pure de la volonté est *l'obéissance* ; cette affirmation, qui se présente de prime abord comme un paradoxe, suit en vérité une logique toute kantienne. En effet, Kant entend par « volonté » le désir qu'a tout être humain de se conformer aux lois ; or la meilleure manière d'obéir aux lois, c'est de se les donner comme lois personnelles. L'internalisation des lois est donc une forme d'autolégislation ou d'autonomie ; procédé que Ricœur résume bien en disant que « l'obéissance véritable, c'est l'autonomie »[31]. En d'autres termes et toujours selon Ricœur, la norme à la première personne est capitale dans la conception kantienne de l'autonomie.

L'internalisation des normes trouve un lieu d'exploration propice dans des architectures dérivées de l'approche BDI. Il est ainsi possible[32] d'enrichir le modèle BDI standard avec des croyances et buts *normatifs*. Parmi les croyances normatives, il faut distinguer entre les croyances principales, les croyances d'extension et les croyances de sanction. La croyance normative principale (*main normative belief*) édicte le contenu déontique : une action est permise, défendue ou obligatoire dans un contexte donné. La croyance normative d'extension (*normative belief of pertinence*) associée à la croyance principale précise à quel ensemble d'individus celle-ci s'applique. La croyance de sanction (*norm enforcement belief*) dit les sanctions, positives ou négatives, qu'entraîne respectivement la conformité à, ou la transgression de, la croyance principale. Les buts normatifs sont des buts internes greffés sur une croyance normative ; si celle-ci vient à disparaître, le but qui se fonde sur elle ne lui survit donc pas.

Un agent normatif, dans ce modèle, peut alors être défini comme un agent qui est au courant de ces trois types de normes et qui, quand il doit générer ses buts et planifier ses actions, est capable de prévoir l'effet des sanctions et d'en tenir compte pour calculer le gain escompté des actions possibles. Lorsqu'un tel agent obéit à la norme tout en se passant de la prise en compte de la sanction, nous pouvons dire qu'il a *internalisé* la norme : il croit qu'il y a une norme (croyance principale), et qu'il fait partie des individus auxquels la norme s'applique (croyance d'extension) ; cela lui suffit pour se conformer. L'internalisation de la norme, dans ce modèle, se construit sur deux piliers. Le premier pilier est ce que les auteurs appellent le principe du calcul parcimonieux : tous les calculs que nous venons de décrire impliquent en effet une charge cognitive. En d'autres termes, l'obtention et le traitement de l'information entraînent des *coûts*. L'obéissance à la norme sans reste permet alors de faire des économies en matière de temps de décision et d'exécution. Le deuxième pilier est le *relief* (*salience*) de la norme : ce pilier prend en compte les observations que fait l'agent sur l'importance de la norme dans la société ; la notion renvoie au rôle du contrôle social exercé par rapport à une norme, au taux d'observance, etc. ; nous verrons plus bas comment elle a été opérationnalisée.

Le prototype que les auteurs proposent se présente comme une simulation à base d'agents[33] dans laquelle il y a trois types d'agents : des agents stratégiques, des agents internalisateurs, ainsi que des

[31] P. RICŒUR, *Soi-même comme un autre*, p. 245.
[32] Voir l'article de G. ANDRIGHETTO, D. VILLATORO et R. CONTE, *Norm internalization in artificial societies*.
[33] Il faut dire que les auteurs sont avares d'indications sur l'implémentation de leur modèle. Le moteur BDI utilisé est « fait maison » : EMIL-A, dont nous trouvons une présentation – dans les grandes lignes – dans l'article de G. ANDRIGHETTO,

éducateurs. Le premier type d'agents sont les agents stratégiques : ils ignorent l'existence de normes et se contentent de calculer les bénéfices. Cependant, ils prennent la norme en compte de façon implicite, car leur calcul intègre les punitions qui leur ont été infligées dans le passé suite à des comportements indésirables. Le deuxième type d'agents sont les éducateurs. Ils se comportent comme les agents normatifs décrits plus haut. Ils ont la particularité d'être, en début de simulation, les seuls à déjà disposer d'une connaissance de croyances normatives. Ces agents ont beau connaître l'existence des normes et de leurs sanctions, ils n'ont aucune connaissance a priori de la probabilité de survenance d'une punition. Ceci dépend du degré de surveillance, dont ils peuvent se faire une idée grâce à l'observation de l'environnement. Ils prennent alors leurs décisions en comparant le degré de surveillance au degré de tolérance au risque, propre à chaque agent.

Le troisième type d'agents sont les internalisateurs. Un internalisateur commence comme un agent normatif du type « éducateur », à ceci près qu'il n'a aucune connaissance initiale sur les normes. Au lieu de cela, il est doté d'un mécanisme de relief de la norme. Le calcul du relief fait intervenir non moins de 12 facteurs que nous simplifions en trois catégories : le nombre de défections observées, le nombre de punitions observées et le nombre de messages éducatifs reçus. Un message éducatif accompagne la sanction donnée par un agent normatif (éducateur ou internalisateur). Il contient une référence explicite aux normes sur lesquelles l'agent s'est basé pour infliger sa punition, fournissant ainsi à l'agent puni une occasion d'apprentissage.

La simulation se déroule dans une topologie sociale, ou chaque agent ne peut interagir qu'avec ses proches voisins. Lors de chaque pas de temps de la simulation, les agents jouent une sorte de dilemme des prisonniers avec un voisin choisi aléatoirement. Ce jeu comporte deux phases, dont la première est classique : chaque agent décide de coopérer ou de faire défection. Lors de la deuxième phase, les agents qui ont subi une défection lors de la première peuvent infliger une sanction. Elle peut être assortie d'une évaluation normative qui précise quelle norme a été violée.

À l'issue d'un jeu, chaque agent met à jour ses croyances. Chaque agent normatif intègre ainsi les messages déontiques concernant des normes qu'il ne connaissait pas. À partir de là, les chemins des éducateurs et des internalisateurs bifurquent : les éducateurs mettent simplement à jour leurs croyances de punition en calculant le ratio entre le nombre de défecteurs impunis et le nombre total des défections. Comme nous venons de le voir, le mécanisme des agents internalisateurs est beaucoup plus complexe, puisqu'ils recalculent le relief pour chaque norme dont ils savent l'existence. L'internalisation de la norme se fait quand le relief atteint un certain seuil et que le calcul coût-bénéfice dépasse une certaine valeur. À ce moment, nous pouvons dire que le respect de la norme l'a emporté, c'est-à-dire qu'à l'avenir l'analyse cout-bénéfice ne sera plus effectuée. La conformité est devenue un automatisme. Il faut insister que même lors de l'application automatique d'une norme, son relief continuera à être évalué : si celui-ci tombe en-dessous du seuil, la flexibilité reprend ses droits.

---

M. CAMPENNÌ, R. CONTE et M. PAOLUCCI, *On the Immergence of Norms*. Plus frappant encore, la plate-forme de simulation utilisée pour le prototype n'est même pas mentionnée dans l'article. Nous ne pouvons que conjecturer qu'il s'agit d'EMIL-S, qui a été conçu pour simuler l'évolution d'agents EMIL (voir le site web du projet de recherche EMIL, financé par la Commission Européenne : http://emil.istc.cnr.it/).

Dans le prototype, le coût des sanctions est fixe, que ce soit pour les infliger ou pour les recevoir. Le degré de risque est fixé à 30% en début de simulation. Le nombre d'éducateurs est constant aussi, toujours égal à 10. Le nombre total d'agents est toujours 100. Les paramètres variables sont essentiellement la proportion entre agents stratégiques et agents internalisateurs ainsi que la distribution des punitions. Cette distribution des punitions reflète, de façon agrégée, les conditions environnementales. Elle n'influence pas la prise de décision des agents (sanctionner ou pas), mais bien l'effectuation des décisions. Même si les auteurs restent assez flous sur ce qu'une telle distribution veut réellement dire, l'idée qui a présidé à son introduction est d'éprouver la stabilité de la norme lorsque des changements structuraux dans l'environnement surviennent. Ainsi, dans la distribution « *step down* », les sanctions disparaissent brusquement après un certain nombre de pas de simulation. Le résultat prévu – et effectivement observé – est que les internalisateurs résistent plus longtemps à la disparition de la norme que les autres types d'agents. En jouant sur le nombre d'internalisateurs et les distributions de sanctions, il devient alors possible d'explorer des questions liées entre autres à la stabilité de la norme.

Le prototype appelle certainement quelques réserves : commençons par regretter l'implémentation de la topologie sociale : en gros, il s'agit ici d'un graphe total et statique : total, puisque tous les agents sont voisins de tous les autres ; statique, car le graphe n'est jamais mis à jour. Il s'agit pour ainsi dire d'une non-topologie, qui fait perdre un bénéfice majeur de la simulation à base d'agents et qui est son inscription dans l'espace. En outre, même si nous comprenons l'idée derrière les distributions de punitions, nous restons sur la désagréable impression qu'il s'agit d'un passe-droit, qui dissocie les décisions de sanction de la prise de sanction effective. Nous voyons ainsi poindre de nouveau cette question lancinante du *qui ? qui sanctionne ?* et l'intention collective redevient cette boîte noire que la SBA est pourtant censée éclairer. Dernière limitation : étant donné que le nombre d'agents normatifs non-internalisateurs reste constant, les auteurs se sont interdits par là même d'explorer un enjeu majeur : la morale est-elle affaire de pure connaissance ? Suffit-il de connaître la norme ou, comme le font les internalisateurs, faut-il y adhérer en en faisant un but digne d'être poursuivi pour lui-même ? Finalement, nous devons également regretter d'avoir opté pour une variante du dilemme des prisonniers pour illustrer le modèle : tout but reste ainsi subordonné à la maximisation chiffrée d'un bénéfice abstrait. Il y a donc une téléologie, mais implicite et muette sur ses enjeux.

Malgré ces quelques limitations d'ordre surtout technique, Il faut bien garder à l'esprit qu'un tel modèle complexifie grandement le rapport à la norme, comparé aux approches qui ne considèrent que la sanction prise comme un critère d'optimisation sans plus. Les auteurs font donc droit à l'inspiration toute kantienne qui est la leur :

> *Sub-ideally, norms are often complied with because they are enforced by a system of sanctions. But ideally, they are meant to be observed because [they] are norms and should be complied with for their own sake.*[34]

Nous ne voudrions pas conclure cette section sur l'internalisation des normes sans faire mention d'une autre architecture BDI prometteuse dans cette lignée de recherche. Il s'agit du n-BDI. Pour une

[34] G. ANDRIGHETTO, D. VILLATORO et R. CONTE, *Norm internalization in artificial societies*, p. 328.

présentation détaillée, nous devons renvoyer le lecteur vers les références en note infrapaginale[35] ; nous n'en retiendrons ici que quelques traits pertinents pour notre discussion. Le point de départ de cette architecture est d'introduire des contextes (ou unités) multiples dans la palette cognitive des agents. La véritable nouveauté de ceci est que chaque contexte peut définir son propre langage logique, tant sur le plan syntaxique que sémantique. Ainsi, le contexte des croyances recourt à un langage propositionnel ; le contexte des désirs à un langage modal. Les deux contextes adoptent une sémantique floue, c'est-à-dire que les valeurs de vérité (le vrai et le faux) sont remplacées par un nombre décimal allant de 0 à 1[36]. Dans les cas des croyances, cette sémantique reflète le degré de certitude, permettant ainsi de modéliser des « croyances » au sens commun du terme ; dans le cas des désirs, elle reflète l'intensité. Citons encore, pour mémoire, que l'architecture connaît aussi le contexte de planification - qui n'est rien d'autre que la bibliothèque des plans – le contexte de communication, éventuellement un contexte de réputation.

Nous voudrions surtout nous arrêter ici sur les deux contextes normatifs de l'architecture que sont le contexte d'acquisition des normes (NAC, pour *norm acquisition context*) et le contexte de conformité aux normes (NCC, pour *norm compliance context*). Le contexte NAC peut être qualifié de « base de données » des normes : toutes les connaissances qu'a un agent sur les normes y sont stockées, y compris le relief – au sens que nous avons déjà vu – qu'a pris chaque norme. Pour être tout à fait précis, le relief de la norme constitue la sémantique de ce contexte. Le contexte de conformité aux normes s'intéresse, quant à lui, aux normes instanciées (*norm instances*), c'est-à-dire les normes dont l'agent estime qu'elles s'appliquent effectivement à lui. Ici encore, la sémantique est floue et dénote la *pertinence* de la norme.

Il faut savoir gré à cette architecture d'être attentive au passage d'un contexte à l'autre. Ainsi il n'est possible d'instancier une norme, c'est-à-dire faire entrer une norme dans le contexte de conformité, qu'en conjuguant la connaissance normative du contexte NAC avec des croyances personnelles de l'agent, stockées dans le contexte des croyances. Dans ce modèle, l'internalisation de la norme se fait donc via des règles de passage normatives. Les auteurs – tout en mettant en garde que la liste pourrait s'allonger – voient trois règles de passage d'internalisation des normes, qui toutes impliquent une mise à jour du contexte des désirs à partir du contexte NCC. La règle d'obligation crée un désir positif d'atteindre l'objet de l'obligation et symétriquement, la règle de prohibition crée un désir négatif ; l'internalisation des permissions, quant à elle, ne crée pas de nouveaux désirs, mais restreint le champ d'application des autres normes.

Avec l'architecture du n-BDI, nous nous détachons donc formellement de la norme comme pure connaissance, mais – il faut l'avouer– le prix à payer pour ce résultat est élevé. En effet, il n'implique rien de moins que le dédoublement de la cognition entre « les faits » et « les normes », le descriptif d'une part et le prescriptif, d'autre part. Les auteurs y voient l'avantage de la flexibilité ; nous devons répondre que nous ne voyons pas pourquoi une croyance normative devrait être séparée d'une

[35] Voir l'article de N. CRIADO, E. ARGENTE, P. NORIEGA et V. BOTTI, *Towards a Normative BDI Architecture for Norm Compliance*, ainsi que la thèse de doctorat de N. CRIADO, *Using Norms to Control Open Multi-Agent Systems*, particulièrement les pages 77-109.

[36] Une présentation très abordable de la logique floue se lit chez V. MATHIVET, *L'Intelligence Artificielle pour les développeurs*, pp. 95-123.

croyance « autre », dès le moment où l'architecture utilisée permet de mettre à jour les croyances et leur degré de certitude. Il ne faudrait par conséquent guère s'étonner qu'une telle dichotomie révèle bientôt ses limites ; mais la piste mérite d'être creusée, car elle promet une expression formelle adéquate de la distinction entre un automatisme comportemental, et une « authentique » délibération, au sens de la *phronèsis*.

### 3.3.2. L'épreuve de la norme

Un article de Jean-Gabriel Ganascia[37] nous servira d'entrée en la matière de la norme comme « épreuve », au sens de Ricœur. L'article en question part de la controverse qui opposa Benjamin Constant à Kant sur le « droit de mentir ». Kant défendit mordicus la position selon laquelle le mensonge était toujours un mal moral, car minant l'idée même de vérité. Constant, en revanche, proposa la vue selon laquelle des principes généraux peuvent être complétés par des principes particuliers, qui prennent le dessus dans certaines circonstances. L'argument de Kant contre cette position fut que – au moins à l'époque – aucune logique n'existait où des cas portant contradiction prévalaient sur des règles générales. Or Ganascia se propose de reprendre à nouveaux frais cette discussion, en ripostant qu'une telle logique peut être définie de nos jours, sous la forme d'une logique non-monotone. De telles logiques sont nées pour répondre aux besoins de la survenance de situations inattendues, ou encore pour interroger le raisonnement du sens commun (« en règle générale, un oiseau vole »). Une réponse moderne est, entre autres, la technique ASP (pour *Answer Set Programming*, *programmation par modèles stables*[38]).

Vu l'omniprésence de la technologie ASP dans les travaux que nous avons consultés, nous nous permettons de l'aborder en quelque détail[39]. Afin d'introduire ASP, commençons par un exemple dans une technique très proche et que nous avons déjà rencontrée, à savoir Prolog :

*homme(jean).*

*célibataire(X) :- homme(X), not mari(X).*

*mari(X) :- homme(X), not célibataire(X).*

Fixons d'abord le vocabulaire : la première *clause* n'a pas de *corps* ; elle exprime donc un *fait*. Les deux dernières clauses ont une *tête* et un corps ; elles expriment par conséquent des *règles*. Intuitivement, le sens de ce petit programme est clair : pour être un célibataire, il faut être un homme et ne pas être marié ; inversement, pour être un mari, il faut être un homme et ne pas être connu comme célibataire. L'intuition se range ici derrière la logique du premier ordre et n'y trouve donc rien à redire. Or la plupart des implémentations de Prolog vont buter contre une telle définition

[37] J.-G. GANASCIA, *Modelling ethical rules of lying with Answer Set Programming*.

[38] En français, nous avons vu passer toutes les orthographes imaginables pour ensemble(s)(-)réponse(s) ; comme le terme de « modèle stable » (*stable model*) est dans ce contexte un synonyme, nous avons préféré traduire par ce terme, dont le sens et la construction morphologique sont mieux assurés.

[39] Pour présenter ASP, outre l'article précité de Jean-Gabriel GANASCIA, nous avons recours au chapitre de M. GELFOND, *Answer Sets* et à l'article de Th. EITER, G. IANNI et Th. KRENNWALLNER, *Answer Set Programming: A Primer*.

circulaire, où les termes *mari(X)* et *célibataire(X)* se définissent mutuellement. En effet, Prolog parcourt simplement dans l'ordre les clauses à sa disposition : pour prouver *célibataire(X)*, il doit prouver *not mari(X)*. Pour prouver *mari(X)*, il doit prouver *not célibataire(X)*, etc. L'évaluation de Prolog ira donc à l'infini, jusqu'à l'épuisement des ressources système allouées au programme.

Voilà pour Prolog. En ASP, il est possible d'écrire exactement le même programme, avec la même syntaxe. Or la résolution du programme, elle, est très différente par rapport à ce qui se passe en Prolog. Dans une première étape, ASP instancie toutes les variables. Cela donne le programme entièrement instancié – aussi appelé *univers de Herbrand* – suivant :

*homme(jean).*

*célibataire(jean) :- homme(jean), not mari(jean).*

*mari(jean) :- homme(jean), not célibataire(jean).*

Dans une deuxième étape, ASP va collecter tous les modèles possibles du programme, c'est-à-dire des ensembles de faits qui peuvent prétendre à décrire le monde du programme – il s'agit de la base de Herbrand : {homme(jean),mari(jean),célibataire(jean)}, {homme(jean)}, {mari(jean)}, etc.

Pour chaque modèle de la base de Herbrand, il faut retenir seulement les modèles stables, c'est-à-dire ceux qui ne contredisent pas le programme. Pour cela, il faut réduire la négation, qui est une source d'instabilité du programme[40]. Cette réduction, dite *de Gelfond-Lifschitz*, fonctionne comme suit. Pour chaque terme du programme introduit par *not* :

1. s'il est présent dans le modèle considéré, la clause où il figure est supprimée ;
2. sinon, le terme nié est retiré de la règle.

Ainsi en considérant le modèle {*homme(jean), célibataire(jean)*}, la réduction donne le résultat suivant :

[40] Le problème de la négation en informatique a une portée philosophique. Dans l'analyse que Gérard Chazal consacre au sujet, il constate que la négation comme échec revient à traiter le vrai comme « présence », et le faux comme « absence ». L'ordinateur est donc incapable de faire la différence entre sens et dénotation (au sens frégéen), ce qui pourrait expliquer pourquoi – à l'époque où écrit Chazal – une analyse fine des actes de langage négatifs échappe à l'ordinateur. Et le problème, dès lors, est de savoir si cette *incapacité de nier* dont fait preuve l'informatique était éphémère, liée à l'état des technologies de l'époque, ou si au contraire elle fait signe vers quelque chose de plus fondamental, d'une « limite essentielle » de la machine. Voir le chapitre dédié à ce sujet dans G. CHAZAL, *Le miroir automate*, pp. 90-114.

*homme(jean).*[41]

*célibataire(jean) :- homme(jean).*[42]

*ø*[43]

Il faut maintenant confronter le modèle considéré au résultat de la réduction. Dans ce cas, il n'y rien qui change : nous gardons donc {*homme(jean),célibataire(jean)*}. Si ce modèle, appelé modèle minimal, est le même que le modèle sous considération – ce qui est le cas ici – alors le modèle peut être dit stable. Le programme a d'ailleurs deux modèles stables. Outre celui que nous venons de calculer, voilà l'autre modèle stable : {*homme(jean),mari(jean)*}.

Si le procédé paraît un peu lourd, il faut se rappeler que la technique qu'ASP déploie ici ne fait finalement rien d'autre que construire une table de vérité pour notre programme, où seules les valeurs vraies font partie du modèle :

| *homme(jean)* | *célibataire(jean)* | *mari(jean)* |
|---|---|---|
| **vrai** | faux | **vrai** |
| **vrai** | **vrai** | faux |

L'équivalence entre deux programmes peut être établie sémantiquement en ce qu'ils ont les mêmes modèles stables. Il faut encore distinguer entre une équivalence forte et une équivalence uniforme. Dans le cas de l'équivalence forte, chaque union avec un programme tiers maintient l'équivalence entre les deux programmes. En revanche, dans le cas de l'équivalence uniforme, cette condition est affaiblie : il suffit alors que les deux programmes restent équivalents pour chaque union avec un ensemble de faits (ou clauses entièrement instanciées).

Voilà pour le principe d'ASP. Il en découle des conséquences importantes, que nous ne ferons que mentionner ici. Tout d'abord, ASP est entièrement déclaratif : les traces de programmation impérative de Prolog sont gommées. Ainsi, ASP n'est pas sensible à l'ordre des règles, ni non plus à l'ordre des sous-buts dans une règle. L'opérateur de coupure de Prolog n'a plus lieu d'être.

Une autre conséquence, tout à fait décisive, qui découle de la sémantique renforcée de la négation comme échec doit être fortement soulignée : ASP ne s'appuie pas sur l'hypothèse de monde clos ! Rappelons-nous la négation comme échec en Prolog :

*p(a) :- not p(b).*

*p(a)* est vrai si *p(b)* ne peut pas être prouvé. Dans ce cas, comme nous n'avons pas p(b) dans nos faits, *p(a)* est vrai. C'est-à-dire, en Prolog. En ASP, en revanche, la même requête retourne « *unknown* ».

[41] Explication : comme il n'y a aucun terme qui soit nié, il n'y a rien à réduire.
[42] Explication : le terme « mari(jean) » est introduit par *not* et ***ne*** figure ***pas*** dans le modèle considéré. Il faut donc appliquer l'alternative 2 et le supprimer de la clause.
[43] Explication : le terme « célibataire(jean) » est introduit par not et figure dans le modèle considéré. Par conséquent, c'est la première alternative qui est de mise et il faut supprimer la clause entière.

En d'autres termes, ASP repose sur une logique *trivalente*. Afin d'obtenir le même résultat qu'en Prolog, il faut forcer l'hypothèse du monde clos, « fermer » la clause :

*¬p(a) :- not p(b).*

Pour obtenir ce résultat, ASP dispose d'un opérateur inconnu de Prolog, qui est la *négation classique*, ou *forte* : ici, s'il est impossible de prouver *p(b)*, la négation de *p(a)* sera vraie, et *p(a)* donc faux.

Ainsi, avec des moyens conceptuels simples, ASP parvient à modéliser un éventail important de phénomènes, notamment les croyances d'un agent, le raisonnement du sens commun, des références révocables ou encore des préférences et des priorités. Pour revenir à notre exemple du début de la section sur le droit de mentir, l'expression de règles par défaut – qui ont des exceptions – est très naturelle en ASP. Ainsi, un énoncé comme « en règle générale, les oiseaux volent » peut être rendu comme suit :

*voler(X) :- oiseau(X), not ¬voler(X).*

Une telle règle édicte qu'un oiseau vole*, à moins que nous ne sachions explicitement* le contraire, par exemple :

*¬voler(X) :- autruche(X).*

Pour revenir à l'exemple qui a motivé l'introduction d'ASP, il devient très facile d'exprimer l'idée de Constant qu'il y a une pluralité de principes moraux, de généralité plus ou moins grande. Dans chaque situation, il importe de choisir le principe le plus approprié, c'est-à-dire, le principe le plus spécifique qui puisse être appliqué à la situation en question ; en l'occurrence, l'argument veut que l'interlocuteur – l'assassin – doive « mériter » la réponse :

*agir(P,A) :- action(A), not ¬juste(A).*

*¬juste(mentir(P, PP)) :- personne(P), proposition(PP), not ¬mériter(P, PP).*

*¬juste(A) :- conséquences(A,meurtre).*

*¬mériter(P, PP) :- personne(P), proposition(PP), conséquences(savoir(P,meurtre)).*[44]

Cet exemple mérite de plus amples commentaires. Pour commencer, il est clair que la théorie de Benjamin peut recevoir un appui psychologique : l'être humain a tendance à accorder plus de poids à des règles spécifiques qu'à des règles générales[45]. La question serait alors de savoir dans quelle mesure le « logicisme » de Kant est bien fondé de se passer de la plausibilité psychologique. Deuxième remarque, l'exemple montre bien, une fois de plus, combien il est nécessaire que toutes les exceptions soient explicitées d'emblée : la question de savoir si de tels procédés peuvent espérer un jour rendre compte de la complexité du réel reste entièrement ouverte.

[44] Adapté d'après J.-G. GANASCIA, *loc. cit.*, p. 44.

[45] Voir à ce propose l'article de L. B. MULDER, J. JORDAN et Fl. RINK, *The effect of specific and general rules on ethical decisions*.

En l'occurrence, pour qu'il puisse y avoir mensonge, il serait peut-être plus économe de formuler la règle comme quoi le mensonge est une catégorie qui est applicable uniquement quand les interlocuteurs sont en situation d'égalité entre eux : si par exemple, la relation de pouvoir est déséquilibrée entre les interlocuteurs, la liberté de parole de la partie la plus faible est supprimée et donc, il n'est plus en mesure de produire des actes de langage adéquats[46]. Asha Tickoo a qualifié une telle situation de *dominée* (*disempowered*). Elle se produit lorsqu'un interlocuteur renonce à sa liberté de parole ou signifie à l'autre le silence. La domination ne doit pas être explicite, mais peut se révéler par une implicitation (*implicature*). Par voie de conséquence, un échange ne peut être dit « libre » qu'à la double condition que les deux interlocuteurs aient le droit – et le devoir ! – de justifier leurs dires (condition de justification), et de réfuter les justifications de l'autre (condition de réfutation). Il convient encore de nous assurer que la justification et la réfutation sont proportionnelles l'une par rapport à l'autre.

Nous inspirant de cette exigence supplémentaire, nous pourrions alors faire l'économie du prédicat *mériter* en ajoutant un but supplémentaire au prédicat *mentir* :

> *mentir(P1,P2,PP) :- personne(P1), personne(P2), proposition(PP), ¬savoir(P1,PP), not ¬empowered(P1,P2).*

où *¬savoir(P1,PP)* signifie que P1 doit savoir que PP est faux (condition classique pour définir le mensonge) et *not ¬empowered(P1,P2)* signifie qu'il n'y ait pas de relation de soumission (*disempowerment*) de P1 vis-à-vis de P2. Si donc nous nous trouvons face à un assassin, comme dans l'exemple de Kant et Constant, il y aurait un rapport de domination, une relation de pouvoir déséquilibrée, en vertu de quoi l'interlocuteur ne serait pas tenu à dire la vérité.

#### *3.3.2.1. La norme comme contrainte*

Nous voudrions présenter ici – pour illustrer le propos – DALMAS[47] (pour *Deontic Action-Logic Multi-Agent Systems*) qui propose une architecture abstraite et normée d'un SMA. L'architecture abstraite se compose du système multi-agents à proprement parler, *D*, et le système normatif qui le

[46] Ces cas de *disempowerment* sont analysés par A. TICKOO, *On assertion without free speech*. Dans cet article, l'auteur s'intéresse aux rapports de pouvoir qui mettent à mal une communication libre et non faussée. Dans un langage dit « totalitaire », il existe une hiérarchie dans les constructions verbales, où certaines ont davantage droit à l'existence discursive que d'autres. Loin de se cantonner aux régimes communistes ou dictatoriaux, ce type de relation langagière s'instaure dès qu'une relation asymétrique de pouvoir s'installe, par exemple dans une relation de patient à médecin dans un contexte psychiatrique. En effet, toute parole du patient risque d'être interprétée comme symptôme, et donc dispense le thérapeute d'y répondre. Les mêmes relations se créent par ailleurs souvent dans des rapports entre professeur et étudiants, où le pouvoir peut être interprété ici dans un sens très large, comme une asymétrie de connaissances socialement reconnues comme étant expertes. Notons que la relation de pouvoir induite par une asymétrie de connaissances devient problème et enjeu dans la *modélisation d'accompagnement*, dont nous traitons plus loin (§ 3.4.2.6).

[47] Notre présentation – fort simplifiée – de DALMAS se base sur l'article de M. HJELMBLOM et J. ODELSTAD, *jDALMAS: A Java/Prolog Framework for Deontic Action-Logic Multi-Agent Systems*, ainsi que le mémoire de master du premier auteur, *Deontic Action-Logic Multi-Agent Systems in Prolog*.

régit, *N*. Ce qui nous intéresse au premier chef dans cette architecture est, de toute évidence, le développement du système normatif.

Reprenons l'exemple des auteurs afin de remonter, à partir de là, dans le formalisme logique. Celui-ci est très simple : il s'agit d'un système qui ne comprend que deux agents, Morphe et Chroma. Chaque agent se caractérise par deux attributs, deux actions, et une stratégie. Les attributs des agents portent sur leur couleur (noir ou blanc) et leur forme, circulaire ou rectangulaire. Les valeurs des attributs des agents à un moment donné reflètent l'état du système. Les actions que ces agents peuvent accomplir, c'est de muter de couleur ou de forme ; la stratégie de Morphe est de changer de préférence de forme, et celle de Chroma de changer de couleur. À chaque tour, les agents vont poser une action, tout en se laissant guider par la norme qui édicte qu'un agent ne peut faire une action qui aurait pour résultat un état dans lequel les deux agents auraient exactement les mêmes valeurs sur tous leurs attributs. Formellement, cette norme s'exprime comme suit :

$$Diff(\omega_1, \omega_2; s) \rightarrow \neg May\ Do\big(\omega_1, Eq(\omega_1, \omega_2; s)\big)$$

Selon nos auteurs, la norme est une mise en relation de deux conditions : une condition descriptive et une condition normative (ou une conséquence). La condition descriptive exprime dans quel état le système normé doit se trouver pour que la norme s'applique. L'exemple cité ci-dessus a deux conditions d'état : *Diff* et *Eq*, qui portent sur les agents $\omega_1$ et $\omega_2$ dans un certain état *s*. La condition normative, quant à elle, n'est pas une condition d'état mais une condition de situation. Elle porte sur la situation $<\omega,s>$, c'est-à-dire que la situation est l'état d'un système et un agent à qui c'est le tour d'agir.

La condition normative, ici, dit donc quelque chose sur ce que peut faire l'agent $\omega_1$. Pour exprimer la contrainte à proprement parler, le formalisme a recours à quelques briques logiques élémentaires venues des logiques d'action et déontique. De la logique d'action, l'architecture retient l'opérateur d'action binaire $Do(\omega,p)$ qu'il convient d'interpréter comme « ω veille à ce que p ». La logique d'action se combine à la logique déontique au travers des opérateurs *Shall* et *May*. La sémantique de l'opérateur *Shall* se laisse résumer en trois postulats simples :

1. $(p \rightarrow q) \rightarrow (Shall\ p \rightarrow Shall\ q)$
2. $(Shall\ p \wedge Shall\ q) \rightarrow Shall(p \wedge q)$
3. $Shall\ p \rightarrow \neg Shall\ \neg p$

L'opérateur de permission *May* se définit alors simplement comme suit :

$$May\ p \leftrightarrow \neg Shall\ \neg p$$

Voilà comment le système normatif se définit en DALMAS. Nous n'entrerons pas trop dans les détails du système multi-agents à proprement parler ; définissons-le simplement par le nonuplet suivant, dont les trois premiers termes sont des ensembles et les six derniers des foncteurs :

$$D = <\Omega, A, S, A_f, ds, ps, cs, \gamma, \tau>$$

Nous nous contenterons de brièvement définir chacune des composantes de cet aspect de l'architecture :

- $\Omega$ est l'ensemble des agents.
- *A* représente l'ensemble des actions *a*, définies comme des fonctions qui, exécutées par un agent (élément de Ω) dans un état donné (élément de S), aboutissent à un nouvel état :

  $a : \Omega \times S \rightarrow S$

- *S* représente l'espace d'états.
- $A_f$ est la fonction qui renvoie, pour un agent $\omega$ dans un état *s*, l'ensemble des actions réalisables (à comprendre dans le sens où une action réalisable est une action *accessible* à l'agent).
- *ds* est la fonction qui renvoie, pour un agent $\omega$ dans un état *s*, l'ensemble des *structures déontiques*. Une structure déontique est un sous-ensemble d'actions réalisables qui partagent la même modalité déontique (permissible, interdit, obligatoire).
- *ps* est la fonction qui renvoie, pour un agent $\omega$ dans un état *s*, l'ensemble des *structures de préférence*. Une structure de préférence est un ensemble d'actions réalisables ordonnée selon leur utilité pour l'agent.
- *cs* est la fonction qui renvoie, pour un agent $\omega$ dans un état *s*, l'ensemble des actions parmi lesquelles l'agent peut choisir. En général, l'ensemble ne contiendra qu'un seul élément, à savoir l'action permise que l'agent préfère.
- Si toutefois l'ensemble de choix contient plusieurs éléments, la fonction $\gamma$ permettra de forcer le choix.
- Finalement, $\tau$ est la fonction de prise de tour qui indique, pour tout agent, l'agent qui agira après lui.

Un agent à qui c'est le tour choisira alors son action préférée qui soit à la fois faisable et non explicitement défendue (*prohibited*).

L'architecture, nous l'avons dit, est dite « abstraite ». De fait, avant d'obtenir un système concret tel que celui de Chrome et Forma, il faut instancier tous les paramètres de l'architecture abstraite. C'est ce que les auteurs ont illustré, dans leur article, par un serveur Prolog et une petite application graphique en Java permettant de remplir les paramètres.

Parmi les questions de recherche qu'offre l'architecture DALMAS, les auteurs proposent notamment le problème suivant :

> *For a specific DALMAS D, which normative system would give the most efficient behaviour of the system?*[48]

Les auteurs, ici, comptent exploiter la séparation entre architecture abstraite et système normé pour « expérimenter » – au sens que peut revêtir ce mot dans le cas des simulations à base d'agents que

[48] M. HJELMBLOM, *op. cit.*, p. 46.

nous avons vu au deuxième chapitre (§ 2.3.1) – ce qui se passe en changeant d'ensemble de normes, tout en gardant constant les agents, leurs actions, leurs préférences, etc. De telles expériences participent, en somme, de l'idée de la norme comme épreuve, en tant qu'elles permettent d'anticiper les effets d'un changement de normes dans un système donné.

Nous n'insisterons pas davantage sur cette architecture ; ce qui nous importe ici, c'est la façon d'envisager la norme qui s'en dégage. Il est clair que l'aspect dynamique de la simulation – le comportement des agents – est porté par le compromis entre normes et stratégies, entre prohibitions et préférences. La norme, en d'autres termes, est une *contrainte,* une *connaissance* par rapport au problème à résoudre : elle co-construit l'espace de recherche logique[49]. Nous en voulons pour preuve l'accentuation très prononcée de la norme comme prohibition : la norme est interdiction, négativité ; elle est limite à ne pas franchir, borne à une visée dont la dynamique revient aux agents eux-mêmes.

Or la norme comme négativité se prête bien aux exigences de la calculabilité. En effet, la démarche essentielle de l'ordinateur est la limitation du nombre de solutions à envisager, alors que de prime abord, toutes les solutions sont pareillement plausibles. Une telle observation pourrait avoir des retombées sur la façon dont nous comprenons l'apprentissage humain, aussi. En effet, nous aurions pu croire que pour l'homme, l'intelligence est une forme de créativité, dans la mesure où son « espace » de choix est a priori très restreint et que la chose essentielle, dans l'apprentissage, est l'élargissement de l'éventail des possibles. Or la leçon de l'informatique est différente : la pensée intelligente ressemblerait plus à un type de *sélection*[50], à l'instar de la sélection darwinienne, qui tue des idées et des hypothèses au lieu de tuer le vivant.

En revanche, ce qui semble rester absolument étranger à une conception de la norme comme contrainte, ce sont les phénomènes de transgression de la norme d'une part, et de conflit entre normes, d'autre part. En modélisant son problème sous forme de contrainte, l'ingénieur adopte en quelque sorte une vue très particulière de la conformité à la norme : il ne s'agit pas d'une conformité adaptative, *sufficising*, caractéristique du vivant et de sa complexité, mais d'une conformité *optimale*, comme un certain idéal qui devient but à atteindre.

#### *3.3.2.2. La norme comme contrat : sanctions et répercussions*

Une façon plus « flexible » de considérer les normes est de les concevoir comme des *contrats* : par un contrat, les contractants s'engagent « librement » à fournir un service ou à exécuter certains comportements ou travaux. Si un des contractants ne respecte pas sa part du marché, le contrat prévoit des clauses de rupture, détaillant les répercussions en cas de violation. Dans la classification

[49] Notre traitement de la contrainte se base ici sur le chapitre d'A. FARINELLI, M. VINYALS, A. ROGERS et N. JENNINGS, *Distributed Constraint Handling and Optimization*, dans G. WEISS, *Multiagent Systems*, pp. 547-578.
[50] M. DUBOIS, *La métaphore et l'improbable*, p. 71.

de Madl-Franklin[51], nous nous situons clairement sur le niveau délibératif : la prise en compte des sanctions se fait en prévision, ce qui distingue ces normes du niveau réactif. Elles portent sur des actions, ce qui les distingue du niveau métacognitif.

Nous trouvons cette conception dans un article[52] où l'auteur formalise la notion de contrat sous forme d'*autorisations* et d'*obligations dirigées*. Les obligations dirigées engagent un agent *i* vis-à-vis d'un agent *j* pour qu'une situation soit réalisée ou une action accomplie. Le contrat stipule également des autorisations : si l'agent *i* manque à son devoir contractuel, *j* peut – selon les cas – demander des actions supplémentaires de *i*, annuler certains des devoirs qui l'obligent lui-même vis-à-vis de *i*, ou encore prendre l'initiative d'une action réparatrice. L'autorisation, dans le chef de l'auteur, est comme une permission, mais à effet performatif. Par exemple, si *j* est autorisé de demander un paiement après la violation du contrat par *i*, de ce fait même se crée une obligation pour *i* de donner suite à cette demande.

L'idée de la pénalité est ici donc clairement établie. Or les façons dont les agents feront effectivement respecter les clauses contractuelles restent à définir. En effet, comme les contrats sont « librement » conclus entre agents, il est impératif de veiller au grain, de suivre la bonne exécution des contrats. C'est pourquoi, dans les systèmes qui s'inspirent d'une conception contractuelle, des mécanismes de garantie doivent être prévus. À titre d'illustration de ces implémentations, tournons-nous vers MaNEA[53], une architecture multi-agents où le respect des engagements vise à diminuer l'incertitude qui règne entre agents et qui sont nombreux et hétérogènes. À cette fin, MaNEA ajoute deux agents d'un type particulier : le gardien et le gestionnaire des normes (resp. *norm enforcer* et *norm manager*). Le gardien des normes surveille le comportement des agents : il détecte des violations et y réagit en infligeant des pénalités ou, le cas échéant, des récompenses. Pour s'y retrouver, il dispose d'une liste de normes actives, ainsi que d'une liste de rôles, dont nous remettons la discussion à plus tard. Pour rester dans la logique des systèmes multi-agents, nous pouvons légitimement poser la question de la façon dont le gardien de la norme connaît les contrats. Dans MaNEA, la réponse est l'introduction d'encore un autre agent – c'est le gestionnaire des normes – qui détient une sorte de registre des normes, avec pour chacune les conditions de validité (notamment dans le temps).

Le gestionnaire maintient ce registre en s'abonnant aux messages émis par le système de gestion des organisations (OMS ou *Organization Management System*). Comme son nom l'indique, le système OMS est en charge des organisations entre agents. Il propose ainsi des services structuraux, qui permettent d'ajouter ou de supprimer des normes, des rôles et des groupes au sein d'une

---

[51] Cf. la section « Comportements éthiques implicites et explicites » (§ 1.3), où nous avons vu l'enrichissement que proposent les auteurs éponymes pour venir à bout de la dichotomie traditionnelle entre comportements implicites et explicites.

[52] Voir Fr. DIGNUM, *Autonomous agents and norms*.

[53] Notre présentation se fonde sur le chapitre *MaNEA: A Distributed Architecture for Enforcing Norms in Open MAS*, dans N. CRIADO, *Using Norms to Control Open Multi-Agent Systems*, pp. 193-239. En deux mots, MaNEA ajoute une couche normative sur une plateforme multi-agents existante, Magentix2. Celle-ci – à la manière de JADE, que nous avons vu au deuxième chapitre (§ 2.2.3) – met l'accent sur la passation de messages entre agents autonomes. Pour ce faire, elle offre une infrastructure de communication basée sur FIPA ACL, un support étendu pour des agents Jason, ainsi que sa propre variante de la gestion des organisations et rôles (rappelons-nous Moise), dont le nom mérite citation : THOMAS (*Teams and Hierarchies for Open Multi-Agent Systems*). Le lecteur intéressé peut se référer aux ressources disponibles sur le site web suivant : http://www.gti-ia.upv.es/sma/tools/magentix2/index.php.

organisation ; des services informatifs, qui renseignent sur l'état courant d'une organisation donnée ; finalement des services dynamiques, qui permettent à des individus d'assumer un rôle particulier dans une organisation, ou de signaler leur départ.

Même si nous remettons à plus tard une discussion approfondie des organisations, il importe de souligner deux choses : d'une part, ni le gestionnaire des normes, ni l'OMS ne rétablissent une forme d'autorité omnisciente : leur charge est étroitement liée aux normes, accords et contrats conclus en vertu d'un protocole particulier ; il est à la limite loisible aux agents de s'arranger en dehors des services fournis par l'OMS ; le gardien et le gestionnaire seraient dès lors hors-jeu. La deuxième remarque porte sur la nature du contrat : contrairement à une certaine tradition philosophique, le contrat ne prétend pas ici à un statut premier, mais dérive entièrement des structures « sociales » qui le rendent possible. Nous aurons l'occasion d'approfondir ce thème dans la section consacrée aux organisations (§ 3.3.5).

#### *3.3.2.3. La norme comme indication d'ordonnancement*

Que doivent faire nos pauvres agents tellement épris de pureté logique lorsqu'un conflit entre normes se fait jour ? Est-il seulement possible, dans un système multi-agents, de modéliser un conflit entre normes ? Quelle portée donner à un tel conflit, sachant qu'à aucun moment, le téléologique n'interviendra – au moins pas explicitement ? C'est le sujet abordé par l'article[54] que nous allons parcourir dans cette section. Le conflit entre normes y est défini comme une chose qui est, pour une organisation virtuelle donnée, à la fois obligatoire (ou permis) et défendue. Nous ne nous arrêterons pas sur les organisations virtuelles ici : contentons-nous de dire que, dans cet article, une organisation virtuelle est modélisée comme une machine à états finis, dans laquelle les actions des agents individuels donnent lieu à des transitions d'état.

Plus pertinent pour notre propos est le traitement de la norme, définie comme une construction normative assortie de trois métadonnées temporelles, à savoir la date de déclaration ou d'introduction de la norme, sa date d'activation ainsi que sa date d'expiration. La construction normative spécifie quel agent, avec quel rôle, a l'obligation, la permission ou l'interdiction de faire quelque chose qui se présente sous la forme d'un prédicat logique. Ce prédicat peut être lui-même soumis à des contraintes, qui apportent diverses spécifications ou précisions sur la norme.

Le traitement du conflit est illustré par l'exemple, très simple, d'un jeu de blocs : le seul prédicat y est *shift(X,Y,Z)* et il représente l'action de décaler un bloc X se trouvant sur Y de sorte que X se trouve désormais sur Z. Une construction normative peut alors être formalisée comme suit :

$$F_{A:R} shift(X, Y, Z) \wedge X = a$$

[54] M. J. KOLLINGBAUM, W. VASCONCELOS, A. GARCÍA-CAMINO et T. J. NORMAN, *Conflict Resolution in Norm-Regulated Environments Via Unification and Constraints*.

Une telle formule dit qu'il est interdit (*F* pour *forbidden*) à tout agent A dans tout rôle R d'opérer l'action *shift(X,Y,Z)* pour autant que *X* soit le bloc *a*[55]. Donnons maintenant une deuxième norme qui entre en conflit avec la première :

$$P_{A:R} shift(X, Y, Z) \wedge X = a \wedge Y = r$$

Cette norme dit qu'il est permis à tout agent A dans tout rôle R d'opérer l'action *shift(X,Y,Z)* pour autant que *X* soit le bloc *a* et *Y* le bloc *r*. À première vue, il s'agit donc ici bel et bien d'un conflit entre deux normes, l'une interdisant ce que l'autre permet, à savoir décaler le bloc *a*.

L'algorithme de détection de conflits commence, dans un premier temps, par essayer l'unification des termes, c'est-à-dire qu'il cherche une substitution des variables libres des deux normes de manière à dériver une proposition identique dans les deux cas. C'est tout à fait possible ici : prenons la substitution qui remplace X par a, Y par r et Z par u, nous obtenons dans les deux cas *shift(a,r,u)*. Si l'unification réussit, il y a potentiellement conflit. Dans un deuxième temps alors, l'algorithme cherche à ajuster la portée de normes en limitant les plages de valeurs possibles. Prenons un autre exemple pour illustrer ce point :

$$P_{A:R} p(c, X) \wedge X > 50$$

$$F_{a:b} p(Y, Z) \wedge Z < 100$$

La première norme édicte qu'il est permis à tout agent *A* dans tout rôle *R* d'opérer *p(c,X)* pour autant que *X* soit plus grand que 50 ; la deuxième norme dit qu'il est interdit au seul agent *a* dans le rôle *b* d'opérer *p(Y,Z)* au cas où *Z* est plus petit que 100. La substitution permettant ici l'unification est, par exemple, *{A/a, R/b, Y/c, X/Z}* ; il y a donc conflit potentiel, mais uniquement pour la plage de valeur de $\mathrm{X} > 50$ ou $\mathrm{Z} < 100$. Finalement, l'algorithme de détection recherche un chevauchement dans les périodes respectives d'activation, délimitées par les bornes d'activation et d'expiration des normes : pour qu'il y ait conflit, il faut effectivement que les deux normes soient actives en même temps au cours d'un certain laps de temps.

Une fois qu'un conflit est détecté, il s'agit d'essayer de le résoudre. Pour ce faire, l'algorithme de résolution de conflits va restreindre (*curtail*) la portée des contraintes associées aux normes conflictuelles. Formalisons le conflit entre normes $n_1$ et $n_2$ unifiées par la substitution σ par *conflict(n1, n2, σ)*. La restriction de portée (*curtailment*) de la norme $n_1$ par rapport à la norme $n_2$ s'obtient en conjuguant les contraintes de $n_1$ avec les contraintes niées de $n_2$ sur lesquelles la substitution *σ* a été appliquée. La formulation du procédé peut paraître kabbalistique, mais il s'agit en réalité d'une chose fort simple : reprenons l'exemple des normes énoncées ci-dessus. Si nous voulons restreindre la norme $\mathrm{P}_{A:R}\mathrm{p}(\mathrm{c}, \mathrm{X}) \wedge X > 50$ par rapport à la norme $F_{a:b} p(Y, Z) \wedge Z < 100$, il suffit de passer par les étapes suivantes :

1. Considérer la contrainte de $n_2$, *Z < 100*
2. Appliquer la substitution (*σ = {A/a, R/b, Y/c, X/Z}*) à la contrainte, ce qui donne *X < 100*

[55] Rappelons-nous que dans les notations qui suivent la convention de Prolog, les variables s'écrivent à l'aide de majuscules (*A, B,…,X, Y, Z*) et les constantes d'individus par des minuscules (*a, b, c, r, u*).

3. Nier cette contrainte : $\neg(X < 100)$
4. Conjuguer cette contrainte niée aux contraintes de la norme à restreindre. Nous obtenons donc : $P_{A:R}p(c, X) \wedge X > 50 \wedge \neg(X < 100)$, ce qui peut être simplifié en $P_{A:R}p(c, X) \wedge X \geq 100$.

Reste à savoir, bien sûr, une fois qu'un conflit a été détecté, quelle norme va faire l'objet de la restriction. Pour guider l'algorithme dans ce choix, des stratégies de restriction (*curtailment policies*) doivent être associées à chaque paire de normes. De telle stratégies peuvent, à titre d'exemple, donner la préséance à la norme dont la date de déclaration est la plus ancienne, ou au contraire la plus récente. Notons que ces stratégies peuvent à leur tour se voir enrichir de contraintes, rendant ainsi possible un contrôle très fin de la résolution des conflits.

Dans un ensemble de normes sans conflits, un agent peut évaluer la conformité de ses actions par rapport à l'état normatif global. Pour jauger la conformité de son action par rapport à l'ensemble des normes, l'agent recourt à un mécanisme fort semblable que celui que nous venons de décrire : par unification des termes, il détecte les normes applicables à son action ; puis il vérifie si la norme est actuellement active. Nous voyons ainsi élaboré un cadre normatif qui conjugue la résolution de contraintes – que nous avons vue à l'œuvre plus tôt, dans la première section de ce sous-chapitre (§ 3.3.2.1) – aux indications d'ordonnancement contenues dans les métadonnées des normes.

Dans les pistes de recherche que les auteurs mentionnent en fin d'article figure en premier lieu l'exploration des stratégies de restriction possibles ; la tradition juridique en ayant déjà légué trois : préséance de la loi la plus récente (principe du *lex posterior*), préséance de la loi émise par l'autorité la plus compétente (principe du *lex superior*), finalement préséance à la loi la plus spécifique (*lex specialis*) ; la recherche en systèmes multi-agents a déjà exploré les stratégies de restriction par négociation ; les auteurs voudraient s'atteler à l'inscription de telles stratégies dans le cadre de l'organisation virtuelle dont font partie les agents. Même si la richesse d'expressivité d'un tel cadre normatif est déjà appréciable, il reste toutefois à confirmer qu'il pourra dépasser l'étape du simple prototype. En effet, de l'aveu des auteurs[56], la vérification de la cohérence de normes munies de conditions d'activation et d'expiration reste pour l'heure une tâche computationnellement ardue.

### 3.3.3. Création de la norme

Selon Ricœur, la visée éthique constitue « l'inspiration » de la norme : ainsi une interdiction de voler trouve son inspiration dans la propriété privée, principe dont l'importance dans nos sociétés est bien connue. Or lorsque les travaux en SMA parlent de « l'émergence » de la norme, il convient de s'entendre et de ne pas prendre l'un pour l'autre : « inspiration » et « émergence » ne coïncident *pas*. Prenons, à titre d'exemple, le « jeu des t-shirts »[57] : chaque agent, dans ce jeu, dispose de deux t-shirts, l'un rouge, l'autre bleu. À l'issue du jeu, le but est que tous les agents portent la même couleur. En début de partie, le choix des couleurs que portent les agents est réparti aléatoirement. À

[56] Cf. l'article de la même équipe : M. J. KOLLINGBAUM et T. J. NORMAN, *Norm Adoption and Consistency in the NoA Agent Architecture*, p. 183.
[57] Exemple tiré de M. WOOLDRIDGE, *An Introduction to MultiAgent Systems*, pp. 174-175.

chaque tour, les agents sont couplés : dans chaque couple, les agents voient la couleur de l'autre. Chaque agent décide ensuite pour soi s'il garde sa couleur ou s'il en change.

Observons que ce qui émerge, c'est un accord collectif sur une branche de l'alternative. La norme est vue comme un problème de coordination pur : la nature du choix importe peu, seule importe qu'il y en ait un. L'émergence conçue en ces termes *présuppose* donc le moment téléologique pour se concentrer sur la *diffusion sociale* de la norme au moyen d'une *acquisition individuelle*. L'acquisition, elle, se joue dans une interaction où il s'agit surtout pour les agents de se *coordonner*. Bref, là où Ricœur pense dérivation conceptuelle, l'émergence – le créneau du SMA rappelons-nous – parle devenir historique.

Comment créer des normes[58] ? À l'instar de ce nous avons vu au premier chapitre (§ 1.5), il y a deux façons de faire : elles peuvent être créées dans un mouvement soit ascendant, soit descendant. Dans le cas d'une création descendante (*top-down*) ou prescriptive, il faut encore distinguer – dans un cadre SMA – entre deux approches : une approche statique (*off-line*) et une approche dynamique (*on-line*). L'approche statique est celle qui est la plus familière aux informaticiens, dans la mesure où les normes sont ici largement spécifiées avant l'exécution de quoi que ce soit.

#### *3.3.3.1. Création statique*

Un exemple de cette approche – descendante et statique – peut être trouvé dans la conception de protocoles dits éthiques[59]. L'auteur part du problème qu'il appelle celui de la « cohérence éthique » (*ethical consistency problem*) : étant donné des organisations hétérogènes, peuplées d'agents contraints par des principes éthiques différents, comment serait-il possible de contraindre ces mêmes agents par un même ensemble de principes afin d'obtenir un effet éthique cohérent ? Pour y répondre, l'auteur définit une hiérarchie entre normes, partant du niveau des *principes* jusqu'à celui des *contraintes*, en passant par ceux des *exigences* et des *protocoles éthiques*.

La terminologie utilisée par l'auteur ne doit pas nous induire en erreur : pour lui, le niveau des « principes éthiques » ne se réfère pas (nécessairement) à une inspiration téléologique ; il s'agit simplement de la première source de normes du système qu'il s'agit de modéliser ; première, car ses normes relèvent du niveau le plus général : il peut s'agir de valeurs, mais ce premier niveau peut déjà contenir un règlement de travail, des codes de bonne conduite, une charte déontologique, etc.

Afin de spécifier les principes éthiques en termes plus analytiques, il faut décomposer tout principe en détaillant l'ensemble de ses observables, la classe des opérations auxquelles il fait appel, ainsi que la contrainte normative qu'il exprime. Les exigences éthiques qui en découlent sont obtenues en redéfinissant la contrainte normative en termes de *préconditions* et ce, pour chaque opération concernée. En appliquant l'exigence à un système entièrement spécifié, nous obtenons un protocole.

---

[58] Pour la distinction entre création ascendante et descendante, nous nous basons sur la thèse de doctorat de N. CRIADO PACHECO, *Using Norms to Control Open Multi-Agent Systems*, notamment la section *Norm Creation Process*, pp. 55-62.
[59] Voir le chapitre qu'y consacre M. TURILLI, *Ethical Protocols Design*, dans M. ANDERSON et S. L. ANDERSON, *Machine Ethics*, pp. 375-397.

Enfin, pour passer du protocole à l'implémentation, il faut combiner celui-ci avec la spécification fonctionnelle de l'agent.

L'approche est certes louable pour le soin apporté aux spécifications. Cependant, elle n'en est pas moins affublée d'un certain nombre de faiblesses. Le système n'est guère adaptatif, ni respectueux de l'autonomie des agents. La faiblesse principale, toutefois, est ailleurs. L'obligation de spécifier chaque opération revient à abandonner toute forme d'universalité : la « garantie » éthique non seulement est statique, mais elle est morcelée au travers des implémentations – diverses et variées – des différentes opérations. Dernière faiblesse, vu le caractère abstrait de la spécification, il reste fort vague comment il faudrait s'y prendre concrètement pour joindre le protocole à l'implémentation. De fait, l'article ne dépasse guère la vertueuse injonction selon laquelle toutes les opérations doivent être pensées et formulées de manière à « prendre en compte » le protocole.

#### *3.3.3.2. Création dynamique descendante*

Eu égard aux faiblesses de la création statique, il est aisé de comprendre pourquoi elle est assez peu utilisée. Tournons-nous dès lors vers les méthodes descendantes *dynamiques* de création de normes. Dans cette catégorie, nous trouvons les approches contractuelles, dont l'article que nous allons étudier maintenant[60] est un exemple illustratif. Cet article part de l'idée de *conventions* de haut niveau, qui sont de deux types : des *règles d'interprétation* d'une part et des *normes conditionnelles* (*prima facie*) d'autre part. Parmi les règles d'interprétation, il faut encore distinguer deux sortes. Une première sorte s'attache à décrire, en termes généraux, un vocabulaire commun. Si, par exemple, un agent doit vendre des biens de première nécessité à un prix « raisonnable », une règle d'interprétation pourrait expliquer ce qu'il faut entendre par ce terme. Une deuxième sorte de règle d'interprétation décrit les effets déontiques implicites de certaines actions, en termes d'obligations ou d'autorisations. Ainsi l'action de vendre implique une autorisation, pour le vendeur, de réclamer un paiement à l'acheteur, lequel se voit alors dans l'obligation de payer.

Les normes conditionnelles, quant à elles, se présentent le plus souvent comme des interdictions. Elles peuvent se présenter également comme des permissions ; il s'agit dès lors le plus souvent d'une exception faite à l'égard d'une interdiction de portée plus générale. Pour finir, ces normes peuvent aussi prendre la forme d'obligations, par exemple : « des agents devraient être coopératifs les uns envers les autres ». De telles obligations ne peuvent pas être exécutées telles quelles ; elles sont plutôt à concevoir comme des critères de choix qui guident les agents pour prendre leurs décisions, toutes autres choses étant égales par ailleurs.

Ce que toutes les conventions ont un commun, c'est leur portée générale : leurs conditions d'applicabilité s'inscrivent dans des situations ; elles ne sont pas liées à l'accomplissement d'actions concrètes. Pour cela, entre en scène la notion de « contrat » : un tel contrat décrit, en termes d'obligations et d'autorisations, les attentes entre deux agents suite à une action (par exemple, une vente). Une obligation peut porter sur une action ou sur un état à réaliser. Dans le dernier cas, l'agent

[60] Cf. Fr. DIGNUM, *Autonomous Agents with Norms*.

à qui incombe une telle obligation doit se créer des sous-buts afin de réaliser l'ensemble des actions qui peuvent induire l'état souhaité.

Mentionnons, pour mémoire, que l'article connaît encore un troisième niveau de normes, à savoir le niveau privé, interne aux agents. L'auteur prévoit en effet la possibilité d'inscrire certaines normes directement dans le comportement de certains types d'agents. L'important ici, c'est de voir que la convention reste de l'ordre du donné, voilà pourquoi cette approche est toujours descendante. Toutefois, la convention est explicite et distincte de l'implémentation des agents ; il est donc relativement aisé d'en changer. Mais surtout, l'ensemble des conventions ne doit pas être exempt de conflits : le conflit peut être toléré, vu que la décision de se conformer à la convention appartient à l'agent, en fonction des circonstances.

#### *3.3.3.3. Création dynamique ascendante*

Dans les approches ascendantes, nous retrouvons l'émergence des normes. Il faut distinguer ici entre une émergence par observation et une émergence par apprentissage cognitif. L'émergence par observation n'est rien d'autre que l'imitation, pure et simple, du comportement que l'agent observateur estime dominant. Nous n'en dirons pas plus et nous concentrons ici sur la création par apprentissage cognitif. Dans cette approche, les normes sont inférées logiquement, notamment à l'aide d'un système de logique inductive[61].

Nous trouvons un exemple parlant de cette dernière approche dans un article[62] consacré à l'apprentissage inductif de normes. L'enjeu, pour les auteurs, part du constat que la spécification artisanale de normes comportementales est une importante source d'erreurs, qui aurait beaucoup à gagner de l'aide d'une assistance automatisée. Dans l'arsenal notionnel déployé par l'article, nous trouvons des évènements et des « fluents ». Ces fluents sont des propriétés qui caractérisent l'état du système à un moment donné : un fluent peut être modélisé par une fonction ou par un prédicat qui comporte un argument temporel. En ASP, cela peut s'écrire comme suit :

$holdsat(f, t_i).$

Les fluents sont de deux types : les fluents normatifs, d'une part, les fluents du domaine d'autre part. Les fluents normatifs se laissent encore diviser en fluents de pouvoir, des permissions et des obligations. Les évènements sont, eux aussi, de deux sortes : les évènements normatifs et les évènements exogènes. Les évènements normatifs peuvent être des actions ou des violations. Dans le langage ASP, nous pouvons écrire : *event(e), evtype(e,obs), evtype(e,act), evtype(e,viol)*. Une séquence d'évènements exogènes est appelée une *trace*. En ASP, nous écrirons :

$observed(e,t_i).$

[61] Rappelons-nous, nous avons déjà rencontrée la logique inductive chez les Anderson au premier chapitre (§ 1.4.2), qui ont utilisé cette technique pour déduire des règles à propos de la prise de médicament assistée en se basant sur l'interrogation des spécialistes.

[62] D. CORAPI, M. DE VOS, J. PADGET, A. RUSSO et K. SATOH, *Norm Refinement and Design through Inductive Learning*.

Pour chaque trace, il est possible de calculer un modèle normatif en trois étapes. Premièrement, il faut appliquer la relation de génération $\mathcal{G}$ telle que :

$$\mathcal{G} : 2^{\mathcal{F}} \times \mathcal{E} \rightarrow 2^{\mathcal{E}_{norm}}$$

C'est-à-dire que la relation $\mathcal{G}$ calcule l'ensemble des parties de l'ensemble des évènements normatifs à partir de l'ensemble des parties de l'ensemble des fluents et de l'ensemble des évènements. Deuxièmement, toutes les violations sont générées : il s'agit soit d'évènements non permis, soit des obligations non honorées. Troisièmenent, la relation de conséquence $\mathcal{C}$ est appliquée.

$$\mathcal{C} : 2^{\mathcal{F}} \times \mathcal{E} \rightarrow 2^{\mathcal{F}} \times 2^{\mathcal{F}}$$

La relation de conséquence détermine le nouvel état en qualifiant tous les changements de fluents en termes d'initiation (ASP : *initiated(f, $t_i$)*) et terminaison (ASP : *terminated(f, $t_i$)*). Il s'agit alors d'apprendre des règles en sachant la façon de calculer un modèle normatif. L'apprentissage se fait par cas, où un cas est défini comme la trace de tous les évènements observés, une théorie et des exemples.

Illustrons de suite avec l'exemple donné par les auteurs : chaque agent détient un bloc d'un fichier qu'il est seul à connaître. En interagissant avec les autres agents, il doit faire l'acquisition (*download*) des blocs qu'il ne connaît pas encore afin de connaître le fichier dans sa totalité. Or, l'acquisition d'un bloc invalide la permission qu'a un agent d'en accepter un autre. Le partage (*share*) d'un de ses blocs réinstaure cette permission. La violation de cette norme a pour conséquence que l'agent se voit retirer le pouvoir d'acquérir d'autres blocs. Cet ensemble de règles est un garde-fou pour garantir la réciprocité des échanges.

Ainsi, une situation simple qui ne comporte que deux agents peut être entièrement constituée en 14 règles et un état initial de 10 faits « holdsat », dont nous ferons grâce au lecteur. Pour démarrer l'apprentissage inductif, il faut supprimer aléatoirement certaines règles, au plus 3. L'apprentissage se base sur des déclarations de mode, qui spécifient la forme des règles à apprendre. En l'occurrence, il s'agit de prendre soit des têtes de clause (*modeh*), soit des corps (*modeb*). Le résultat attendu doit être vrai dans le modèle stable qui résulte de l'union de la théorie et des traces. Si ce n'est pas le cas, l'apprentissage doit créer des règles de manière à rendre le résultat attendu vrai.

Supposons la suppression de la règle suivante, qui édicte qu'un agent ne peut faire une acquisition qu'après avoir partagé :

*initiated(perm(myDownload(Agent;Block)), I) :- occurred(myShare(Agent), I).*

La trace – c'est-à-dire les exemples qui sont à la base de l'apprentissage – se présente comme :

*observed(download(alice, bob, x3), 0).*

*observed(download(bob, alice, x1), 1).*

Alors le programme d'apprentissage crée les hypthèses suivantes, dont voici la première :

*initiated(perm(myDownload(A,_)),C) :- occurred(myShare(A),C).*

et la deuxième hypothèse, comprenant un fait et une règle :

*terminated(perm(myDownload(_,_)),_).*

*initiated(perm(myDownload(A,_)),C) :- occurred(myShare(A),C).*

Où la première hypothèse est celle à laquelle nous nous attendions, alors que la deuxième – tout en étant compatible avec les traces – exprime une hypothèse beaucoup plus forte : avant qu'une permission d'acquisition puisse être donnée, il faut que, au préalable, plus aucun agent n'ait une telle permission. La génération d'une telle hypothèse supplémentaire s'explique, en l'occurrence, simplement par le fait que le cas traité ne comporte que deux agents.

Un tel procédé pose avec acuité la question de savoir si l'induction permet de *créer* une règle nouvelle, ou si, au final, elle ne fait que restituer une règle implicite ? Nous aborderons plus amplement cette question au paragraphe suivant.

### 3.3.4. Diffusion de la norme

L'auteur distingue encore une troisième forme de création de normes, par apprentissage social, c'est-à-dire que les agents acquièrent des normes par « crainte » des punitions que d'autre agents du groupe auquel ils appartiennent pourraient leur infliger. Or, l'apprentissage social est d'une autre nature que les deux autres formes. Dans l'imitation, il y a une authentique création de norme, ne fût-ce que celle qui édicte de coller à ce qui a été observé – et encore, nous formulons la norme de façon très générale : plus qu'une norme de grégarisme, la norme n'est alors rien d'autre que la description du comportement érigée en modèle à suivre. Dans le cas de l'apprentissage cognitif, il y a un véritable travail conceptuel : les comportements deviennent autant de propositions de normes. En revanche, dans l'apprentissage social, il n'y pas de conversion d'une observation en norme. Si autrui me punit pour un comportement, c'est que la norme existe déjà. Il n'y a donc point émergence, ni même modification, mais seulement apprentissage individuel de la norme : c'est l'objet de la présente section.

Nous nous arrêterons sur deux façons – pour un groupe – de faire respecter ses normes. La première façon est héritée de la théorie des jeux : elle consiste à « punir » un contrevenant aux normes de façon abstraite, quantitative[63]. Ainsi, le « jeu des normes » part d'une situation proche d'un dilemme de prisonniers : un agent peut soit coopérer, soit faire défection. S'il fait défection, il en tire un gain (*payoff*, par exemple 3), tandis que les autres agents subissent une perte légère (par exemple -1) ; en revanche, s'il coopère, le score de chacun reste stable. Le jeu diverge cependant d'un dilemme classique par la possibilité que les autres agents voient la défection – avec une certaine probabilité à paramétrer – et décident de punir le tricheur. La pénalité est généralement très lourde (par exemple -9). Infliger une telle pénalité a cependant un coût que doit supporter l'agent vengeur (par

[63] Nous nous baserons ici sur le chapitre *Promoting Norms* de R. AXELROD, *The Complexity of Cooperation*, pp. 40-68.

exemple -2). Un agent est donc caractérisé par son audace (*boldness*), qui détermine sa propension à tricher, ainsi que par sa volonté punitive (*vengefulness*). Le jeu ainsi décrit peut être simulé dans un tournoi : partant d'un choix aléatoire de joueurs, l'étude porte sur l'évolution de la population au fil des parties.

Nous pouvons alors observer une dynamique en accordéon : en début de tournoi les niveaux d'audace et de volonté punitive sont moyens. Dans un premier temps, le niveau d'audace chute de façon draconienne, à cause du risque rédhibitoire d'écoper une lourde pénalité. Ensuite c'est au taux de volonté punitive de chuter, puisque le coût de la punition est relativement élevé et sans retour sur investissement notable pour celui qui la pratique. Or dès que la volonté punitive s'effondre, il redevient intéressant de tricher, et l'accordéon peut de nouveau s'ouvrir. L'exemple est toutefois très simpliste : à la réserve que nous avons déjà exprimée quant au sens des quantifications s'ajoute ici le rapport moins évident que ne le semble croire l'auteur entre la notion de norme et l'opposition toute binaire entre « collaborer » et « faire défection ». Des approches plus élaborées de la même idée aboutissent alors à des systèmes où la norme est considérée comme contrat, comme nous l'avons déjà vu.

L'acquisition ne doit pas toujours prendre un aspect punitif : les récompenses sont également possibles. Il n'en demeure pas moins que tant le bâton que la carotte soient toujours de nature quantitative. Or dans une approche strictement quantitative, le respect de la norme se confond avec la maximisation de l'utilité et la notion même de norme s'efface. En d'autres termes, un téléologisme utilitariste, diamétralement opposé en ceci à une conception aristotélicienne, ne peut accommoder la déontologie sans se renier[64]. Dans d'autres approches plus expressives[65], l'acquisition revêt également un aspect plus qualitatif, en ceci qu'elle en appelle aux émotions des agents. Les auteurs définissent l'émotion fonctionnellement sur le niveau intra et inter-agents. Au plan intra-agent, l'émotion a une fonction double : d'une part, elle sert d'indicateur qui « informe » l'agent sur un évènement dans l'environnement qui requiert une réaction immédiate et adaptative ; d'autre part, elle induit les changements physiologiques nécessaires pour que l'agent réponde adéquatement à ce que la situation requiert de lui. Au plan inter-agents, l'émotion fournit des indices informatifs à l'interlocuteur pour améliorer l'interprétation réciproque, mais aussi des indices directifs : motivations ou sanctions (« mieux vaudrait changer de sujet… »). En d'autres termes, il est possible de donner une *incitation positive*, un *encouragement*, à son interlocuteur. La récompense – qui dans sa forme la plus fruste est une simple consolidation de l'amour-propre, une fleur offerte à notre éthos – fait ainsi également partie du mécanisme d'acquisition.

L'illustration que les auteurs proposent renvoie à une situation tout à fait similaire à celle retenue par notre discussion sur l'approche quantitative : un agent peut collaborer ou faire défection, d'autres agents peuvent observer soit un comportement conforme, soit un comportement qui enfreint les règles établies. La punition d'un comportement non-conforme prend alors la forme d'une

[64] Nous n'ignorons pas que rien n'interdit l'approche utilitariste de porter sur les normes : il s'agit alors de déterminer la norme qui maximise l'utilité. Il est clair, cependant, que l'utilitarisme devient alors un mécanisme secondaire, qui se résume à une heuristique d'arbitrage entre normes en concurrence, et ce de façon tout à fait *ad hoc*.

[65] Voir notamment les articles de Chr. von SCHEVE, D. MOLDT, J. FIX et R. von LÜDE, *My Agents Love to Conform*, ainsi que J. FIX, Chr. von SCHEVE et D. MOLDT, *Emotion-based norm enforcement and maintenance in multi-agent systems*.

création d'émotion à valence négative : le mépris (*contempt*), le dédain (*disdain*), le dégoût (*disgust*). L'agent fauteur ressentira alors des sentiments de honte (*shame*), culpabilité (*guilt*) ou d'embarras (*embarrassment*). Ces sentiments à leur tour peuvent contribuer à l'internalisation de la norme : nous retrouvons alors une problématique vue plus haut (§ 3.3.1). Notons, cependant, que même si les auteurs mentionnent la possibilité de l'encouragement – et que rien ne s'y oppose dans leur prototype – ils se consacrent quasi exclusivement à la sanction punitive[66]. L'idée est cependant là : une émotion, *fonctionnelle et symbolique*, permet par des moyens relativement simples une expressivité dans la modélisation dont est absolument dépourvue une approche purement quantitative.

### 3.3.5. Les organisations et leurs rôles

John Rawls a introduit la notion de « voile d'ignorance » en éthique : dans sa vue contractualiste, il faut négocier la position sociale de chacun dans *l'ignorance totale* de sa propre situation dans la société historique. Quoique les idées de Rawls aient pu être influentes, leur fondement paraît mal assuré. Ainsi Ricœur a argumenté[67] que les droits d'une personne ne précèdent pas le lien sociétal et que la position de chacun contient nécessairement des éléments d'appréciation socio-historique ; en d'autres termes, une part importante de contingence historique est inextricablement liée à l'évaluation que nous pouvons faire de la distribution idéale des rôles, devoirs et avoirs de tout un chacun. L'image de Rawls met néanmoins le doigt sur une exigence éthique qui pèse sur les fondements du vivre-ensemble : nous nous situons donc sur le plan impersonnel, celui des organisations, dont l'État est traditionnellement la forme la plus aboutie.

Or nous avons vu, au deuxième chapitre (§ 2.2.2), que les architectures multi-agents se sont enrichies d'une « couche » organisationnelle ; l'exemple que nous avions cité était la plate-forme Jason, où l'extension Moise permettait de modéliser des structures de type organisationnel, avec tout ce qu'une organisation apporte en termes de rôles, de droits, etc.

Il en temps de parcourir plus en détails Moise, et notamment le rôle que peut y jouer la logique déontique. L'article sur lequel nous allons nous appuyer maintenant[68] prend effectivement Moise comme point de départ pour modéliser un scénario de gestion de crise, où il s'agit de coordonner différents services de secours dans une situation d'urgence. Mais d'abord, rappelons la base de Moise : cette plate-forme permet de modéliser des organisations selon trois dimensions : une dimension structurelle, une dimension fonctionnelle ainsi qu'une dimension normative.

La dimension structurelle spécifie les rôles, les groupes et les liens d'une organisation, y compris les relations de compatibilité entre les rôles, ainsi que le nombre de rôles qui peuvent être répartis entre

[66] La priorité – très visible dans les approches quantitatives à cause de leur transparence – donnée aux aspects punitifs de la norme est assez frappante. Est-il cependant nécessaire que l'acquisition prenne toujours un aspect punitif ? La norme serait-elle non seulement *épreuve* – comme le voulait Ricœur – mais aussi *chicotte*, qui meurtrit nos chairs récalcitrantes à la sociabilité ?

[67] P. RICŒUR, *Soi-même comme un autre*, p. 213.

[68] O. BOISSIER, Fl. BALBO et F. BADEIG, *Controlling multi-party interaction within normative multi-agent organizations*.

les différents groupes. Les liens de l'organisation établissent des devoirs de communication ou d'autorité entre les rôles. La dimension fonctionnelle définit tous les buts individuels ou collectifs possibles dans l'organisation. Ces buts sont regroupés en un ensemble de plans ou schèmes sociaux : il s'agit d'arborescences qui hiérarchisent entre eux les différents buts individuels et collectifs. Lorsqu'un agent est appelé à participer à un tel schème, la mission spécifie, sur la base des rôles des agents, la rétribution des buts du schème sur les différents agents appelés à participer.

La dimension normative, quant à elle, définit un ensemble de normes. Une norme est ici modélisée au travers d'un identifiant, une condition d'activation *c*, le rôle *r* et la mission *m* concernés, et la modalité déontique (permission ou obligation). L'idée est que la dimension normative jette le pont entre les deux autres dimensions : quand la condition d'activation *c* est remplie, tout agent ayant le rôle *r* a l'obligation (ou la permission) de participer à la mission *m*. Une norme est dite « active » quand la condition d'activation c est remplie ; elle est dite « accomplie » quand l'expression déontique a pu être vérifiée ; sinon, elle est dite « inaccomplie ».

La difficulté soulevée dans l'article est que Moise, en tant que tel, ne prévoit aucune modélisation pour les modes de communication entre agents en fonction de leur rôle dans l'organisation. À ce titre, il lui manque quelque chose d'essentiel, car une organisation est-elle autre chose qu'une *mise en relation indirecte*, là où la réciprocité directe entre agents ne peut être établie[69] ? Afin de pouvoir modéliser de façon pertinente le plan impersonnel, il faut donc ajouter des moyens de communication indirecte à Moise. Heureusement, C'est pourquoi il est heureux que les auteurs aient voulu enrichir la plateforme en ajoutant une dimension supplémentaire, la dimension communicative, qui spécifie un ensemble de modes de communication à appliquer sur les liens organisationnels entre les rôles de la dimension structurelle. Un lien possède trois propriétés : la source du lien, la cible, et le groupe dans lequel le lien est défini. Les modes de communication définissent, eux aussi, trois propriétés : d'abord, le type de communication, direct ou indirect, c'est-à-dire basé sur l'identifiant de l'émetteur seul ; ensuite, la direction : nous avons le choix entre unidirectionnel ou bidirectionnel ; enfin, le protocole technique de communication : pour ne citer qu'un exemple, il peut s'agir de FIPA ACL, protocole multi-agents que nous avons déjà rencontré au deuxième chapitre (§ 2.2.3).

Les modes de communication se rapportent aux liens par un nouveau type de normes, les normes de communication, qui portent sur les liens entre les différents rôles. Pour exprimer ces normes, les auteurs réutilisent une spécification communicationnelle existante : EASI (pour *Environment as Active Support for Interaction*). EASI est basé sur le concept de filtres environnementaux. Il y a trois types de filtre : un filtre sur les destinataires, un filtre sur le type ou sur le contenu des messages, enfin un filtre sur le contexte. Ce dernier filtre peut notamment être utilisé pour exprimer des contraintes sur l'émetteur du message ; il peut également faire référence au but ou à la mission poursuivie.

[69] Cf. P. RICŒUR, *Soi-même comme un autre*, p. 234 : « L'institution en tant que régulation de la distribution des rôles, donc en tant que système, est bien plus et autre chose que les individus porteurs de rôles. Autrement dit, la relation ne se réduit pas aux termes de la relation. ? Mais une relation ne constitue pas non plus une entité supplémentaire. Une institution considérée comme règle de distribution n'existe que pour autant que les individus y prennent part. »

Il devient ainsi possible d'exprimer des choses comme : un agent dans le rôle de coordinateur engagé dans la mission *m1* est obligé d'utiliser le lien l1 avec le mode de communication mc1, où *mc1* est un mode de communication direct, bidirectionnel, recourant à FIPA ACL. La dimension communicative permet ainsi un routage très expressif de messages à l'intérieur de l'organisation. Il est important de souligner que les agents ont un accès réflexif à la dimension communicative. Quand un agent reçoit un message, il sait si ce message lui a été adressé via un mode direct ou indirect ; et le cas échéant, de quelle manière il est censé y répondre. Une vie organisationnelle beaucoup plus dense en relations s'offre ainsi aux agents qui évoluent dans un tel cadre. L'autonomie, pouvons-nous dire en guise de conclusion, comporte aussi un volet collectif, volet que seul un cadre multi-agents peut adéquatement modéliser.

Attirons, pour finir ce survol théorique des SMA, l'attention sur le chemin parcouru par de tels systèmes : le paradigme, sur fond du souci d'affranchir le flux de contrôle d'une intelligence centrale, semble par une curieuse ironie du sort destiné à se positionner sans cesse face à l'unité perdue. C'est ainsi que des couches de savoir commun, d'interfaces partagées entre le monde extérieur et les agents, ne cessent de se multiplier à mesure que ces technologies se développent, comme s'il fallait toujours développer des relations plus intimes entre ces entités nées d'une séparation consentie à regret.

## 3.4. Études de cas

### 3.4.1. La tentation de l'intelligence centrale

Nous avons vu se dessiner une tension, dans les systèmes multi-agents, entre contrôle central et contrôle décentralisé. Pour notre recherche, à chaque fois qu'une intelligence (ou contrôle) centrale est utilisée, la question du « qui ? » revient : pouvons-nous architecturer un système social organisé d'une certaine complexité, « intelligence » si l'on veut, sans faire appel à l'équivalent d'un dieu horloger ? Dans cette interrogation, la SF peut nous aider. Dans la suite de cet exposé, nous aurons recours aux deux œuvres suivantes : la nouvelle du poète Marcel Thiry *Le Concerto pour Anne Queur*[70] et le roman de Gilbert Hottois, *Species Technica*[71].

Mais avant de continuer, il n'est pas inutile de nous expliquer sur ce recours à la littérature de science-fiction. Isabelle Stengers[72] compare la littérature de science-fiction à des expériences de pensée, dont elle fait ressortir la spécificité en les opposant aux expériences de pensée « logicistes », dont la chambre chinoise de Searle est l'emblème. Le *monde* dans lequel est située la chambre n'existe que pour les besoins de l'argument philosophique ; entièrement clos sur sa propre abstraction, il est par là même évanescent, fragile. À de telles expériences de pensée, syllogismes

[70] Publiée pour la première fois en revue en 1949, la nouvelle sera incluse dans le recueil *Nouvelles du Grand Possible* de 1960. Nous nous référons ici à l'édition parue en 1987 dans la collection patrimoniale « Espace Nord ».
[71] Écrit en 1981, le roman n'a paru que vingt ans plus tard, aux éditions Vrin.
[72] I. STENGERS, *Science-fiction et expérimentation*, dans G. HOTTOIS, *Philosophie et science-fiction*, pp. 95-113.

travestis en images, la SF expérimentale fait exister des mondes denses, foisonnants et riches, dans lesquels l'auteur ne fait que *déléguer*, dans une démarche exploratoire, des *observateurs partiels*.

La nouveauté explorée par la SF se joue dans deux registres : en plus des innovations proprement techniques, qui sont des conditions nécessaires à l'intrigue, le genre éclaire les retombées des nouveautés techniques sur les manières de vivre, de percevoir et d'être affecté des personnages qui peuplent le monde mis en scène. C'est dire que la SF explore les conséquences concrètes d'une hypothèse au départ abstraite. Ce faisant, elle contribue à un *diagnostic des possibles* – elle ouvre des voies à la pensée, là où des démarches probabilistes, visant le ou les cours d'évènements les plus plausibles, ont trop vite tendance à exclure d'emblée de larges pans du possible ; vue sous cet angle, la SF contribue à *ralentir la pensée*, vertu cardinale pour tout exercice de la philosophie.

Force est de constater, toutefois, que dans l'article que nous venons de citer, Stengers ne pense pas d'abord – ou en tout cas pas seulement – à la SF d'anticipation technique : pour ne citer qu'un exemple, elle commente le roman *The Mists of Avallon*[73], où l'auteur explore, à travers la prêtresse Morgane, d'autres rapports à la vérité que la fascination gréco-chrétienne pour l'Un. Les romans d'histoire ou fantastiques sont, en effet, tout aussi aptes que la SF proprement dite à servir de terrain au type d'exploration prôné par Stengers, tout autant capables de créer un monde possible, dans lequel un observateur peut être dépêché pour récolter des faits pertinents à l'interrogation problématisée[74].

Sur ce point, nous pouvons dire qu'un auteur comme Gilbert Hottois va plus loin, médite plus en avant la portée de la SF pour la philosophie[75]. Hottois la place dans le contexte d'une réaction contre l'inflation langagière dans la pensée de la seconde moitié du vingtième siècle : qu'elle soit d'obédience phénoménologique ou analytique, la philosophie d'après-guerre a tendance à ne plus porter que sur des productions langagières ; ce faisant, en abandonnant sa visée référentielle, elle se coupe du monde, s'enferme dans un ordre symbolique qu'elle érige en absolu. Or la SF fait indubitablement preuve d'une volonté référentielle : même si les dimensions psychanalytiques, socio-politiques etc., ne sont pas absentes de son univers, elle constitue le témoin d'un imaginaire

---

[73] *Ibid.*, pp. 101-102.

[74] La tentation est forte de juger les mondes *possibles* autres que les nôtres comme *improbables* et, de là, comme peu *crédibles*, voire comme peu *plausibles*. Or cette tentation n'a rien de rationnel : comme le dit si bien l'ingénieur épris de rationalité du roman de Max FRISCH: « Ich brauche, um das Unwahrscheinliche als Erfahrungstatsache gelten zu lassen, keinerlei Mystik ; Mathematik genügt mir. Mathematisch gesprochen: Das Wahrscheinliche (daß bei 6 000 000 000 Würfen mit einem regelmäßigen Sechserwürfel annäherend 1 000 000 000 Einser vorkommen) und das Unwahrscheinliche (daß bei 6 Würfen mit demselben Würfel einmal 6 Einser vorkommen) unterscheiden sich nicht dem Wesem nach, sondern nur der Häufigkeit nach, wobei das Häufigere von vornherein als glaubwürdiger erscheint. Es ist aber, wenn einmal das Unwahrscheinliche eintritt, nichts Höheres dabei, keinerlei Wunder oder Derartiges, wie es der Laie so gerne haben möchte. Indem wir vom Wahrscheinlichkeit sprechen, ist ja das Unwahrscheinliche immer schon inbegriffen und zwar als Grenzfall des Möglichen, und wenn es einmal eintritt, das Unwahrscheinliche, so besteht für unsereinen keinerlei Gründ zur Verwunderung, zur Erschütterung, zur Mystifikation. » (*Homo faber*, p. 22). Ce qui donne à ce passage toute sa saveur, c'est que la langue allemande ne distingue pas fondamentalement – comme le fait le français – entre le *probable* mathématique et le *vraisembable* de l'opinion, mais les confond sous le même vocable « wahrscheinlich ».

[75] Nous nous basons ici sur son ouvrage *Généalogies philosophique, politique et imaginaire de la technoscience*, plus particulièrement sur les chapitres *Origine philosophique et science-fictionnelle de la technoscience* (pp. 25-57) et *La technoscience illustrée dans la science-fiction* (pp. 158-240).

centré sur le futur et la technoscience. Pour Hottois, le concept de « technoscience » est capital, car il met le doigt sur une caractéristique importante de la science moderne : celle-ci n'est ni essentiellement représentation, ni discours. Elle abandonne la pose contemplative (ou théorétique) pour mettre résolument en avant le rapport *opératoire* de l'homme vis-à-vis du cosmos.

Ces diverses thématiques – dépassement du logocentrisme, rapport opératoire à la vérité, primat du futur – se rejoignent ultimement dans la question de la *fin* de l'homme. La fin de l'homme doit être comprise ici non comme sa finalité ou son destin, mais de façon beaucoup plus littérale : comment et quand, dans quelles circonstances, l'espèce humaine viendra-t-elle à prendre fin ? Une telle fin peut être envisagée comme un anéantissement pur et simple, suite à quelque catastrophe. Le plus souvent cependant, la SF nous donnera à voir l'une ou l'autre forme de *mutation* : en donnant naissance à autre chose qu'elle-même, l'humanité fait face, dans les œuvres de SF, à une forme radicale d'altérité, que Hottois appelle « abhumaine »[76]. L'alterité abhumaine constitue comme l'horizon, la ligne de fuite, de la technoscience, en tant que celle-ci exerce alors sa saisie opératoire sur l'être humain lui-même. C'est ainsi que se trouve posée la *question éthique fondamentale* : l'homme est-il *l'avenir* de l'homme ?

Voilà la toile de fond sur laquelle se détache l'intérêt philosophique pour la SF. Même si cette approche n'est pas exempte de critiques[77], nous ne pouvons que confirmer que la nouvelle que nous allons commenter maintenant porte un regard original – et très imagé – sur cette même question. Le protagoniste de l'histoire, le docteur Cham, est un scientifique actif dans une université belgo-américaine au Kivu. Il y a inventé une technique de « revivification » des morts, mais sous une forme absolument singulière : hormis le cerveau et l'ossature, toute la matière biologique du cadavre est dissoute dans un bain acide ; le squelette est alors armé de capteurs visuels et sonores, ainsi que d'un système d'irrigation artificiel du crâne. La nouvelle s'attache à décrire la courte période de cohabitation des « Secs » et des êtres humains, avant que les relations se dégradent au-delà du réparable et qu'un conflit armé éclate. Cependant, alors que tant aux États-Unis qu'en France, les colonies des Secs sont assiégées, ces derniers optent pour une issue radicale : un suicide collectif. Après quoi, toutes les traces de la présence des Secs – leurs écrits, les technologies nécessaires à la « dessiccation », etc. – sont détruites.

---

[76] En écrivant ces lignes, Gilbert Hottois n'avait pas encore connaissance du mouvement transhumaniste et utilisait presque indifféremment « abhumanisme » et « transhumanisme ». Il dit cependant préférer le terme d'« abhumanisme » car celui-ci est moins sujet à la confusion finaliste : la pensée transhumaniste projette souvent dans le projet de transformation de l'homme une dimension quasi religieuse, dans la mesure où elle y voit le *destin* de l'homme (sur le lien entre transhumanisme et destin cosmique de l'homme, voir J.-Y. GOFFI, *Les transhumanismes, la technique, la terre et l'espace*).

[77] Force est de constater que si la caractérisation de Stengers avait tendance à brasser trop large, Hottois tend à restreindre excessivement le champ de la « bonne » SF : ainsi il l'oppose à la « politique-fiction » (*Généalogies philosophique, politique et imaginaire de la technoscience*, p. 33) qu'il fait remonter à la nouvelle *Micromégas* de Voltaire. La critique de l'inflation langagière ne semble en outre pas tout à fait convaincante pour parler d'un genre qui reste très traditionnellement romanesque, très attaché à la mimèsis et à la mise en intrigue, qui sont l'apanage de l'ordre symbolique. À cet égard, le refus de la figuration porté par le mouvement futuriste – qui précède de peu l'apparition de la SF moderne – aurait peut-être mieux convenu au propos de Hottois ; il est cependant clair que ces considérations nous éloignent par trop du sujet du présent mémoire.

Bien sûr, la nouvelle invite tout d'abord à une réflexion sur le transhumanisme : le point de vue « majoritaire » dans le roman étant celui du docteur Cham, homme profondément religieux, qui voit dans le procédé qu'il a mis au point une victoire de « l'âme immortelle » sur la « chair » ; ce vocabulaire religieux est omniprésent dans sa bouche et sa pensée : le passage de l'être humain à squelette ambulant, même si extérieurement source d'effroi, est ressenti comme une « résurrection » (p. 197), une « délivrance » (p. 198). Le succès est d'ailleurs au rendez-vous. Libérés des pesanteurs de la chair, les Secs font preuve d'une intelligence hors mesure :

> *Il [le docteur Cham] goûtait à la fois, après des années de transes scientifiques et religieuses, un premier repos, le succès éclatant de sa découverte et l'espoir de voir ses miraculés préparer sur la terre le règne de l'Esprit. Et cet espoir commençait à se réaliser sous ses yeux avec des développements si rapides qu'il hésitait à s'en convaincre. La spéculation qu'il avait conçue sur les possibilités de l'intelligence humaine, une fois qu'elle serait dégagée de toutes les lourdeurs de la matière, se révélait juste en son principe, mais combien timorée dans sa mesure ! Les dix-huit pensionnaires de la Malmaison dépassaient tous les jours, de toute la projection de leur intuition « délivrée », les étapes qu'il avait aménagées pour la marche de leur éducation scientifique et morale. Il avait dès longtemps préparé pour eux une sorte d'initiation à la somme des connaissances humaines ; et voici que les hommes-cerveaux en franchissaient les degrés avec une vitesse inégale, suivant les dons respectifs qu'ils avaient possédés dans leur vie première, mais toujours stupéfiante.*[78]

À part le transhumanisme religieux exprimé par le docteur Cham[79], d'autres points de vue sont également évoqués dans la nouvelle : ainsi un courant – minoritaire – parmi les Secs, les Virginiens, sont attachés à la nostalgie de la chair. Ils se regroupent autour d'Anne Queur, jeune fille tuée accidentellement par son père et le seul être humain de la nouvelle à subir l'opération sans son consentement. Le point de vue pragmatique, proprement « abhumain » selon Hottois, se trouve également exprimé :

> *[…] la Fondation scientifique belgo-américaine, à laquelle appartenait l'université dirigée par Cham, avait à l'insu de l'inventeur africain, pendant que la presse le vilipendait, que les cardinaux conféraient à Rome et que les ambassadeurs remettaient des aide-mémoire, décidé de son sort et de l'avenir de sa découverte. […] Après avoir pesé d'une part le coût très élevé de la survie (l'appareillage d'un cadavre valait une fortune et sa combustion de sang synthétique coûtait beaucoup plus cher que l'alimentation d'un vivant), et d'autre part*

[78] M. THIRY, *Nouvelles du Grand Possible*, p. 201.

[79] Tout porte à croire d'ailleurs que la voix du transhumanisme religieux est celle du poète lui-même ; nous avons trouvé une confirmation dans l'œuvre poétique, plus précisément dans la pièce *Prose des cellules He La* parue en 1969 (M. THIRY, *Œuvres poétiques complètes*, pp. 407-412). Dans cette « prose », Thiry se fait l'écho de l'histoire d'une prénommée Helen Lane (de son vrai nom Henrietta Lacks), victime d'un cancer particulièrement agressif. Une souche de cellules proliférantes fut prélevée sur elle en 1948, dont les descendants allaient bientôt peupler tous les laboratoires biomédicaux de la planète. Ainsi donc, Helen « fut marquée entre toutes les femmes » (p. 410), élue pour l'immortalité, « alors que son corps n'est même plus cadavre » (ibid.). Son corps n'est plus, mais ses cellules sont devenues à elles-mêmes toute une galaxie, au plus grand profit de la médecine : « Tu fus livrée au grand hourra silencieux des hordes cellulaires libérées, Et maintenant que ton martyre t'a gagné la métamorphose Tu vis pour nous éparse parmi nous en rose poussière d'étoiles charnelles dans nos laboratoires… » (p. 411). Une fois de plus donc, vocabulaire religieux et abhumanisme spectaculaire se conjuguent.

*l'extraordinaire rendement intellectuel de ces créatures qui ne connaissaient pas le sommeil et qui montraient pour toutes les sciences des aptitudes monstrueuses, les conseillers anonymes jugèrent que ce bilan laissait un solde bénéficiaire et que la découverte était rentable. Ils en conclurent que cette force nouvelle des Purs Cerveaux ne devait pas demeurer aux ordres du Dr. Cham, monopolisée, éduquée par lui seul, employée à cette seule fin zélatrice qui était la sienne ; elle devait, cette force, entrer dans le circuit des valeurs que manœuvraient les sénats scientifiques pour organiser une humanité meilleure.*[80]

Bien sûr, le rejet total de cette nouvelle forme de vie – et qui va l'emporter à la fin de la nouvelle – est très présent aussi ; elle s'exprime le plus souvent de façon grossière, par un attachement quasi bestial aux voluptés de la chair :

*La chair, c'est tout ce qu'ils ont, c'est tout ce qu'ils sont. Bien grossière, bien malheureuse, bien disgraciée ; mais à eux, et leur seul mode d'exister.*[81]

En privilégiant une vue dualiste de l'homme, qui distingue radicalement l'esprit-raison et le corps-volupté, Thiry se situe dans une lignée que nous pouvons qualifier de kantienne, à comprendre comme un rationalisme transhumain[82] : la raison est universelle, transcende l'homme, dans la corporéité duquel elle se trouve tout autant limitée qu'elle n'y est réalisée :

*[…] rien de ce qui est spécifiquement humain n'est réellement digne ; ce qui est lié à la particularité humaine désigne très précisément ce avec quoi il faut rompre pour être authentiquement rationnel. Kant parle de la « grossièreté de la machine et de la texture dans la nature humaine » et en fait la « cause de cette inertie qui maintient les capacités de l'âme dans un épuisement et une impuissance constants ». En tant qu'homme, l'homme constitue un être raisonnable particulièrement défectueux, dont Kant se plait à dresser la liste des défauts.*[83]

Cette conception de la rationalité a d'importantes retombées sur la moralité : l'homme est soumis à l'exigence morale de se dépasser pour découvrir les lois morales, dépassement qui dans les faits sera peut-être impossible :

*[…] on doit douter qu'un homme puisse être effectivement moral, dans la mesure où l'exigence posée par la raison pratique se révèle précisément transhumaine là où l'homme ne peut pourtant pas être plus que lui-même. Condamné à n'être qu'humain, l'homme aperçoit une exigence morale que son humanité devrait lui interdire de réaliser dans les faits.*[84]

---

[80] M. THIRY, *Nouvelles du Grand Possible*, pp. 217-218.
[81] *Ibid.*, p. 256.
[82] L'idée d'un transhumanisme kantien se trouve exprimé chez Th. GRESS et P. MIRAULT, *La philosophie au risque de l'intelligence extraterrestre*, pp. 93-155. Leur lecture de Kant se fonde sur l'inspiration – très appuyée au début de sa carrière – que Kant a trouvée dans les réflexions de Fontenelle et de Huygens sur la possibilité et les conséquences philosophiques de formes de vie rationnelles non-humaines, extraterrestres.
[83] Th. GRESS et P. MIRAULT, *op. cit.*, p. 114.
[84] *Ibid.*, p. 131.

La question du transhumanisme n'est pas la seule que soulève la nouvelle de Thiry ; il y a encore l'image - tout à fait intrigante – du « Vase ». Le Vase est une réalité qui va progressivement s'imposer à la communauté des Secs. Tout commence par le mode de communication entre Secs : leurs cerveaux, délestés de « l'épais travail des digestions et des décrétions multiples » (p. 211), se communiquent *sans langage*, par la radiation même de l'activité cérébrale. Le corollaire – somme toute logique – d'une telle forme de communication directe est la perte de la capacité de mentir :

> *Cham trouva là d'autant plus de raisons de haïr les tyrannies de la matière organique que les Cérébraux, il s'en aperçut avec émerveillement, ne pouvaient pas mentir. C'est leur pensée intime qui était automatiquement transmise, sans l'intervention de leur volonté. Et lorsque le professeur demanda, en manière d'expérience, d'essayer de former une espèce de mensonge mental, de penser « ce mur est noir », alors qu'il était blanc, il en résulta une superposition d'ondes qui brouillait le message et révélait la fraude. Donc le mensonge tenait à la chair !*[85]

Ce mode de communication instaure un mode de vie où chacun donne à la collectivité ses idées, ses pensées, ses savoirs, sans acception de ce qu'il recevra en retour. Cette culture de potlatch mental généralisé donne lieu à une « âme collective » (p. 241), une « pensée collective » (p. 242), que les Secs vont appeler « le Vase » (p. 242). Ces échanges incessants vont connaître leur apogée quand la communication mentale va elle-même être renforcée techniquement, par une sorte d'amplificateur des radiations :

> *Jusqu'alors les Cerveaux ne s'étaient communiqué mutuellement par ces ondes silencieuses leur teneur et leur travail qu'à la condition d'être en présence, c'est-à-dire à courte distance les uns des autres et se voyant ou se sachant proches. Cham alourdit encore le crâne des squelettes d'un engin qu'il y fit loger et qui intensifiait la portée des rayons mentaux jusqu'à les rendre perceptibles sur toute la surface du globe. Ainsi la conscience commune devenait vraiment totale ; elle ne dépendait plus des rencontres, des échanges plus ou moins fréquents de la vie quotidienne. C'est immédiatement que le Vase s'accroissait des trouvailles, des progrès qu'avaient réalisés la pensée de n'importe quel squelette en n'importe quel point du monde.*[86]

Thiry nous fait alors entrevoir ce que ce mode de vie a en propre, la beauté qui est la sienne, malgré l'horreur de son apparence macabre :

> *Notre imagination s'effare de ce commerce incessant d'idées, s'entrecroisant en tous les sens et tissant autour de la terre un réseau continuel de raisonnements et de songeries. Elle se représente mal comment les Cerveaux ainsi noyés dans cet immense trafic de concepts pouvaient les démêler et les organiser.*[87]

C'est peut-être là aussi que Thiry accède à la plus grande originalité technique, tout en se servant d'une image aujourd'hui désuète, les émissions radiophoniques :

---

[85] M. THIRY, o*p. cit.*, p. 211.
[86] *Ibid.*, p. 246.
[87] *Ibid*.

> *Il est concevable que l'attention jouait le rôle de sélecteur. On peut essayer de se figurer le Pur Cerveau régnant sur les ondes mentales comme nous régnons sur les ondes hertziennes ; son attention déterminait la longueur des ondes qu'il voulait se rendre sensibles, comme nous appelons à notre gré un bulletin d'information ou un concert de musique de chambre, la voix de Londres ou celle de Paris. Pour mieux adapter la comparaison avec la radio, il faut admettre en plus que l'auditeur dispose non seulement des émissions en cours à l'instant où il tourne le bouton de son appareil, mais aussi d'une collection universelle d'enregistrements dans laquelle il peut puiser, comme les Cerveaux pouvaient puiser dans la réserve totale, la mémoire du Vase.*[88]

Le Vase nous rappelle vaguement l'internet, toutefois il ne faut pas oublier la date de première publication de la nouvelle : 1949 ; elle est donc contemporaine des premiers ordinateurs, précède de peu les ordinateurs à transistors (années '50) et elle précède de beaucoup la présentation d'Arpanet (l'ancêtre direct de l'internet actuel) en 1972. Toutefois Thiry met ici le doigt sur le mode propre d'efficacité de l'informatique, qui – si nous suivons Jacques Ellul[89] – est la mise en réseau, dans le style très expressif qui est le sien :

> *[…] il est parfaitement vain et inutile de parler de l'ordinateur pris comme une unité. […] Considérer un ordinateur c'est en rester à la mentalité du badaud à la foire qui va voir l'homme-tronc ou la femme à barbe.*[90]

Au fil de la nouvelle, l'individualité des squelettes, vestige pour ainsi dire historique, finit effectivement par s'effacer pour laisser place à une intelligence centrale, dont les squelettes ne sont que des émanations locales. Le Vase devient alors non seulement un réceptacle de connaissance, mais un véritable organe de décision commune :

> *Quand ainsi fut unifiée en une seule conscience, un seul jugement et un seul vouloir la pensée éparse des squelettes répandus sur les continents, quand la somme des connaissances et des aspirations contenues dans le Vase se fut synthétisée dans le dessein de sauver l'humanité en lui enseignant à se déprendre de la gangue charnelle, le Dr Cham, qui figurait l'exécutif de cette intelligence décrétante, put passer à l'action.*[91]

Ainsi, et moyennant une considérable ambiguïté sur les prérogatives exactes du Vase, nous assistons à une perte d'individualité mentale qui accompagne la perte du corps et ce non seulement dans le domaine de la connaissance pure et simple, mais aussi dans les domaines que nous aurions probablement voulu garder privés, dont au premier chef la volition.

Les cerveaux interconnectés reviennent dans le roman de Gilbert Hottois, *Species Technica*. Situé dans un futur proche lors de sa rédaction, devenu entretemps notre présent, le roman a pour argument central la disparition du fils et de la femme du philosophe André Gillian. À l'occasion d'une

---

[88] *Ibid.*, pp. 246-247.
[89] Voir les pages denses consacrées au rôle de l'ordinateur dans l'organisation technicienne chez J. ELLUL, *Le Système technicien*, pp. 101-115.
[90] *Ibid.*., p. 111.
[91] M. THIRY, *op. cit.*, p. 247.

invitation à parler de son dernier livre lors d'une conférence dans un institut scientifique de pointe richement doté, Gillian va – presque par hasard, suite à une erreur technique – découvrir une face plus sombre de l'institut. De fil en aiguille, en parcourant le réseau auquel appartient l'institut, il va découvrir un projet aux ambitions dantesques, « le Fils de l'Homme ». Ce projet est piloté par le docteur Samuel Spinrad, qui s'en explique en ces termes :

> *Le Fils de l'Homme est un être complexe, mosaïque. Entendez-moi bien : cet être n'est pas le successeur de l'homme. C'est un chaînon. Mais un chaînon indispensable car le successeur de l'homme n'est pas à la portée de la pensée de l'homme. L'homme n'a pas la capacité d'imaginer, d'anticiper conceptuellement sa propre transcendance. C'est aussi un instrument. Mais un instrument au sens large. Nous sommes tous d'une certaine façon des instruments. Des instruments du Temps. Le Fils de l'Homme, c'est l'instrument de l'évolution le plus sophistiqué que l'on ait jamais conçu.*[92]

Kant n'eût probablement pas apprécié voir s'instrumentaliser l'homme au profit d'une finalité supérieure, fût-ce le Temps ; nous nous bornerons à la remarque que le docteur Spinrad introduit très clairement une finalité dans l'évolution en métaphorisant le temps qui passe. Nous retrouvons d'ailleurs le thème de l'autonomie de l'agent exprimé comme « fil téléologique » (§ 1.7.2.2 et 2.1.3), tant il est vrai que le Cyborg n'existe qu'en vertu de la finalité qui lui a été fixée :

> *La seule unité du Cyborg mosaïque réside dans le but qui lui est assigné, de l'extérieur et par nous pour ce qui est des composantes cybernétiques, de l'intérieur, comme un désir ou une volonté existentielle, pour ce qui est des composantes neurologiques humaines. Et ce but, c'est la détermination de la mutation technologique de l'homme, l'identification prospective de la « species technica » pour reprendre l'heureuse expression que vous utilisez. Le Cyborg mosaïque n'existe pas en fonction de lui-même, comme un individu. Il n'est qu'en fonction d'un projet de transcendance. Comme une collectivité s'absorbe dans l'entreprise commune.*[93]

Or cette finalité a un prix. Le Cyborg mosaïque doit en effet être « nourri » d'une certaine quantité de cerveaux humains pour mener à bien son projet. Pour être précis, le Cyborg n'est rien d'autre qu'une interconnexion d'unités de calcul informatiques classiques d'une part et cérébrales, d'autre part. Laissons encore la parole au docteur Spinrad :

> *[Le Cyborg mosaïque est] constitué par un réseau d'ordinateurs couplé à un jeu de composantes neurologiques humaines. Imaginez la puissance de conception et de résolution d'une pareille entité ! Composite, elle n'a aucune identité personnelle, bien qu'elle pense au plein sens humain de ce mot, puisqu'elle comprend une importante quantité de cette matière qu'on appelle grise. Et là est le trait de génie : c'est cette carence même d'identité qui est motrice de la quête du dépassement de l'anthrope.*[94]

Or la matière grise nécessaire à la fabrication du Cyborg provient d'êtres humains non consentants. Gillian apprend avec horreur – c'est le dénouement de l'enquête policière du roman – que son propre

---

[92] G. HOTTOIS, *Species Technica*, p. 146.
[93] *Ibid*., p. 147.
[94] *Ibid.*

fils a été ravi pour nourrir le moloch. Laissé seul avec le Cyborg, il cherche à savoir, dans un dialogue poignant, si ce dernier garde quelque trace des créatures qui lui ont été immolées :

> *— Je dis que je suis le père de… l'un des cerveaux qui vous constituent.*
>
> *— On comprend mais on ne saisit pas la pertinence ni la finalité de cette information. Les cerveaux qui nous co-constituent sont sans identité personnelle. Votre adresse est sans destinataire.*
>
> *Au prix d'un grand effort, Gillian s'imposa de poursuivre le dialogue infernal. Il fallait qu'il sût, en toute certitude, à quoi s'en tenir.*
>
> *— Mais ne reste-t-il pas, dans la mémoire associée à ces cerveaux partiels, le souvenir d'une ancienne identité ?*
>
> *— Aucun souvenir personnel. De tels souvenirs auraient été des corps étrangers, nuisibles à la quête et causes de réactions de rejet. Ils ont été tous éliminés lors de la phase préparatoire d'intégration.*
>
> *— Le nom de « Pierre Gillian » ne signifie donc rien pour vous ? Aucune trace, aucune association ?*
>
> *— Aucune. Si vous voulez des informations sur notre préhistoire, vous devez interroger, en manuel, la base de données G.X.4, à l'adresse « Projet Fils de l'Homme. Historique ».*[95]

Quel résultat se donne-t-il à voir dans cette tentation radicale de l'Un ? Dans le roman de Hottois, le résultat est résolument inquiétant : l'individu y est radicalement et absolument *supprimé*, vivier de cervelle qu'il devient même sans son consentement, au profit d'une fin supérieure à laquelle il n'adhère pas… La nouvelle de Thiry est à cet égard autrement nuancée : le monde des Secs y est décrit comme un foisonnement d'idées, non dépourvu d'intérêt, voire de beauté. Leur suicide collectif prend des allures d'un épouvantable gâchis, *d'une occasion ratée pour l'avenir de l'homme*[96].

Or dans les deux cas pourtant, des cerveaux sont « mis en commun » afin d'atteindre un but qui transcende les individus. Certes, à partir de ce même argument, les chemins divergent : ainsi chez Thiry, le but est découvert sur le tard – un danger suprême qui plane sur une humanité sous l'empire de la Chair ; les cerveaux gardent en outre leur personnalité, leurs souvenirs, leur liberté de mouvement. L'intelligence centrale n'apparaît que petit à petit, *émerge* pour ainsi dire de la perte de la corporéité. Il en va tout autrement chez Hottois : l'intelligence centrale prime ; les cerveaux apportés à son tribut ont été soigneusement dépouillés à l'avance de tout souvenir personnel. Ici, nul doute n'est permis : les cerveaux ne sont qu'une sorte de « matière première », dont toute individualité a été bannie. C'est d'ailleurs cet état de choses qui va convaincre Gillian de détruire le

---

[95] *Ibid*., pp. 151-152.

[96] En relisant la nouvelle, nous devons cependant nuancer le tragique de cette affirmation ; quelques notations discrètes, en effet, placent la nouvelle sous le signe du *phénix* : ce qui met le pouce à l'oreille, c'est la dédicace de la nouvelle à Jean Hubaux, auteur de renom d'une étude sur *Le mythe du phénix dans la littérature grecque et latine* (cf. M. DELCOURT, *In Memoriam*). Ainsi est suggérée – discrètement il est vrai – la possibilité d'une renaissance de la rationalité portée par les Secs, un jour où les mentalités de l'homme seront plus mûres.

Cyborg mosaïque. Il n'est donc pas défendu de parler, dans le cas de Thiry, d'un « désir positif de se transcender », pour reprendre le vocabulaire de Gilbert Hottois[97] : pourquoi pas dans la technique, plutôt que dans le langage ?

Car du langage, les deux histoires se méfient : chez Thiry, le langage humain est source de mensonge, que la communication directe entre esprits rend impossible (pp. 210-211) ; un Sec ne peut mentir qu'en usant d'une communication – langagière – écrite (pp. 237-238). Chez Hottois, l'épisode de Waha est instructif à cet égard (pp. 133-135) : Waha est un orang-outan à qui a été injecté le matériel génomique du langage. Le résultat est un bonimenteur verbaliste et infréquentable ; Hottois a par ailleurs pris un malin plaisir à donner aux répliques délirantes de Waha d'indéniables accents littéraires, de type avant-gardiste, lâchant les amarres de la référence dans un but d'expressivité pure.

La défiance à l'égard du langage mérite que nous nous y arrêtions un instant, étant donné qu'elle touche de près des thèmes développés au deuxième chapitre. La simulation, nous l'avons vu dans la section « Interfaces entre théorie et expérience » (§ 2.3.1), se fraie en effet un chemin *non discursif* dans les pratiques scientifiques. Plus précisément, la simulation s'est profilée comme un *degré de liberté* entre théorie et expérience – du moins si nous nous en tenons au point de vue de l'observateur. Rappelons-nous la distinction introduite – au deuxième chapitre encore, dans la section « De l'environnement à la simulation » (§ 2.2.3) – entre points de vue de l'observateur, de l'utilisateur et du joueur. L'observateur veut comprendre le monde ; l'utilisateur veut opératoirement agir sur lui ; quant au joueur, les effets qu'il recherche affectent surtout sinon exclusivement sa propre personne.

Nous pouvons dire maintenant que ces trois points de vue sont autant de modes de vie, de manières d'être, qui accueillent diversement en leur sein le langage. Nous venons de voir le rôle que peut avoir la simulation dans le mode de l'observateur, celui qui cherche à *connaître*. Passons sur le mode de l'utilisateur qui – presque par définition – n'entretient que des rapports lointains avec le langage. Il nous reste à aborder la manière du joueur, celui qui cherche le plaisir. Le plaisir peut être, à l'évidence, d'origine langagière : c'est le domaine de l'humour, pour ne citer que lui. Or dans ce mode d'être également, la simulation peut introduire une composante non-langagière, pourvu que des moyens techniques suffisants soient mis en œuvre. Nous en voulons pour témoin un autre récit de SF, *Excession*[98], dans lequel notre galaxie est administrée par un conseil d'intelligences artificielles, alors que l'humanité continue surtout à vaquer à ses mesquines préoccupations interpersonnelles – amours, jalousies, etc. Or il se fait que ces intelligences artificielles ont un passe-temps, à savoir la *métamathique*, l'exploration de mondes virtuels, des « champs de réalité intrinsèquement inconnaissables » dont l'intérêt tient aux infinies variations introduites par rapport au monde physique – rappelons-nous l'importance qu'a, en SBA, la variation induite par les paramètres d'entrée. La métamathique n'est en somme rien d'autre qu'une simulation grandeur nature, un « amusement sans fin ».

---

[97] G. HOTTOIS, *Transcendances symboliques et techniques*, dans ID., *Philosophie et science-fiction*, p. 134.

[98] Roman d'Iain Banks, dont nous avons pris connaissance par le rapport qu'en fait G. HOTTOIS, *Généalogies philosophique, politique et imaginaire de la technoscience*, pp. 209-213 ; la métamathique y est abordée pp. 212-213.

L'observateur voit ainsi la simulation à visée scientifique ; le joueur la voit dans les mondes virtuels auxquels la ludicité lui donne accès ; l'utilisateur recourra quant à lui volontiers à la simulation pour optimiser ses processus industriels ou logistiques. L'un de ses points de vue doit-il cependant prévaloir sur les autres ? Nous ne le pensons pas, ce serait encore une tentation de l'Un, *tentation d'ailleurs souvent exploitée en SF, au profit de l'utilisateur* : pour reprendre la terminologie de Hottois, le mode opératoire au réel s'affranchit alors de tout souci théorétique. Un tel affranchissement se trouve par exemple chez John Campbell, auteur et théoricien important de la SF : selon lui, la vérité est une catégorie dénuée de pertinence pour juger une théorie scientifique ; celle-ci n'est intéressante qu'en tant qu'outil pour intervenir dans le monde[99].

Nous pourrions croire, de prime abord, une telle vision du monde authentiquement fonctionnaliste – dans le sens où cette notion a été prise tout au long du premier chapitre – mais il n'en est rien : l'approche fonctionnaliste cherche à *comprendre* un comportement en l'imitant ; le tout-opératoire ne cherche pas à comprendre, seulement à créer de nouveaux modes de saisie. Nous voyons ainsi reparaître le spectre du « simulacre » que nous avons amplement commenté au deuxième chapitre sous la rubrique « Émergence du sens ou simulacre ? » (§ 2.3.2). En effet, nous sommes en plein dans la « fiction » de l'ingénieur comme Stengers l'a relevé : fiction, mais fiction efficace, une « culture de l'effet » qui ne manque pas d'aboutir à une conception magique du monde : ainsi le même Campbell qualifie volontiers la science de « magie qui marche » (*magic that works*)[100] ; nous pouvons retrouver la même idée chez les néo-sorcières décrites par Stengers : la magie y est comprise comme une « fiction efficace »[101].

De fait, les allusions « magiques » ne manquent pas dans les œuvres de science-fiction ; tout semble s'y trouver, des conceptions franchement délirantes jusqu'aux allusions ironiques. Dans les traitements qui nous semblent délirants, mentionnons le roman *L'énigme de l'univers* de Greg Egan[102]. Le roman relate les péripéties de la diffusion de la *Théorie du Tout* (ou TDT), dont la seule formulation peut avoir des effets directs dans le réel, exactement comme une parole magique : ainsi l'apprentissage maladroit de la théorie peut provoquer des maladies ; les effets directs d'une telle théorie vont jusqu'à évoquer des limitations produites par la théorie sur les lois de la physique elle-même[103].

Pour en venir aux œuvres que nous avons commentées, *Species Technica* se borne à mentionner l'existence d'un groupement de personnes vivement intéressées à développer les capacités

---

[99] Cf. G. HOTTOIS, *Généalogies philosophique, politique et imaginaire de la technoscience*, pp. 171-172.
[100] *Ibid.*, p. 176.
[101] Cf. le chapitre « *Reclaim* » dans Ph. PIGNARRE et I. STENGERS, *La sorcellerie capitaliste*, pp. 182-191. Précisons toutefois que la magie y est présentée comme une *technique*, comme un registre d'action qui ne se confond nullement avec une activité scientifique, en tout cas comme étrangère à l'ordre de la croyance ou de la garantie. Contrairement donc aux auteurs de SF, Stengers est attentive à ne pas confondre la culture de l'effet et la culture scientifique : ainsi lorsqu'elle fait état du baquet de Mesmer, *deux questions distinctes* sont posées : l'efficacité thérapeutique du baquet mesmérien ainsi que de toutes les thérapies « du toucher » traditionnelles d'une part ; le *pouvoir d'expliquer* ces mêmes guérisons, d'autre part (cf. I. STENGERS, *Au temps des catastrophes*, pp. 110-111, 114-115).
[102] Œuvre longuement analysée par G. HOTTOIS dans *Généalogies philosophique, politique et imaginaire de la technoscience*, pp. 214-240.
[103] Respectivement à la page 215 et 238 de G. HOTTOIS, *op. cit.*

parapsychologiques de l'homme[104]. Nous ne dépassons guère là l'énoncé notionnel. Chez Thiry, en revanche, le thème reçoit un traitement plus riche : tout d'abord, il exploite la possibilité – que la nouvelle partage avec la poésie – qui est de créer des renvois, des liens entre les pièces qui font *recueil*. Ainsi ce n'est sans doute pas un hasard si une nouvelle du recueil dans lequel *Le Concerto pour Anne Queur* a été inséré, traite ouvertement du thème : dans *Distances*, nous voyons apparaître une spirite, Mlle Ambert, qui communique avec les morts – la mort, dit-elle, n'est rien d'autre qu'une très grande distance[105]. Dans *Le Concerto* lui-même, le thème affleure également, lorsque la science des Secs prend des proportions quasi mesméristes :

> *Cham, à vrai dire, semblait manquer de moyens industriels pour les entreprises qu'il se proposait ; les laboratoires qu'on a trouvés dans les Malmaisons, et qui furent immédiatement détruits dans la fureur agnostique des gens de Gabriel, paraissent avoir été fort simples ; mais Cham devait disposer de procédés qui nous sont entièrement inconnus même dans leurs principes et qui semblent avoir tenu du magnétisme et du psychisme plutôt que de la chimie ou de la physique.*[106]

Nous voyons ainsi autant de mises en récit – de mises en chair – de thèses et d'hypothèses abordées tout au long de ce mémoire, à commencer par l'image de l'homme qui inspire nos éthiques : que désire l'homme, cet être de désir ? Le changement perpétuel en guise de liberté ? Le langage là-dedans n'est-il qu'une technique « du pauvre », accessible lorsque tous les autres moyens – intellectuels ou économiques – font défaut ? L'approche fonctionnaliste de type ingénieur peut-elle prétendre à une connaissance de type scientifique ? Il faudrait bien sûr s'entendre sur le sens à donner à un tel concept épistémologique très général. Quoi qu'il en soit, les œuvres de SF éclairent d'une lumière nouvelle le chemin parcouru par le paradigme multi-agents : que ce soit en génie logiciel ou dans les sciences qui recourent aux techniques informatiques, nous avons vu que l'abandon du contrôle social est vécu comme un *renoncement* ; il nous est désormais possible d'envisager le même parcours comme un *acquis*, un apprentissage de la spécificité de l'homme. S'ouvre alors un espace de possibles insoupçonnés, un axe autour duquel un nuage de points peut être exploré.

Aucune de ces questions ne sera cependant tranchée ici ; que ce soit sur la nature du savoir qui se dégage de la simulation multi-agents, sur la tension, pour ne pas dire l'aporie, entre diverses manières de saisir l'homme dans sa spécificité, l'exploration reste ouverte. D'ailleurs, une philosophie de l'altérité radicale, cosmique, reste à faire[107]. En littérature cependant, nous avons vu Maurice Maeterlinck faire une expérience authentique de l'altérité radicale, dans la figure de la fourmi qui se caractérise par son excès d'altruisme. À partir de là, il reconstitue la forme de vie de la fourmi, avec tous ses travers, ses défauts… ses qualités, évidemment, aussi. De son côté, la SF nous

[104] G. HOTTOIS, *Species Technica*, pp. 82-83.
[105] M. THIRY, *op. cit.*, pp. 50-52, 67-70.
[106] *Ibid.*, p. 248.
[107] C'est également la conclusion à laquelle aboutissent Th. GRESS et P. MIRAULT, *op. cit.*, p. 183.

permet, par la confrontation à des formes d'altérité radicales non terrestres[108], de concevoir autrement les liens entre éthique et image de l'homme.

## 3.4.2. Simulation à base d'agents et prise de décision

L'intelligence dont font preuve les agents en SMA n'a rien de contemplatif : elle est instrumentale, destinée à produire des effets dans l'environnement ; elle est instrument de *décision*. Il ne faut donc pas s'étonner si la simulation à base d'agents, dont le but est de modéliser une dynamique, peut servir d'outil à la prise de décision d'un observateur externe. Or des décisions se prennent à tous les échelons d'initiative et de pouvoir, sur un ample éventail de thématiques. Tenter, comme nous le faisons dans les lignes qui suivent, d'éclaircir dans quelles circonstances la SBA peut accompagner le *processus décisionnel* jettera également une lumière nouvelle sur les forces et limites de cette technique. Suivant en ceci notre source[109], nous ne nous intéresserons qu'à la prise de décision dans le contexte d'une décision *publique* sur une thématique qui touche au développement durable.

### *3.4.2.1. Processus décisionnel*

Comment prendre une décision éclairée[110] ? Dans un premier temps, il faut se documenter sur le système afin de s'en faire une représentation adéquate, identifier les contraintes, répertorier les actions possibles, estimer leurs effets. Dans un deuxième temps, ces effets doivent être valorisés à l'aide de critères de valeur, dérivés des préférences des décideurs, et agrégés et priorisés afin de rendre possibles des choix en fonction de certaines règles. Caractérisée de la sorte, la prise de décision paraît comme un double mouvement : le mouvement cognitif, qui se fait au début, cherche à ouvrir autant que possible le débat ; le mouvement axiologique le referme en hiérarchisant (selon les préférences) les possibles ; elle se fait à la fin du processus de décision.

---

[108] À ce titre, la SF n'est peut-être qu'une illustration d'un genre plus large, un type de littérature d'idées qui porte une réflexion sur l'altérité radicale. Alors que la SF, selon Hottois, est en premier lieu une réflexion sur la technoscience et que l'abhumanisme dérive de cette problématique première, dans cette autre conception primerait justement l'interrogation d'autres formes de vie, dont au premier chef la vie extraterrestre (que nous pourrions appeler la variante « exotique » de l'interrogation) tout comme la forme abhumaniste (la variante « inquiétante étrangeté » de l'interrogation). Une telle conception serait plus satisfaisante que celle de Hottois et ce, sous plusieurs angles. Tout d'abord, elle permettrait de rapprocher la SF avec certains romains historiques et/ou fantastiques (ainsi que nous l'avons vu faire Stengers plus haut). Ensuite, elle permettrait aussi de rendre compte que l'interrogation de l'altérité radicale puise dans des sources pré-techniques, comme le Golem ou Pygmalion, tout en évitant le risque d'être confondue avec la « gadgétophilie ». Finalement, elle présente encore l'avantage de distinguer nettement la SF de la politique-fiction : distinction dont Hottois sent la nécessité mais qu'il lui est bien malaisé d'opérer : or du moment que prime l'enjeu de l'altérité, il est évident que l'enjeu de la politique-fiction est aux antipodes de celui de la SF : la politique-fiction, en effet, ne s'intéresse à l'avenir que dans la mesure où celui-ci lui permet d'amplifier la (ou les) dérive(s) qu'elle veut dénoncer : ce faisant, elle ne fait que retrouver le Même, l'objet de sa dénonciation, dont elle donne à voir les effets délétères, mais à travers un miroir grossissant.

[109] Il s'agit d'un rapport commandité par le SPP Politique Scientifique à l'Institut pour un développement durable : P.-M. BOULANGER et Th. BRÉCHET, *Modélisation et aide à la décision pour un développement durable*.

[110] *Ibid.*, pp. 14-17.

Or les outils classiques d'aide à la décision se concentrent sur la fin du processus, le moment axiologique : ils valorisent, agrègent et priorisent les effets des politiques en fonction des préférences des décideurs. Il en va ainsi des analyses coûts-bénéfices, ou des analyses coûts-efficacité. Se contenter de ces outils, ce serait méconnaître une étape décisive de la prise de décision : la recherche d'action possibles et l'estimation de leurs effets. Afin d'assister le décideur dans cette recherche, les modèles cognitifs sont là pour modéliser les alternatives.

#### *3.4.2.2. Critères pour une modélisation cognitive*

Le modèle, vu sous son angle cognitif, ne dit rien de l'axiologie (ou, dans notre vocabulaire, de la téléologie) : elle étudie les effets des décisions sur le milieu étudié. Elle éclaire la prise de décision, mais ne la remplace pas[111]. Cela ne veut *pas* dire, évidemment, que le modèle cognitif soit *éthiquement neutre* : les *types d'effets* auxquels le modélisateur s'intéresse, en effet, influenceront pour une bonne part les conclusions qu'il sera possible de déduire du modèle utilisé. Voilà pourquoi d'ailleurs les auteurs de l'étude appellent de leurs vœux un métamodèle, qui modéliserait tous les aspects pertinents dans une analyse de durabilité.

En attendant un tel cadre rigoureux et exhaustif, ils distinguent cinq critères qu'un modèle cognitif devrait pouvoir prendre en compte[112], que sont l'interdisciplinarité, la prise en compte de l'incertitude, la prise en compte du long terme et de l'équité intergénérationnelle, l'interaction ou l'interdépendance entre niveaux de pouvoir et, finalement, la participation des parties prenantes. Nous nous écarterons, dans les lignes qui suivent, de la présentation que font les auteurs de leurs critères en ceci qu'il nous semble que les critères proposés se dédoublent à chaque fois en *deux exigences distinctes* : une *exigence d'expressivité*, d'une part, une *exigence procédurale*, d'autre part. L'exigence d'expressivité se manifeste dans des critères qui imposent au modèle la prise en compte de certaines entrées ou de certaines sorties : il importe alors que le modèle puisse prendre en compte certaines données (entrées) – ou qu'il puisse représenter certains effets pertinents à la prise de décision ou d'explorer des leviers d'action (sorties). En revanche, les critères procéduraux s'intéressent au *comment* de la modélisation ; ils imposent des exigences quant à la bonne tenue, à la manière de procéder, de la modélisation.

Le premier critère, l'interdisciplinarité, traduit l'exigence de se mettre à l'écoute de la complexité du réel. À ce titre, il se comprend de deux façons : premièrement, il peut être vu comme une exigence de pluridisciplinarité, en ce que la construction, l'utilisation et l'interprétation du modèle exigent un véritable dialogue interdisciplinaire. Ainsi compris, le critère est clairement de nature procédurale. L'autre sens du critère renvoie à une exigence d'expressivité : il faut que le modèle incorpore des connaissances venant d'horizons disciplinaires variés. En outre, il doit y avoir des interactions entre les variables d'état relevant de disciplines différentes ; il faut, en outre, que ces interactions ne soient pas « à sens unique », ce qui revient donc à exiger la présence de « boucles rétroactives » – chères à la pensée systémique – entre variables.

---

[111] *Ibid.*, pp. 143-146.
[112] *Ibid.*, pp. 25-33.

De même, nous retrouvons, sous la bannière du deuxième critère – la prise en compte de l'incertitude – deux exigences jumelées : les auteurs distinguent en effet entre deux sortes d'incertitude à prendre en compte : tout d'abord, l'incertitude épistémique : l'état de nos connaissances est tel que nous ignorons la valeur de certains paramètres quantifiables, que nous ignorons certaines relations entre variables, que nous pouvons même ignorer l'existence de certaines variables pourtant cruciales à la compréhension du phénomène étudié (incertitude sur la complétude). Gérer une telle incertitude est une exigence d'ordre procédural : par exemple, sachant que la SBA est fort sensible au calibrage des données initiales, l'exigence veut que le modèle augmente sa fiabilité en procédant par échantillonnage. Le deuxième type d'incertitude dont traitent nos auteurs est ontologique : l'incertitude est alors irréductible, inhérente au réel même ; prendre en compte un tel type d'incertitude relève de l'expressivité du modèle, qui doit alors permettre des variations stochastiques[113].

Le troisième critère est la prise en compte du long terme. Les auteurs ont d'abord à l'esprit un critère d'expressivité : le modèle doit pouvoir représenter les effets de nos politiques sur les générations à venir. La prise en compte du long terme s'entend alors comme un critère d'équité intergénérationnelle[114]. Les auteurs insistent d'ailleurs que ce critère ne doit pas être confondu avec une capacité de « prédire » l'avenir : si certains modèles projettent la démographie d'une ville sur une durée allant jusqu'à 250 ans (!), leur prétention n'est évidemment pas de se prononcer sur l'état du monde au 24e siècle ; leur ambition est plutôt de vérifier *la cohérence dans le temps* d'une dynamique, ce qui n'est, somme toute, rien d'autre que la définition même d'un développement durable. Plutôt qu'une prédiction, il s'agit d'une *anticipation*, fonction que nous avons déjà rencontrée dans le cadre de la science-fiction (§ 3.4.1). L'anticipation doit être entendue comme un *exercice*, à la manière des sportifs qui s'entraînent dans une salle de sport : ils savent fort bien qu'ils n'auront jamais à poser ces gestes répétitifs dans un moment plus ou moins lointain de leur existence ; en revanche, ils estiment qu'en s'exerçant de la sorte, ils se donnent la capacité corporelle et émotionnelle de faire face aux épreuves que la vie leur prépare[115]. Sur son versant procédural, le critère exige que les modèles soient explicites quant à leur cadre de validité temporel. Exigence de simple hygiène académique peut-être, mais il s'avère que les modèles n'indiquent que rarement leur domaine de validité temporel, alors que celui-ci peut parfois être fort bref, au point de ne pas dépasser l'année en cours.

Quatrième critère, la « glocalité » ou l'interdépendance entre niveaux de pouvoirs : plus la problématique est globale, plus il y a interactions entre différents niveaux décisionnels. L'effet global visé n'est parfois rendu possible que par une multitude d'actions locales ; à l'inverse, une décision prise en un lieu peut avoir des répercussions fâcheuses en un tout autre lieu. Nos auteurs demandent, avant tout, que les modèles puissent rendre compte de ces interactions croisées. Il faut cependant aussi relever une autre exigence qui renvoie à ce critère, à savoir celle de pouvoir représenter les

[113] Ajoutons à ceci qu'il existe un troisième type d'incertitude, l'incertitude morale, qui n'est cependant pas d'ordre cognitif et dons nous reparlerons plus loin, vu qu'il met en exergue le problème de l'intégration axiologique.

[114] Très beau critère au demeurant, car en lui se rencontrent deux conceptions du temps : le temps qui passe ; le temps en tant que triple présent : les agents naissent, prennent de l'âge, cèdent la place à une nouvelle génération pour qui l'expérience de la vie est à refaire.

[115] L'analogie paraît sous la plume de D. H. GUSTON, *"Daddy, Can I Have a Puddle Gator?": Creativity, Anticipation, and Responsible Innovation*, dans R. OWEN, J. BESSANT et M. HEINTZ, *Responsible Innovation*, p. 111.

décideurs comme *internes* au modèle, comme des entités qui peuvent être affectés par lui, exigence qui, aux dires des auteurs[116], n'est que rarement satisfaite dans les modèles actuellement utilisés. Notons que les deux exigences portent sur les contenus attendus du modèle : ce critère est donc le seul à être exclusivement de nature expressive ; il recoupe d'ailleurs en bonne partie la discussion entre niveaux micro et macro (que nous avons déjà présentée au deuxième chapitre, § 2.3.2).

Cinquième et dernier critère, la prise en compte des parties prenantes[117]. Faire appel à ces dernières peut avoir des raisons diverses. En premier lieu, l'implication des parties prenantes peut se faire au nom d'une certaine idée de la démocratie ou de l'émancipation (raison *normative*) ; elle peut se faire pour des raisons stratégiques, pour faciliter l'adoption d'une décision (raison *instrumentale*). Il convient de distinguer soigneusement la raison instrumentale d'une tentative de manipulation : dans le cas d'une participation instrumentale, le pouvoir public part du principe que le public est déjà d'accord avec le principe (par exemple, le nécessaire effort contre le réchauffement climatique) ; que la négociation sur l'objectif est déjà close ; les pouvoirs publics endossent alors un rôle de plaideur d'une cause entendue (*issue advocate*). Une dernière raison possible pour faire appel aux parties prenantes est dite *substantive* : il s'agit alors de mobiliser des connaissances sociales ou locales que seules possèdent les parties prenantes et qui sont décisives pour la réussite des objectifs du projet. Que la raison soit normative ou stratégique, elle relève dès lors d'une exigence procédurale adressée au modèle. L'exigence d'expressivité – moins décisive pour ce critère – se trouve dans la possibilité pour le modèle de prendre en compte les connaissances locales (en guise d'entrées) ; de fournir une visualisation parlante des résultats (en guise de sortie).

Les auteurs insistent d'ailleurs sur l'importance de la visualisation. Elle a partie liée avec la question de l'intégration cognitive, dont les auteurs font grand cas[118] sans toutefois vraiment la définir. Elle semble devoir être comprise ici comme un principe unificateur des entrées et sorties : si toutes les données d'entrée sont localisables dans l'espace, les effets doivent l'être aussi. Si les entrées sont exprimables en termes de probabilité, les effets doivent l'être aussi : c'est la notion de risque. Il s'agit, en somme, d'un *langage commun* – d'où la remarque des auteurs que l'intégration joue un rôle essentiel pour les critères d'interdisciplinarité et de la participation des parties prenantes.

À ce titre, l'intégration cognitive est l'envers et le pendant de la spécialisation et de l'abstraction : elle est le nom de l'exigence de synthèse et de retour au réel sur laquelle fonder un cadre général – une métathéorie – d'analyse des incidences de durabilité (*sustainability impact assessment*). Concrètement, l'intégration peut être fondée sur un principe unificateur théorique, qui est alors probabiliste ou systémique. Le principe d'intégration peut aussi être pragmatique – les auteurs pensent notamment au rôle joué par l'espace dans ce contexte : du fait de son haut degré de réalisme, il favorise fortement l'appropriation du modèle par les parties prenantes.

[116] P.-M. BOULANGER et Th. BRÉCHET, *op. cit.*, p. 121.
[117] Les auteurs sont plutôt laconiques dans leur traitement de ce critère ; nous l'amplifions un peu en empruntant des matériaux à la discussion consacrée au dialogue public par K. SYKES et Ph. MACNAGHTEN, *Responsible Innovation – Opening Up Dialogue and Debate*, dans R. OWEN, J. BESSANT et M. HEINTZ, *Responsible Innovation*, pp. 94-97.
[118] P.-M. BOULANGER et Th. BRÉCHET, o*p. cit.*, pp. 12-13.

### *3.4.2.3. Critères pour une modélisation axiologique*

Alors que le mouvement cognitif du processus décisionnel se veut résolument ouvert, l'étape suivante en est une de conclusion ; l'étape cognitive explore le système sous tous les angles, toutes les coutures, tente d'aller aussi loin que possible dans la découverte d'actions et l'estimation de leurs effets. L'analyse axiologique, elle, va valoriser et prioritiser ces effets en fonction des préférences – des valeurs – des décideurs. La modélisation axiologique est une problématique que le rapport sur lequel nous nous basons ici ne fait que mentionner ; nous ferons de même, nous contentant de quelques remarques.

Tout d'abord, la valorisation des effets appelle à son tour une intégration, non plus cognitive cette fois mais axiologique. Un « classement » des actions, en effet, ne peut se faire sans un principe d'ordre. Dans le cas du développement durable, où le décideur est assumé être une autorité publique, le principe d'intégration semble le plus souvent de nature budgétaire. Insistons : le budget n'est pas ici conçu comme une valeur en soi mais un principe intégrateur : il nous renseigne sur l'attribution des *moyens* : quels moyens pouvons-nous libérer pour quelles fins ?

Ensuite, l'axiologie telle que nos auteurs la conçoivent a une dimension téléologique prononcée : il s'agit de formuler une certaine vision de « la vie bonne » en société. Plus précisément, une analyse axiologique se construit comme un triangle dont les angles sont les moyens, les fins, les valeurs qui inspirent les fins[119]. Chaque aspect peut ici faire l'objet d'un examen approfondi : les moyens mis en œuvre ont-ils une probabilité élevée de réaliser les fins, de façon proportionnelle ? Peuvent-ils provoquer des effets secondaires ? Sont-ils compatibles avec les valeurs ? Les valeurs elles-mêmes n'échappent pas à l'examen, dans la mesure où elles sont le lieu d'un type d'incertitude particulier.

Car, enfin – dernière remarque – nous avons vu, au paragraphe précédent, deux types d'incertitude : épistémique d'une part ; ontologique d'autre part. Or il est tout à fait possible de discerner un troisième type d'incertitude, une incertitude proprement morale : même si tous les faits pertinents sont connus, il est envisageable que nous ne sachions pas que choisir parmi les actions possibles : une telle situation peut se présenter lorsque les alternatives, tout en s'excluant mutuellement, sont toutes pareillement plausibles en tant que valeur, comprise ici comme *motivation*[120]. Une telle incertitude, cependant, n'est plus cognitive mais proprement axiologique.

### *3.4.2.4. Classes d'outils de modélisation cognitive*

Apprécier les atouts – ainsi que les faiblesses – de la SBA ne prend sens qu'en la comparant à d'autres techniques de simulation. Certes, les techniques de modélisation sont le plus souvent conçues avec des objectifs précis, parfois très éloignés des préoccupations du développement durable. Il faut en

---

[119] Nos auteurs passant très brièvement sur les principes de l'analyse axiologique, nous nous inspirons ici de M. WEYEMBERGH, *Max Weber et Ulrick Beck : de la première à la deuxième modernité ?*, dans C. KERMISCH et G. HOTTOIS, *Techniques et philosophies des risques*.

[120] Cf. V. BHARGAVA et T. WAN KIM, *Autonomous Vehicles and Moral Uncertainty*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, p. 6.

outre garder à l'esprit que l'exercice qu'est la comparaison entre différentes techniques de modélisation est un problème essentiellement ouvert : ni méthodologie universellement reconnue, ni pierre de touche philosophale, ni même étalon de mesure éprouvé ne sont disponibles pour une telle entreprise.

Une des thématiques phare du développement durable est l'aménagement du territoire, étant donné que s'y côtoient beaucoup d'enjeux : l'énergie, l'eau, la biodiversité, la mobilité, la sécurité, la santé et la pauvreté, pour ne nommer que les principales préoccupations. Or il se fait que ce n'est pas le domaine le mieux loti en matière de simulation[121] . L'outil de référence y est TRANUS, qui se fonde sur une analyse en « secteurs » (il faut entendre par là des types de ressources) sur laquelle il greffe différentes mesures économétriques. Un secteur peut être local ou transportable ; lorsqu'il est transportable, le coût de transport est calculé à l'aide d'un graphe de transport – c'est la seule concession que fait l'outil à la prise en compte de l'espace, pourtant un aspect capital en développement durable. À aucun moment, TRANUS ne sort de l'économique, que ce soit dans son langage, dans le dialogue entre spécialistes ou dans la rétroaction entre les variables. Le comportement économique maximisateur a seul droit de cité.

Par voie de conséquence, afin de gérer durablement l'aménagement du territoire, il faut se mettre à la recherche de techniques alternatives. Parmi celles-ci, trois types de modélisation sont jugés particulièrement intéressants pour la prise en décision en matière de développement durable : il s'agit des réseaux bayésiens, des modèles de dynamique des systèmes, ainsi que la simulation à base d'agents. Regardons chacun de ces modèles de plus près.

Sur le fond, un réseau bayésien n'est autre qu'un graphe acyclique orienté et fortement connecté : la sémantique des arcs repose sur celle de la « causalité » dans la terminologie bayésienne[122]. Les nœuds y représentent des probabilités, connues ou déduites par application du théorème de Bayes. Un modélisateur, dès lors, traduit tous les phénomènes interdépendants (ou supposés tels) dans un tel graphe et pondère les nœuds grâce aux informations dont il dispose. Une fois que cet exercice est fait, l'outil se chargera de calculer les probabilités des nœuds restés vierges. L'application principale des réseaux bayésiens est l'évaluation des risques. L'outil, en effet, excelle dans la gestion de l'incertitude.

Les modèles de dynamique des systèmes ont historiquement partie liée à la naissance du développement durable. De tels modèles représentent la réalité à modéliser comme un ensemble de sous-systèmes, chacun doté de stocks et de flux. Alors que les taux de flux décrivent les entrées et les sorties, les niveaux de stock fluctuent au gré d'un jeu d'intégrales. La dynamique observable dans un tel modèle est l'interaction entre les différents stocks : ils peuvent se renforcer ou s'amortir mutuellement. Un tel modèle est de prime abord assez rigide : sa structure est invariable ; son comportement entièrement déterministe. Cependant il excelle dans des situations où il s'agit de comprendre des interactions complexes entre systèmes, ou lorsqu'il s'agit de simuler les effets

[121] En Belgique notamment, le recours à la simulation en matière d'aménagement du territoire serait, selon les dires des auteurs, inexistant (P.-M. BOULANGER et Th. BRÉCHET, *op. cit.*, p. 99).
[122] Comme la théorie bayésienne est de nature strictement statistique, le terme de « corrélation » aurait sans doute été plus adéquat.

induits par des *délais* d'information ou de réaction. Ce type de modèles est, en d'autres termes, très approprié pour simuler des phénomènes sur le long terme.

C'est le cas de l'outil de simulation *Urban Dynamics*[123], où les villes sont considérées comme un ensemble de trois sous-systèmes en interaction : la population, les activités économiques et l'espace bâti. Des secteurs en croissance et des habitations de qualité attirent une population qualifiée ; au contraire, des logements insalubres et une activité économique en berne la repoussent. Ce qui rend ce modèle intéressant sur le plan du développement durable, c'est sa notion de « situation normale », c'est-à-dire la situation où les interactions entre les différentes variables s'équilibrent. Or une telle situation normale peut correspondre à un état qualifié de durable– il est vrai, de façon tout à fait *a priori* – et ainsi examiner tous les écarts par rapport à l'état de référence. Cet avantage prend tout son sens sur le long terme : *Urban Dynamics* peut explorer jusqu'à 250 ans d'évolution ! Même si un tel exercice n'a pas vocation à être utilisé comme pronostic, il permet de vérifier si un jeu d'interactions données peut prétendre à la stabilité, à un comportement cohérent et compréhensible.

Toutefois, ni la dynamique des systèmes, ni les réseaux bayésiens ne prennent en compte l'espace. Or la SBA, elle, excelle justement dans ce domaine. Elle a, par voie de conséquence, de sérieux atouts en matière d'aménagement du territoire[124]. L'exemple cité par les auteurs, PolSim (pour *Policy Simulator*), est une composante du logiciel commercial CommunityViz[125], lui-même une extension de la plate-forme ArcGIS. L'outil permet de suivre l'évolution de résidents et des entreprises sur un plan d'affectation du sol, en prenant en compte un nombre important de données : caractéristiques et fonctionnalités des immeubles, besoin en main d'œuvre et en ressources des entreprises, besoin en services et biens des personnes, etc. Les mouvements migratoires sont en outre influencés par des évènements qui influencent la composition des ménages : naissances, mariages, décès, etc. Chaque personne est située dans un réseau de relations familiales, professionnelles et de voisinage. L'outil se montre très intéressant sous deux angles. Premièrement, la prise en compte de l'espace est excellente, dans la mesure où il s'agit d'une extension de l'application SIG la plus renommée sur le marché. Autre atout de taille : avec PolSim, la modélisation s'émancipe pleinement du « tout-économique ». Ainsi l'évaluation de la qualité de vie des ménages s'y fait à deux niveaux : bien sûr, l'évolution de la situation économique des ménages est prise en compte. Cependant, cette évaluation est complétée par l'adéquation entre le style de vie et les *valeurs du ménage* : ainsi, une dimension subjective, liée aux préférences des acteurs, s'introduit dans les équations.

Dans les paragraphes qui suivent, nous mettrons les qualités de la SBA en relief à travers des exemples plus fournis. Le premier exemple mettra en avant les qualités réalistes de la SBA, notamment sa capacité à intégrer des apports interdisciplinaires divers et variés ; le deuxième s'intéressera davantage à sa faculté d'inclure les parties prenantes dans la prise de décision.

---

[123] *Ibid.*, pp. 124-129.

[124] Ce n'est pas le lieu ici de développer la comparaison chiffrée à laquelle procèdent nos auteurs ; contentons-nous de signaler que la SBA se retrouve en une première position tout à fait confortable dans le classement des différents modèles ; le lecteur intéressé peut se référer à l'ouvrage cité, pp. 87-90.

[125] Voir *op. cit.*, pp. 135-136. Note historique : PolSim a été retiré de CommunityViz en 2003, c'est-à-dire l'année même de la publication du rapport (http://www.georgejanes.com/PDF/History_of_CommunityViz.pdf).

### *3.4.2.5. La SBA contre la pandémie du Covid-19*

L'actualité belge et internationale a, pour ainsi dire, rattrapé ce mémoire : en fin d'année 2019, le coronavirus SARS-coV-2 fait son apparition dans la province chinoise de Hubei. Elle atteindra rapidement l'Europe, prendra l'ampleur que nous savons, donnant naissance à la pandémie Covid-19. Les politiques mises en œuvre pour combattre la pandémie entravent de façon parfois importante les libertés individuelles et collectives : confinement, couvre-feu, restrictions importantes de déplacement… même les domiciles privés, pourtant réputés inviolables par ailleurs, se voient imposer des contraintes parfois draconiennes sur leur organisation interne.

Notre propos, ici, ne sera pas de nous faire le champion des libertés, ni de surenchérir sur l'insécurité juridique ou l'esprit technocrate dont bon nombre des mesures se voient parfois accusés. Nous ne contesterons pas le principe selon lequel des populations sont prêtes à troquer une part de leurs libertés en vue d'un retour à la normale en matière sanitaire. Nous nous poserons une seule question : de quelle *justification rationnelle* peuvent se prévaloir les mesures prises ? Sont-elles seulement passibles d'une telle justification ? Précisons : lorsque nous demandons une justification, nous demandons des éléments permettant de dire que les mesures sont adéquates par rapport au mal combattu. C'est le principe de la *proportionnalité*, d'autant plus pertinent ici que selon la durée, l'extension etc. du confinement, les résultats peuvent prendre une tout autre forme[126].

Le thème du mémoire oblige : nous nous intéresserons à une simulation à base d'agents, COMOKIT, créée sous l'impulsion de l'Institut de Recherche pour le Développement (IRD)[127] afin de fournir un outil d'aide à la décision sur une échelle résolument locale. La simulation doit permettre d'interroger les effets de politiques inscrites dans des situations concrètes : le confinement de tel quartier est-il plus efficace que celui d'un village entier ? La fermeture des écoles aplanit-elle les pics de transmission ? Comment le port du masque influe-t-il sur la dynamique de l'épidémie ? Quel pourcentage de la population peut être autorisé à vaquer à ses occupations pendant un confinement ? Etc.

---

[126] Le bourgmestre d'Ixelles, Christos DOULKERIDIS, dans une tribune parue dans *La Libre Belgique* du 21 février 2021, dit bien le fond du problème : « […] nous ne pouvons pas ériger en système les règles actuelles. Tant sur le plan humain que sur le plan démocratique. Sur le plan humain, ni les jeunes, ni les familles, ni les personnes isolées, ni les personnes âgées, ni les femmes et les hommes forcé(e)s de ne pas travailler, de ne pas se retrouver, de ne pas voyager, de ne pas fréquenter leur école ne pourront supporter plus longtemps sans réagir. Et nous ne pouvons pas non plus faire de nos forces de police des femmes et des hommes qui vont finir par se faire haïr par la majorité de notre société si on continue à leur demander de faire respecter et de sanctionner des comportements qui en définitive sont parmi les plus humains. Sur le plan démocratique, la fragilité du régime d'exception dans lequel nous nous sommes installés a été rappelée à juste titre par de nombreux acteurs et actrices de notre société. » Dans un autre contexte, celui des drones tueurs, le principe de proportionnalité a été fortement mis en relief par R. ARKIN, *Governing Lethal Behavior in Autonomous Robots*, pp. 47, 185 et suivantes.

[127] Notre présentation de COMOKIT se base sur les informations glanées sur le site web du projet, https://comokit.org/, ainsi que sur les articles de B. GAUDOU ET A., *COMOKIT: A Modeling Kit to Understand, Analyze, and Compare the Impacts of Mitigation Policies Against the COVID-19 Epidemic at the Scale of a City* ; A. DROGOUL ET A., *Designing social simulation to (seriously) support decision-making* ; ainsi que sur le document de la main d'A. BRUGIÈRE ET A., *O.D.D. description of the COMOKIT model*.

Le modèle connaît trois types d'agents : des bâtiments, des individus, ainsi que l'autorité, un type spécial d'agent en charge de prendre des mesures afin de contenir la propagation du virus. Les bâtiments, quant à eux, sont nécessaires pour modéliser des phénomènes tels que la contamination environnementale – nous revenons ci-dessous sur la notion – les loisirs ou la capacité hospitalière.

Au total, COMOKIT cherche à intégrer non moins que cinq sous-modèles différents : le premier sous-modèle est celui de la dynamique clinique individuelle et du statut épidémiologique des agents. Les individus peuvent ainsi passer par divers états épidémiologiques : réceptif, latent, asymptomatique, présymptomatique, symptomatique, immunisé (ou mort). Les trois états entre la latence et l'immunisation sont infectieux. Une fois infecté, l'individu peut passer parmi divers états cliniques : besoin d'hospitalisation, besoin de soins intensifs, remis (ou, malheureusement encore, mort). Ce sous-modèle – inspiré de SEIR[128], le modèle épidémiologique de référence – requiert plusieurs paramètres : le temps de latence, la période asymptomatique et symptomatique, le taux de transmission en cas d'infection symptomatique et asymptomatique, la probabilité d'être asymptomatique, la probabilité de mourir, etc.

Le deuxième sous-modèle représente la transmission directe de l'infection d'agent à agent. Ce sous-modèle requiert des données abondantes de nature sociale afin de représenter adéquatement la vie en collectivité, au sein de laquelle le taux de transmission est plus élevé : ainsi dans un ménage, le risque de transmission est plus élevé, même comparé aux autres habitants du même immeuble. Les informations requises ici ont trait à la composition des ménages, à l'emploi, l'âge, etc.

Le troisième sous-modèle s'intéresse toujours de près aux agents individuels, notamment à leurs activités quotidiennes, que celles-ci soient domestiques, professionnelles et scolaires, ou encore liées aux loisirs. Cette modélisation prend la forme d'un emploi du temps dont chaque individu est doté et qui se présente comme un agenda d'activités quotidiennes : aller travailler, aller à l'école, faire des courses, etc. Ces activités peuvent être partagées entre plusieurs individus, par exemple aller au restaurant entre amis ; elles se déroulent toujours dans un bâtiment. La prise en compte des activités requiert une granularité temporelle très fine. COMOKIT se distingue ainsi par des « pas de temps » d'à peine une heure, alors que d'autres modèles se contentent généralement d'un demi-jour, voire d'un jour entier. De tels pas de temps, lorsqu'ils sont appliqués aux comportements individuels, impliquent déjà par eux-mêmes un effet agrégatif, que l'emploi d'une SBA est justement censé prévenir.

Le quatrième sous-modèle a trait à la transmission environnementale. En effet, la contamination ne doit pas nécessairement être directe, causée par une interaction entre agents ; elle peut aussi être indirecte lorsqu'elle est induite par la rémanence d'une charge virale dans un milieu bâti. Les foyers de contamination les plus importants dans ce cadre sont le domicile (le ménage), l'école, le lieu de travail et les amis (ou le réseau social). Une telle étude, qui se centre sur le va-et-vient des individus dans des bâtiments, rend nécessaire la prise en compte d'une échelle spatiale relativement petite.

[128] Relevant des modèles compartimentaux, abréviation de *Susceptible, Exposed, Infectious, Recovered*. Ces modèles se distinguent entre eux par le nombre d'états d'un cycle individuel : SIS, SIR, SIRD, etc. (cf. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology).

Elle a besoin de plusieurs *shapefiles*[129] qui renseignent sur l'espace à modéliser et sur l'emplacement des bâtiments.

Le cinquième et dernier sous-modèle a pour objet l'action publique (*policy design and implementation*). L'agent représentant l'autorité a ainsi à sa disposition un ensemble de mesures qu'il peut appliquer : il peut décréter un confinement, la fermeture des écoles ou des lieux de travail, imposer aux personnes infectées de rester chez elles, etc. De telles mesures peuvent être limitées à une zone particulière (*SpatialPolicy*) ou à un pourcentage de la population (en vertu d'une restriction partielle – *PartialPolicy*) ou encore pendant une période déterminée (*TemporalPolicy*). Chaque agent, avant d'entamer une activité, demande la permission à l'autorité. Celle-ci évalue l'ensemble des mesures actives avant de marquer son accord. Même si l'individu n'obéit pas toujours (l'obéissance peut être paramétrée via un taux de conformité), le modèle fait donc l'hypothèse d'une connaissance parfaite des mesures applicables.

L'autorité ne prend bien sûr pas ces mesures à l'aveugle : à chaque pas de temps, il peut tester un certain nombre de personnes afin de se faire une idée de l'état épidémiologique de la population. Il faut souligner à ce propos que l'autorité ne peut pas tester toute la population à chaque pas de temps, pour des raisons évidentes de moyens. En outre, les résultats des tests présentent une marge plus ou moins importante d'erreur. Il s'ensuit que l'autorité n'a qu'une *vue partielle* sur le nombre d'infections de la population ; ce qui reflète aux yeux des auteurs assez fidèlement la réalité, dans la mesure où la plupart des individus infectés sont asymptomatiques et passent donc inaperçus. Pour notre discussion, il est important de faire remarquer que les pouvoirs publics sont ici partie intégrante du modèle : tant leur perception que leur action sont modélisées, à la manière de tout autre agent. La chose est assez rare pour être mise en relief.

L'ensemble des sous-modèles ainsi mis en œuvre ont besoin d'une grande quantité de données de nature diverse : spatiales, démographiques et épidémiologiques. Or dans beaucoup de cas, celles-ci ne sont pas disponibles, soit qu'elles ne relèvent pas du domaine public, soit qu'elles sont tout simplement inconnues. La gestion de l'incertitude est ici donc une exigence décisive ; c'est pourquoi COMOKIT permet la génération, plus ou moins aléatoire, dans des plages de valeurs estimées plus ou moins plausibles, des paramètres requis. Ce procédé comble les lacunes dans les données, mais requiert à l'évidence une certaine prudence : notamment, les simulations doivent être exécutées en série en faisant varier l'espace paramétrique afin d'exclure d'éventuels biais.

Telles sont les données d'entrée. Qu'en est-il maintenant des sorties ? En d'autres termes, à quel type de questions le modèle cherche-t-il à répondre ? Rappelons-nous l'enjeu : l'aide à la décision pour enrayer la pandémie. Assez logiquement, le modèle cherche à étudier les effets de l'action publique sur l'évolution spatio-temporelle des infections. Pour ce faire, différents scénarios sont simulés et comparés entre eux : tout d'abord le scénario de référence (*baseline*), celui où aucune mesure n'est prise. Puis, un scénario où le port du masque est rendu obligatoire ou encore un scénario de confinement total. Enfin, dans un but de tester le bien-fondé des hypothèses du modèle, trois scénarios reflètent des politiques effectivement mises en œuvre, à savoir en Corée du Sud, en

[129] Les *shapefiles* sont un format de données géographiques propriétaires d'Esri, la firme de référence en matière de SIG en vertu de sa suite logicielle ArcGIS (cf. https://www.arcgis.com/index.html).

France et à Malte : il s'agit là de trois cas bien documentés et dont les différentes politiques ont abouti à des résultats différenciés. La Corée du Sud fait figure de premier de la classe, car sa politique de tests de dépistage à grande échelle avec mise en quarantaine des ménages positifs a réussi à stopper net la diffusion de la pandémie sur le territoire national. Les effets observés dans les deux autres cas sont plus mitigés : les restrictions importantes des déplacements en France n'ont fait qu'aplanir la courbe de diffusion ; à Malte, où les personnes à risque ont été confinés d'office, l'effet a été une diminution certaine du nombre de morts.

Le modèle parvient à reproduire ces observations avec un nombre limité de répétitions : une cinquantaine d'exécutions suffit pour que la moyenne s'accorde avec les données empiriques. La taille relativement modeste de l'échantillonnage requis ne doit cependant pas nous faire oublier que la simulation d'un tel modèle riche en données monte difficilement en échelle[130]. Une autre faiblesse signalée par les auteurs tient aux lacunes dans les « contenus » de la modélisation : ainsi les activités informelles – sans agenda – sont passées sous silence. Des rassemblements au grand air, en dehors des bâtiments, ne sont pas au rendez-vous non plus. L'absence la plus remarquable est toutefois celle des transports en commun : le modèle part en effet du principe que les contaminations durant les transports sont négligeables. Cette omission est justifiée par le fait que les transports en commun ne font sentir leurs effets qu'à l'échelle nationale et planétaire ou dans les grandes agglomérations urbaines, non pas dans les petites à moyennes municipalités et qui sont la cible principale de COMOKIT[131].

Nous pourrions multiplier les détails, les complications, les subtilités ; l'essentiel, cependant, est de faire ressortir le ferme parti pris des concepteurs du modèle en faveur du réel, au point de rapprocher le modèle d'un système d'information géographique : le rendu visuel de la diffusion spatiale de la maladie le dispute en acuité à cette classe d'outils purement descriptifs que sont les SIG. Dans la discipline scientifique à laquelle les auteurs cherchent à rendre justice – que nous avons vu relever des « sciences du terrain » pour parler comme Stengers – l'abstraction, si prisée qu'elle soit dans d'autres champs du savoir académique,– n'est pas une vertu, mais bien au contraire une tare, un manque flagrant de sérieux, dans la mesure où justement elle est éloignement du réel :

> *What we have witnessed, instead, these last 4 months, is an explosion of agent-based toy models representing,* ad nauseam*, the spread of the virus or similar dynamics within artificial populations without space, without behaviours, without friend nor family relations, without social networks, without even remotely realistic activities or mobility schemes; in short, populations of artificial agents devoid of everything that makes a human population slightly different from a mixture of homogeneous particles. How we, as a community, can claim to*

[130] Certes, une exécution unique est tout à fait abordable : pour nous permettre de nous faire une idée de l'ordre des grandeurs, les auteurs affirment qu'il faut 15 minutes et 800 Mo de mémoire vive sur un ordinateur portable monocœur pour simuler 6 mois d'une bourgade de dix mille habitants, y compris la vue 3D et tous les graphiques.

[131] En effet, plus l'échelle d'intervention est réduite, moins les modèles habituels en épidémiologie (modèles agents sur-simplifiés, modèles épidémiologiques classiques, agrégés sur des compartiments de la population, par exemple SEIR) donnent des résultats fiables.

*inform policy makers, in such a critical context, with such abstract and simplistic constructions is difficult to justify.*[132]

Les simulations dites « simples », même à base d'agents, sont toujours tentées par l'abstraction. Évidemment, en COMOKIT également, les agents sont – selon les dires des auteurs mêmes – des représentations simplifiées à l'extrême de leurs homologues empiriques : des agents, ne sont retenus que les attributs qui aident à les comprendre comme des vecteurs de la maladie. Mais la simplification est ici pilotée par les questions auxquelles le modèle cherche à répondre, et non l'inverse. Ces vertus du modèle sont la raison pour laquelle il peut être utilisé dans un contexte géographique et temporel concret, pour faire des pronostics précis même à court terme : la SBA n'est donc pas limitée à se borner à décrire des dynamiques cognitives générales sur une échelle temporelle indéterminée.

Pour réaliser un tel objectif, il faut cependant des données, tant quantitatives que qualitatives, d'où l'importance du libre accès aux données disponibles pour ce type d'exercice. Pour la prise de décision, la donnée qualitative est souvent de première importance : il s'agit donc de modéliser adéquatement les *comportements* des agents, sur la base de descriptions et classifications sociologiques précises. Ces données qualitatives ne se présentent certes pas, à la manière des données quantitatives, comme un ensemble de paramètres. La donnée qualitative est le cœur même de la simulation, le code écrit par le programmeur-modélisateur pour « donner vie » aux agents.

Finalement, malgré l'ambition réaliste du modèle, il a – comme tout modèle – ses limites : de par ses hypothèses, elle ne saurait nous renseigner sur le bien-fondé de certaines catégories de décisions. Ainsi l'abstraction faite des déplacements et des transports en commun interdit le modèle de se prononcer sur les effets du port généralisé du masque, comme nous l'avons vu dans la Région bruxelloise. L'hypothèse de la connaissance parfaite des mesures interdit de mesurer les effets de la cacophonie entre niveaux de pouvoirs dont la Belgique a fourni un exemple particulièrement vaudevillesque. L'absence d'activités hors milieu bâti bride encore l'interrogation du bien-fondé de l'interdiction des manifestations sportives et culturelles en plein air. Le modèle a choisi quelles questions poser à la réalité ; il importe que les interprétations que le modélisateur en fait, n'outrepassent pas ce cadre.

#### *3.4.2.6. La SBA à l'œuvre dans la modélisation d'accompagnement*

Le modèle COMOKIT que nous venons de voir cherche à produire des connaissances précises, un pronostic pour aider à la décision. Or ce n'est pas la seule façon dont la décision peut être facilitée. Nous en avons déjà touché un mot au deuxième chapitre (§ 2.3.1) lors de la discussion de SELFCORMAS : la SBA est une alliée de choix dans une démarche appelée la *modélisation d'accompagnement*[133]. Le

[132] A. DROGOUL ET A., *Designing social simulation to (seriously) support decision-making*.
[133] Notre source principale pour exposer la démarche est l'ouvrage collectif dirigé par M. ÉTIENNE, *La modélisation d'accompagnement. Une démarche participative en appui au développement durable*.

cas[134] que nous voulons présenter ici a pour lieu l'atoll de Tarawa, qui concentre à lui seul les deux tiers de la République des Kiribati, État archipel dans le Pacifique. L'îlot principal de l'atoll, Tarawa-Sud, connaît non seulement une expansion démographique induite par l'immigration, mais aussi un besoin en eau croissant par tête de population. Or sur l'atoll, l'eau potable provient non pas d'une nappe phréatique comme sur le continent européen, mais dépend de la présence de lentilles d'eau douce, c'est-à-dire une couche d'eau souterraine qui flotte au-dessus de l'eau de mer, plus lourde à cause de son taux salin plus élevé. De telles lentilles sont particulièrement sensibles à des fluctuations dans le climat, ainsi qu'à la pollution.

Afin de faire face au besoin d'eau de la population de Tarawa-Sud, l'État a instauré des réserves d'eau qui permettent d'alimenter le réseau de distribution. Comme les surfaces agraires sont rares, ces réserves d'eau sont mal vues par la population, qui dépend encore majoritairement d'une économie de la subsistance. Les réserves existantes sur les îlots de Bonriki et Buota sont donc une source continue de tension entre les insulaires et les autorités. Un projet financé par la Banque asiatique de développement devrait donner lieu à d'autres réserves sur l'îlot de Tarawa-Nord, risquant d'entraîner les mêmes conflits qu'à Bonriki et Buota.

Pour éviter ces conflits, les autorités firent appel à un processus de médiation connu sous le nom de *modélisation d'accompagnement* : dans un scénario de conflit de ressources, une telle démarche vise la médiation par la production d'une *représentation commune* des enjeux, représentation exprimée sous forme de modèles d'inspiration multi-agents. Dans une telle démarche, les chercheurs s'attachent d'abord à récolter les perceptions des acteurs locaux : il s'agit d'un travail d'acquisition relevant du génie cognitif (*knowledge engineering*).

Après un travail initial de documentation et d'enquête, les chercheurs établirent une liste de groupes sociaux avec leurs meneurs, capables d'exprimer la voix du groupe : furent ainsi identifiés des groupes religieux (catholiques, protestants), culturels (le comité des Anciens, l'association des femmes), administratifs et sportifs. Les interviews menées auprès de ces porte-parole eurent trois composantes : premièrement, une demande de commenter des photos renvoyant à diverses réalités de l'île, que celles-ci soient économiques ou environnementales ; deuxièmement, la demande de dessiner une représentation de l'île avec ses caractéristiques principales ; troisièmement, la personne interviewée fut invitée à relier entre elles des cartes mentionnant des éléments pertinents dans le traitement des eaux. Le but de ce dernier exercice était de se faire une idée des connaissances hydrologiques de la population. Toutes les interviews furent enregistrées. Ces enregistrements furent par la suite analysés par un logiciel d'analyse qualitative pour organiser leur lexique en réseaux associatifs. À partir de ces réseaux de termes, les chercheurs créèrent des ontologies qui étaient à la base du modèle final dans le formalisme UML. Le modèle UML à son tour permit de construire un modèle multi-agents[135].

---

[134] La présentation que nous en donnons ici se base sur deux articles d'A. DRAY et ses collègues : d'une part, *The AtollGame Experience: from Knowledge Engineering to a Computer-Assisted Role Playing Game* ; d'autre part, *Who wants to terminate the game?*

[135] Tous les cas font mention de l'étape indispensable que constitue le(s) diagramme(s) UML. Dans les travaux consultés, nous devons cependant regretter le manque de clarté concernant la manière exacte dont traduction du diagramme UML en une SBA se fait concrètement ; notre perplexité trouve sa source dans les exemples, qui ne donnent jamais à voir autre

Se pose dès lors la question de la validation d'un tel modèle. La voie royale, en modélisation d'accompagnement, pour faire valider de tels modèles, consiste à les faire *jouer*, en version simplifiée, sur des jeux de plateaux[136]. Dans le cas qui nous intéresse ici, jeu et simulation s'entre-mêlent, formant un produit hybride entre jeux qualifiés « à agents humains » et simulations à agents virtuels[137]. Ce jeu assisté par ordinateur consiste en deux plateaux, représentant chacun une île, l'une densément peuplée, à la manière du Tarawa du Sud, siège des institutions gouvernementales, et l'autre faiblement peuplée, dans laquelle les joueurs n'auront pas eu de peine à reconnaître Tarawa-Nord[138]. Chaque île accueille 8 joueurs. Le choix du même nombre de joueurs sur chaque île devait faciliter les interactions entre joueurs, mais le nombre de familles représentait plus fidèlement la réalité démographique des îlots composant le récif de l'atoll de Tarawa : ainsi l'île surpeuplée compte 200 familles et l'île agreste n'en compte que 50. Selon la coutume locale dite de la *famille étendue*, chaque famille terrienne héberge sur ses terres des familles de la même parentèle, pouvant aller jusqu'à 10 familles (qui comptent en moyenne 8 personnes).

Chaque joueur est un propriétaire terrien. Sur le plan professionnel, un joueur peut être fonctionnaire, pêcheur, ou sans emploi, en fonction de quoi il reçoit à chaque tour un certain nombre de jetons. Le ressort du jeu est la pression sur les ressources en terre et en eau exercée par le nombre sans cesse croissant de migrants sur une île déjà fort peuplée. En début de jeu, les joueurs tirent au sort leur situation professionnelle, la taille de leurs familles, ainsi que le lieu de leur domicile sur l'une des deux îles. Ils ne disposent que d'un nombre limité de seaux pour stocker de l'eau. Afin d'augmenter leur capacité de stockage, ils peuvent investir leurs jetons dans divers équipements pour accéder à l'eau : pompes à eau, citernes d'eau pluviale ou réservoirs d'approvisionnement, ces derniers étant gérés par la société d'eau locale.

Pendant leur tour, les joueurs peuvent accroître leurs revenus par le produit de leurs récoltes, cependant l'irrigation nécessaire à l'activité agricole fait augmenter le besoin d'eau. L'objectif de chaque joueur est de pourvoir aux besoins en eau de sa famille. S'il échoue, le nombre de personnes malades (en cas d'eau insalubre) ou mécontents (en cas d'absence d'eau) augmente. L'eau peut être insalubre parce qu'elle est polluée ou salée. Le degré de salinité de la lentille d'eau douce est calculé à chaque tour par le simulateur, dont l'espace virtuel reflète les deux plateaux de jeu. L'entité spatiale minimale du simulateur est une « cellule d'atoll », dotée de quelques attributs locaux relatifs à la lentille d'eau, comme sa qualité et sa profondeur. La lentille d'eau elle-même n'est rien d'autre qu'un agrégat de telles cellules, qui renseigne sur la recharge (ou la déplétion) de l'eau souterraine disponible. Enfin, l'île en tant que telle est également un agrégat de cellules, renseignant sur le climat et notamment la pluviométrie. Sur chaque cellule (ou case du plateau), l'eau peut être douce,

chose que des diagrammes de classes, inaptes à modéliser des comportements (ou interactions) entre entités. Il faut certainement supposer que d'autres diagrammes, plus dynamiques, sont également utilisés. Cependant, même en faisant une telle supposition, l'opération – pourtant de toutes la moins triviale – qui consiste à formaliser un comportement, à relier un effet d'un agent sur une entité, ne semble pas faire l'objet d'une attention particulière.

[136] Les travaux consultés usent du terme – fort malencontreux selon nous – de jeu de rôles (*role-based game*) : or le jeu de rôles prototypique n'est-il pas celui de la petite fille jouant à la dînette ou à être maman ? Ce type de jeux est donc tout à fait dépourvu de la dimension spatiale, pourtant essentielle aux jeux utilisés en modélisation d'accompagnement, puisque c'est la spatialité qui permet une traduction directe du jeu en simulation à base d'agents.

[137] Pour une typologie des possibilités, voir M. ÉTIENNE, *op. cit.*, pp. 75-77.

[138] Afin de garantir l'égale compréhension du jeu par tous, tous les matériaux du jeu étaient bilingues anglais – langue de l'administration – et gilbertin – langue locale parlée.

saumâtre, ou saline. La salinité dépend du climat (plus le temps est sec, plus l'eau devient salée), ainsi que de la proximité de l'océan (ou du lagon). La citerne d'eau se remplit normalement à la fin du tour, sauf en cas de sécheresse. Le réservoir d'approvisionnement quant à lui restera vide si la canalisation entretenue par l'État n'a pu pas acheminer assez d'eau.

L'interaction entre joueurs prend deux formes. La première forme est la négociation lorsqu'un joueur veut déménager d'une île à l'autre. Il doit, pour ce faire, trouver un terrain à acheter : la négociation portera alors sur le prix (en jetons) des parcelles. L'autre forme d'interaction se présente lorsqu'arrivent des évènements qui doivent être gérés collectivement. Parmi ces évènements, citons la décision gouvernementale de créer une réserve d'eau douce sur une île. Une telle décision implique que tous les joueurs concernés doivent trouver à relocaliser leurs proches qui se situent sur le territoire de la nouvelle réserve ; ils perdent en outre toutes les cultures qui s'y trouvent. Un autre évènement possible est la décision gouvernementale consistant à lever une taxe sur l'eau. Un des joueurs devient alors gestionnaire d'eau, chargé de percevoir les taxes. D'autres évènements affectent le débit des canalisations, et donc du rendement des réservoirs d'approvisionnement.

Notons que le jeu se construit sur la migration induite par l'obligation d'héberger sa parentèle. L'enjeu « scientifique » à la base de la démarche d'accompagnement passe au second plan. Ce renversement des priorités est notamment manifeste dans le traitement des causes de la pollution de l'eau douce qui sévit sur l'atoll : comme celles-ci ne faisaient pas l'objet d'un consensus parmi les personnes interviewées, voire étaient source de conflit, les modélisateurs ont pris le parti de laisser au simulateur le soin de semer aléatoirement des foyers de pollutions dans le jeu, tout en veillant à ce que leur distribution soit réaliste. La pollution est ainsi traitée comme une donnée environnementale, dont le pourquoi reste inexploré. Les connaissances sont donc extraites – ni construites ni injectées – par le modélisateur au prix d'un patient travail[139]. Comme le disent les auteurs :

> *[…] we argue that local communities need to be involved not only in the analysis of the results (consultation) or the choice of the possible scenarios (participation) but in the knowledge creation itself (engagement).*[140]

Le modèle ainsi constitué *s'enrichit* des connaissances scientifiques ; il convient de souligner que celles-ci *ne remplacent pas* les connaissances qu'il contient déjà. L'objectif poursuivi par la modélisation est d'abord de déplier l'éventail des points de vue sur une problématique donnée[141]. Seulement ensuite – et dans la mesure du possible – l'effort pourra porter sur la synthèse qu'est une représentation commune, cohérente entre tous les points de vue, y compris celui du concepteur de la modélisation.

Remarquons à ce propos que le modélisateur endosse ici un rôle un peu particulier : ni tout à fait chercheur, ni tout à fait médiateur, ou plutôt les deux à la fois, il devient *participant* dans un effort collectif, qui engage tant des chercheurs et des communautés locales, pour co-construire le modèle

[139] Dans un cas de modélisation d'accompagnement rapporté par M. ÉTIENNE (*op. cit*., p. 29), le travail s'est poursuivi pendant sept ans !
[140] A. DRAY., P. PEREZ, Chr. LE PAGE, P. D'AQUINO et I. WHITE, *Who wants to terminate the game?*, p. 509.
[141] Ce point est fortement mis en relief par M. ÉTIENNE, *op. cit.*, p. 51.

multi-agents. S'il y a là-dedans production de connaissance, c'est celle d'une *représentation partagée* du monde ; celle-ci ne relève pas tout à fait d'une description objective de la réalité, passible de vérité ou de fausseté, mais se fait mieux saisir par les notions ricœuriennes de témoignage et d'attestation. Les auteurs, en effet, développent l'idée selon laquelle l'adéquation du modèle à la réalité doit se comprendre dans « un double processus de traduction et d'interprétation » : la simulation est une traduction du monde réel, simulation dont les résultats doivent être interprétés « en modalités d'action dans le monde réel » :

> *Ce double processus de traduction et d'interprétation constitue un des moteurs essentiels de la modélisation d'accompagnement : une évolution du monde virtuel amène à de nouvelles simulations et à l'opportunité d'en discuter leur signification pour le monde réel, les changements induits sur le monde réel ou au minimum sur les points de vue des acteurs sur celui-ci amènent à réviser sa représentation dans le monde virtuel.* [142]

Si l'enjeu est d'amener les acteurs concernés à revoir leurs représentations du problème, les modèles produits par la démarche doivent donc être vus d'abord comme des supports d'un *apprentissage collectif*[143] : à la rigueur, le modèle peut être scientifiquement « faux » mais être un vecteur fertile d'apprentissage. Une telle conception se comprend évidemment dès lors que les exigences propres à la prise de décision sont prises au sérieux : dans une prise de décision, en effet, la qualité de la décision dépend en priorité du processus de décision lui-même, avec tout ce que cela implique : dialogue entre acteurs concernés, création collective de possibilités d'action, etc[144].

La création d'une représentation commune ne vise pas à la substituer aux représentations plurielles ; il s'agit plutôt d'élaborer un dénominateur commun, par rapport auxquels les différents points de vues peuvent utilement se référer ; si la réussite de la démarche doit être évaluée et saisie en termes de création et d'acquisition de savoirs et de savoir-être, le rôle du scientifique se précise alors : il est considéré comme un acteur *parmi d'autres*, dont le savoir peut être interrogé, remis en question, à tout moment, en fonction du contexte de la concertation. Chaque acteur, scientifique ou non, est *en droit* porteur d'une connaissance qui lui est propre. Ainsi, au travers d'une typologie des savoirs[145] (empirique, académique, technique, institutionnel), s'esquisse une éthique de la recherche scientifique, un « savoir-être » du chercheur :

> *[…] nous considérons que le scientifique, engagé dans une démarche de recherche impliquée comme la modélisation d'accompagnement, entre dans un jeu d'acteurs sociaux dépassant les limites du champ scientifique dans lequel il accepte de se rendre public et de diffuser une certaine image de soi. Sa posture présente une double dimension rhétorique et actionnelle, elle se traduit donc par une prise de position morale, affective, sociale, philosophique et politique qui conduit à des actes. En parlant de la posture du commodien [c'est-à-dire du*

[142] M. ÉTIENNE, *op. cit.*, p. 25.
[143] Les apprentissages peuvent être de nature diverse : organisationnel, communicationnel, technique, etc. Nous renvoyons le lecteur intéressé à *op. cit.*, pp. 235-245, pour une présentation détaillée.
[144] Pour une présentation plus fournie, voir *op. cit*., pp. 153-164.
[145] *Ibid.*, p. 23.

*modélisateur d'accompagnement], nous faisons ainsi référence à une façon particulière de penser la position du chercheur dans les relations entre science et société.*[146]

L'importance du jeu – cette caractéristique au prime abord si déroutante de la démarche[147] – se comprend dès lors mieux. Or il nous reste à comprendre comment la simulation à base d'agents et jeux se conjuguent, se complètent, se renforcent mutuellement. Il faut partir ici de l'idée que la modélisation d'accompagnement se fonde sur ce qu'il faut bien appeler une métathéorie, à savoir la possibilité d'appliquer une *représentation* multi-agents à toute problématique[148] : une telle métareprésentation est utilisable même lorsqu'aucun support informatique n'est présent.

Les avantages du jeu dans le monde « réel » sont bien connus, que ce soit à des fins thérapeutiques, professionnelles…[149] Parfois, les parties dans le conflit sont invitées à jouer le rôle de leur « adversaire » : il s'agit d'apprendre ce qui importe, ce qui compte pour l'autre. Comme c'est un jeu, sans « enjeu », le recul est possible, voire la bonne humeur[150]. En fonction du parcours de modélisation, la SBA peut intervenir en amont ou en aval d'un jeu donné : ainsi il est possible – en amont – de construire un jeu à partir d'une (version simplifiée d'une) simulation afin de faire valider celle-ci par les parties prenantes ; en aval, la simulation permet d'explorer un nombre quasi infini de parties, rendant ainsi possible d'explorer autant de scénarios possibles que les participants ont d'idées[151].

La simulation peut également faire partie intégrante du jeu : ici encore, plusieurs possibilités, en fonction surtout de savoir si toutes les décisions restent aux mains des joueurs humains ou si une partie d'entre elles sont déléguées à l'ordinateur. L'intérêt majeur de la SBA dans ce cadre réside cependant moins dans les agents virtuels à proprement parler qu'à l'environnement spatio-temporel qu'elle fait exister : il est possible, pour chaque tour du jeu, de simuler l'évolution des ressources, de visualiser des indicateurs de performance, sur la base des décisions des joueurs, de toutes les données pertinentes dans le contexte, etc. C'est d'ailleurs ainsi que dans *AtollGame*, le degré de salinité de chaque cellule a pu être recalculé à chaque tour du jeu sans en ralentir le rythme. Même si certaines de ces fonctions sont également à la portée d'un simple tableur, ici encore les plateformes de simulation ont des atouts considérables de visualisation et d'intégration des informations.

La frontière entre simulation et jeu devient parfois bien floue, au point que les simulations sont parfois appelées des « jeux habitables », dans la mesure où elles permettent véritablement d'immerger leurs participants dans des systèmes complexes et de les faire prendre conscience des rôles – actuels ou potentiels – qu'ils peuvent avoir dans la transformation de ceux-ci ; nous retrouvons une problématique déjà entrevue au deuxième chapitre. Dans le contexte de la

---

[146] *Op. cit.*, p. 47. Voir aussi les pages 49-51.
[147] Caractéristique si singulière qu'elle a fait dire aux auteurs que « le recours à des modèles non-informatiques (jeux de plateaux) est (paradoxalement) encore un des aspects les plus originaux de la démarche » (*op. cit.*, p. 42).
[148] Cf. *ibid.*, p. 72-81.
[149] Le lecteur intéressé peut se référer aux sites https://www.gamesforchange.org/ et https://www.serious-game.fr/.
[150] Le lecteur intéressé peut écouter les témoignages sur https://www.commod.org/qui-sommes-nous/videos.
[151] En vérité, différents séquençages entre jeu et simulations sont envisageables, en fonction des besoins et du contexte : voir M. ÉTIENNE, *op. cit.*, pp. 99-101, pour un plus ample développement à ce sujet.

modélisation d'accompagnement, l'immersion ludique devient un critère de réussite de la démarche : en effet, si le modèle est trop abstrait, le risque existe que les participants n'arrivent pas à s'y projeter et qu'ils finissent par le rejeter. À l'inverse, un modèle trop réaliste aura beaucoup plus de chances de susciter la confiance des participants, mais il comporte son propre risque : celui d'échouer à dépayser les joueurs, à les faire entrer dans un monde « nouveau », condition *sine qua non* pour se distancier du cas concret et d'envisager de nouvelles pistes de solution.

En définitive, le véritable résultat recherché par une telle démarche d'accompagnement, c'est un *engagement* ; engagement à produire des connaissances, engagement à chercher des solutions ensemble, engagement à collaborer dans un esprit d'entente… Nous sommes ici aux antipodes d'une démarche d'optimisation d'une fonction objectif donnée : ce qui compte avant tout, c'est la *qualité du processus* qui aboutit au résultat ; la priorité étant que chaque groupe sente ses aspirations, ses craintes et ses problèmes pris au sérieux.

#### *3.4.2.7. Forces et limites de la SBA comme outil d'aide à la décision*

Que retenir de ce – trop bref – survol de la SBA comme outil d'aide à la décision ? La décision, à coup sûr, est d'un autre ordre que la recherche pure : celle-ci s'accommode d'apories, de lacunes, voire de contradictions plus ou moins temporaires ; rien de tout cela lorsqu'il s'agit de prendre une décision : la tergiversation n'est pas permise, même lorsque la plupart des données dont nous aimerions pouvoir disposer nous font défaut. La décision nous engage ; elle a des exigences auxquelles le chercheur, en règle générale, ne songe que dans ses cauchemars.

Qui dit décision, dit décideur : *qui* décide ? Les auteurs du rapport de développement durable qui ont été nos guides dans cette section font remarquer que les décideurs sont pour ainsi dire toujours absents des modèles. Cette absence, selon eux, se défend : il s'agit de modéliser le système tel qu'il évoluerait en l'absence d'intervention extérieure. La décision, dans cette vue, consiste alors à intervenir sur le modèle *de l'extérieur*, un peu comme un *deus ex machina*, qui change les paramètres, ajuste les comportements, restructure ou revalorise les éléments du modèle. Une telle vue présuppose, peu ou prou, l'omniscience du décideur : dans la mesure où le modèle capte tout ce que nous pouvons connaître de la réalité, il semble dès lors aller de soi que tous ces aspects sont connus du décideur sans médiation, sans délai, sans bornes. Or une première force de la SBA doit ici être soulignée : nous avons vu, dans COMOKIT, l'Autorité prendre place *à l'intérieur même* de la simulation : ses décisions dépendaient de la perception qu'elle avait de la situation, qu'elle pouvait en avoir : perception toujours située et, par là même, limitée. Ceci reste vrai, même si l'accès à l'Autorité par les autres agents était, lui, immédiat. En effet, il est facile de voir que cet accès pourrait être problématisé de façon bidirectionnelle. Nous aurions alors une simulation entièrement située de la prise de décision elle-même.

Ce qui est vrai du décideur, est encore plus vrai du modélisateur : *qui* modélise ? Car si les décideurs se font généralement discrets dans les modèles, le modélisateur n'est, lui, même pas mentionné. Son point de vue semble presque par nécessité externe, surplombant la mêlée. Cette perception du rôle du modélisateur change du tout au tout dans une démarche dite de modélisation

d'accompagnement. La déontologie de la démarche oblige celui qui la pratique à expliciter tous ses présupposés, toutes ses hypothèses, afin de les soumettre au tribunal souverain des autres participants. Le mot important, ici, est « autre » : le modélisateur n'est qu'un participant parmi d'autres ; son système de valeurs, sa conception du vrai ou du juste, rien ne devrait être épargné – du moins en principe – par la confrontation avec les autres, qui comme lui sont porteurs d'un savoir : pour reprendre les critères cognitifs du développement durable, il serait inexact de considérer que la modélisation d'accompagnement excelle dans le critère de *participation*, *indépendamment* du critère de *l'interdisciplinarité*. Dans cette démarche, la distinction même entre participation et interdisciplinarité est vue comme *illégitime* !

Si nous nous tournons vers les contenus des savoirs véhiculés par la SBA, nous avons vu tout d'abord qu'elle n'intervient guère dans la modélisation axiologique : en effet, son rôle dans la valorisation des préférences ou les questions budgétaires est modeste. Du moins a-t-elle le mérite de pouvoir exprimer, sur le plan cognitif, tous les effets dont les outils d'intégration axiologique ont besoin. Dans la modélisation cognitive, en revanche, elle excelle dans l'usage désormais presque classique de la SBA et que nous avons vu au deuxième chapitre (§ 2.3.2) : susciter la surprise chez ses utilisateurs en donnant vie à des effets, à des dynamiques, inattendus. En cela, elle constitue déjà une aide précieuse dans l'exploration des actions possibles. Ainsi, dans une variante de Polsim, il suffisait d'augmenter l'information des ménages à faibles revenus pour que ceux-ci trouvent à se loger à des conditions plus en ligne avec leurs préférences personnelles[152].

Vue sous cet angle, en tant qu'elle capte des tendances, des dynamiques globales, la SBA fournit une aide précieuse pour simuler, et par là comprendre, les comportements. À ce titre, la SBA se révèle très appropriée au long, voire au très long terme. De ce fait, elle constitue un outil privilégié dans les exercices d'anticipation[153]. Elle présente également des atouts très importants en matière d'interdisciplinarité : il faut alors penser tout d'abord aux disciplines scientifiques, mais elle permet également de tendre la main à une discipline comme l'histoire. Son expressivité est telle que non seulement les données quantitatives, mais aussi qualitatives – sous forme de comportements codés – ont droit de cité dans la modélisation.

Nous avons lu aussi, dans le rapport de développement durable, que le rôle de la SBA en tant qu'outil de pronostic, aux prévisions à courte durée, serait voué à rester faible. Or COMOKIT vient prouver le contraire : tirant profit de la capacité d'intégration sur base d'une résolution spatiale de plus en plus fine, la SBA se montre ici capable d'une granularité réaliste qui la rend fiable dans des pronostics à court ou à moyen terme. Il est vrai que cet exemple appartient à un de ces contextes très favorables aux outils à bonne résolution spatiale, domaine où règnent traditionnellement les SIG et où la dynamique à étudier est essentiellement celle d'une propagation spatiale.

Dans le cas de COMOKIT, l'accent est clairement mis sur l'interdisciplinarité ; la validation du modèle se fait par un « retour au réel », interprété à l'aune des méthodes statistiques. Dans le cas

[152] P.-M. BOULANGER et Th. BRÉCHET, o*p. cit.*, pp. 131-134.

[153] L'exercice d'anticipation tel que considéré ici n'est pas sans rappeler la SF. Cependant, alors que la SF se soucie seulement de la cohérence d'un monde possible, la SBA ne perd jamais de vue les probabilités qui s'attachent aux mondes qu'elle donne à voir.

d'*AtollGame*, nous avons vu ce surprenant phénomène où la validation du modèle est confiée à un jeu. L'adéquation recherchée n'est pas celle à une réalité conçue comme externe aux participants, mais à la représentation que ceux-ci peuvent en avoir. Que la SBA puisse servir d'outil fiable, fertile et robuste dans deux conceptions scientifiques si éloignées l'une de l'autre – pourtant toutes deux aux prises avec la problématique difficile de la prise de décision – voilà un fait remarquable qui doit très certainement être porté à son crédit.

Les différences entre ces deux démarches ne devraient cependant pas celer une incidence commune majeure sur la SBA en tant que technique : dans les deux cas, nous observons comme un glissement du centre de gravité des préoccupations. En effet, tout se passe comme si la notion *d'environnement* compte tout autant, sinon plus, que celle d'agent à proprement parler[154]. Ceci est tout à fait manifeste pour ce qui est de la modélisation d'accompagnement : dans la mesure où la dimension spatiale en vient à jouer le rôle de matrice de tous les savoirs, elle devient parfois le dépositaire principal de l'intelligence de la simulation. Pour nous qui nous interrogeons sur l'éthique, ce déplacement du centre des préoccupation est en lui-même porteur de son lot de questions : à côté des questions traditionnelles qui porte sur le « qui », le « pourquoi » et le « comment », nous sommes désormais invité à poser des questions foncièrement *contextuelles* : « où », bien sûr, mais aussi « quand ? », « au moment opportun ? », « à l'endroit souhaité ? »[155].

Malheureusement, la grande expressivité de la SBA est également son point faible : elle permet d'implémenter tous les comportements, sans garde-fou ni restrictions. L'absence de cadre formel rend malaisée la validation du modèle, autre que par la correspondance du réel aux résultats (déjà agrégés). Sa sensibilité aux conditions initiales (et à l'ordonnancement des agents) rend délicat son calibrage et condamne le modélisateur à de laborieux échantillonnages. Même sa grande capacité spatiale et visuelle, ses possibilités de représentation graphique proches des SIG, peut paraître comme une faiblesse dès lors que nous considérons qu'aucune unité ne préside aux choix du modélisateur en la matière.

---

[154] Un point qui n'a pas échappé à nos auteurs lorsqu'ils affirment que la SBA permet d'exprimer l'idée même du développement durable, dans la mesure où elle permet de modéliser des transformations de *l'environnement* (P.-M. BOULANGER et Th. BRÉCHET, *op. cit.*, p. 43) : ce que prime n'est pas tant les agents en tant que tels, mais leurs interactions avec l'environnement, compris lui-même comme une topologie.

[155] Comme nous l'avons bien vu à propos de l'affaire des caricatures du prophète Mahomet, faire abstraction de telles questions peut revenir à faire preuve de violence envers l'autre : « Nous pouvons rire de nos propres symboles sacrés, mais pas de ceux des autres – surtout si nous appartenons à un groupe qui les a historiquement opprimés. » (STARHAWK, *Quel monde voulons-nous ?*, p. 109). Même si nous n'avons pas ici l'occasion d'approfondir cette piste, signalons toutefois que *l'image de l'homme* qui s'en dégage – un homme « ancré » ou « enraciné » dans une terre – pour n'être pas nouvelle, est pourtant au cœur de certaines pensées qui nourrissent le mouvement altermondialiste. Nous pensons ici particulièrement à STARHAWK (*Quel monde voulons-nous ?*, pp. 53-60), auteure qui a derrière elle une longue carrière d'activiste, du Vietnam à Seattle. Lorsque l'homme paraît ainsi lié non seulement à sa famille, à ses proches et à sa nation, mais aussi à la terre, à la faune et à la flore qui l'entourent, un homme *sous la dépendance du contexte* qui lui a permis de naître, de vivre et de s'épanouir, la négation de l'espace devient une *figure du mal*, où le mal doit être entendu de deux manières : le *mal commis*, le *mal souffert* (cf. P. RICŒUR, *Le mal*, pp. 21-22) : le réfugié, jeté sur les routes ou croupissant dans un centre fermé, ne porte-t-il pas les stigmates de notre époque ? Le mal commis, où est-il plus manifeste que dans la rapacité des multinationales, des firmes qui, étant sans attaches, n'ont de comptes à rendre à personne ?

Non seulement en matière de comportements, mais également le choix des agents eux-mêmes est entièrement libre : le risque existe dès lors que « n'importe quoi » est pris comme comportement, de même que « n'importe qui » fait figure d'agent. Contrairement aux réseaux bayésiens et aux modèles de dynamique de système, la SBA ne peut se prévaloir d'une ontologie scientifique, ni même d'une robuste métathéorie. La notion d'agent a dès lors quelque chose d'énigmatique : tout est « agentifiable ». Cette objection peut, certes, être nuancée : le modélisateur peut toujours faire appel à des travaux sociologiques, psychologiques – les auteurs de COMOKIT ne s'en sont d'ailleurs pas privés. Il n'en demeure pas moins que rien, dans la modélisation multi-agents, ne les pousse à fonder leur modèle de la sorte. Le choix des agents constitue donc un agenda de recherche urgent, au même titre que les domaines de calibrage et de validation. En attendant, l'arbitraire qui s'y attache risque toujours d'être pris dans les rets des jeux de pouvoir[156]. Qu'une telle situation risque parfois d'ouvrir la porte à des formes plus ou moins subtiles de violence, voilà un danger dont le cas pratique que nous allons étudier dans les pages qui suivent se fait l'écho[157].

### 3.4.3. Les agents sur la route : un cas d'innovation irresponsable ?

Le 18 mars 2018, un dimanche soir à 21h58, Rafaela Vasquez percute Elaine Marie Herzberg au volant d'une Volvo XC90 sur l'avenue Mill, dans la ville de Tempe, en Arizona (États-Unis)[158]. La victime, une SDF de 49 ans et qui décédera sur place suite à ses blessures, traversait l'avenue qui à cet endroit comptait 4 bandes et une piste cyclable. La conductrice de la Volvo, employée par Uber, était censée surveiller la performance de la voiture, modifiée pour être entièrement autonome sur la route, pendant les trajets de test de l'autopilote. Dans cette section, nous survolerons rapidement le contexte de l'accident, avant de nous concentrer sur la faille logicielle qui y a grandement contribué. Nous conclurons sur quelques réflexions quant à la pertinence du paradigme multi-agents pour analyser de tels cas, ainsi que sur la valeur ajoutée des voitures autonomes à l'aune d'une démarche d'innovation qui se voudrait *responsable*.

#### *3.4.3.1 Un accident de la route évitable…*

L'accident n'avait rien d'inévitable. Nous nous attacherons, dans les lignes qui suivent, à relever les circonstances qui ont conduit à cette malheureuse issue. Le but de ce paragraphe n'est pas tant de distribuer les responsabilités, mais de faire sentir que les facteurs institutionnels et humains sont

[156] C'est un risque dont sont bien conscients les modélisateurs d'accompagnement, voir le chapitre de C. BARNAUD, P. D'AQUINO, W. DARÉ, Chr. FOURAGE, R. MATHEVET et G. TRÉBUIL, *Les asymétries de pouvoir dans les processus d'accompagnement*, dans M. ÉTIENNE, *op. cit.*, pp. 125-151.

[157] Le lecteur pressé peut déjà se reporter à la section détaillant la faille logicielle à l'origine de l'accident (§ 3.4.3.3).

[158] Dans cette section, nous nous basons de manière presque exclusive sur les rapports rendus disponibles par le NTSB, agence états-unienne indépendante en charge « d'enquêter sur les causes probables des accidents de transport, de promouvoir la sécurité des transports et d'assister les victimes des accidents de transport et leurs familles ». Même si tous les rapports cités peuvent être retrouvés individuellement dans notre bibliographie, le lecteur intéresse pourra probablement se retrouver plus facilement sur la page « fiche de renseigments » (*docket*) de l'agence consacrée à l'accident et où l'ensemble des documents sont disponibles : https://data.ntsb.gov/Docket?NTSBNumber=HWY18MH010.

tout aussi importants, voire davantage, que la faille technique dont nous reparlerons au paragraphe suivant.

Commençons ce tour par la victime elle-même : elle traversait à un endroit sans passage pour piétons, avec au contraire un marquage explicite interdisant la traversée. À cause des phares de la voiture, la piétonne avait l'occasion de s'apercevoir de la voiture dès que celle-ci s'était engagée sur ce tronçon de route, soit 5,6 secondes avant la collision[159]. Dans 6 sur 15 accidents impliquant des piétons, le piéton est sous l'influence de substances illicites (alcool ou autres). C'est donc sans trop grande surprise que l'examen post-mortem a révélé la présence de psychotropes dans le sang de la victime : la marijuana ne s'y trouvait qu'à un niveau résiduel. En revanche, le niveau de méthamphétamine (mieux connue, sous son appellation argotique, comme le *speed*) indiquait un abus chronique de cette substance. Or celle-ci peut affecter sérieusement la perception et le jugement. Les enquêteurs n'ont pas pu reconstituer l'emploi du temps de la victime avant l'accident : elle était SDF, ses proches très réticents à s'exprimer devant les enquêteurs.

Continuons notre tour par l'État d'Arizona : en 2015, à l'occasion du décret (*executive order*) autorisant la mise en circulation de voitures autonomes à des fins de test, un comité fut instauré, le le « *Self-Driving Vehicle Oversight Committee* », pour veiller sur ces tests et proposer des amendements dans la législation de l'État si cela devait s'avérer nécessaire[160]. De sa création jusqu'en 2018, le Comité s'était réuni deux fois : à la première séance, une lecture fut donnée du décret, un président fut désigné, les membres ont écouté une présentation sur les voitures autonomes. Lors de la deuxième séance, la législation des autres États fut examinée et jugée inadéquate : beaucoup de bureaucratie sans plus-value notoire du point de vue de la sécurité. Aucune action supplémentaire ne fut dès lors entreprise, ni même aucune information collectée auprès des sociétés testant des voitures autonomes. Il faut dire, en effet, que des États comme la Californie ou la Pennsylvanie ont des politiques autrement contraignantes en la matière[161] : ainsi la règlementation de la Californie requiert notamment qu'une société désireuse de faire ce type de tests obtienne un permis et qu'elle fasse rapport de tous les incidents. Toutefois, l'accident mortel n'a pas été pour l'Arizona une raison suffisante pour revoir sa politique en matière de voitures autonomes.

Tournons-nous maintenant vers Uber : le NTSB a monté en épingle la « culture de sécurité inadéquate »[162] du groupe. À plusieurs reprises, Uber a échoué à donner aux considérations de sécurité l'attention requise, supprimant des couches de sécurité qui, avec le recul, auraient pu jouer un rôle déterminant dans l'accident de Tempe. Jusqu'en septembre 2017, les trajets de test étaient effectués en binôme : le deuxième opérateur avait la charge de diverses tâches liées à la documentation du parcours. Or ces tâches de documentation étaient rendues singulièrement plus faciles à accomplir lorsqu'une nouvelle interface homme-machine - sous forme de tablette – fut introduite dans les voitures. Les trajets se firent désormais par le conducteur seul et ce, malgré les bonnes pratiques relatives à ce type de tâche de surveillance[163]. En outre, même si Uber avait la

[159] NTSB, Highway Accident Report, 2019, pp. 48-49. La limite de vitesse – 45 milles par heure, soit environ 72 km/h – était respectée.
[160] NTSB, *Human Performance Group Chairman's Factual Report*, pp. 17-19.
[161] NTSB, *Highway Accident Report*, 2019, pp. 52-56.
[162] NTSB, *Highway Accident Report*, 2019, pp. 57-58.
[163] NTSB, *Human Performance Group Chairman's Factual Report*, pp. 18, 29.

possibilité de contrôler que les conducteurs suivent les consignes de sécurité (grâce à une caméra filmant l'intérieur de la voiture), cette possibilité n'a été exploitée que rarement. Cette omission est d'autant plus surprenante qu'Uber était parfaitement au courant que depuis la suppression du binôme, le nombre de licenciements à cause de l'usage non autorisé de téléphones portables au volant a significativement augmenté[164]. Toujours dans le champ des ressources humaines, Uber négligeait aussi les normes et bonnes pratiques applicables aux conducteurs commerciaux : ainsi le groupe ne procédait pas à des tests de dépistage d'alcool ou de drogues avant l'embauche, ni même après des accidents ; il ne demandait pas non plus un bilan de santé pour prouver l'aptitude à conduire avant l'embauche.

Un autre choix malheureux fut la désactivation du système de sécurité intégré de Volvo. En effet, les voitures Volvo utilisées dans les tests disposent déjà d'un système avancé d'assistance à la conduite, dont un dispositif de freinage d'urgence. Or ces mécanismes avancés de sécurité de Volvo étaient désactivés lorsque la voiture roulait en mode automatique, n'étant réactivés que lorsque le conducteur reprenait le contrôle du véhicule. Selon Uber, la possibilité d'interférences entre le dispositif de Volvo et le sien était à la source de cette décision. Or dans les mesures correctrices post-accident figure l'harmonisation des deux systèmes, ce qui incite à croire que l'interférence n'a pas dû constituer un problème technique majeur. Cette désactivation est d'autant plus malheureuse qu'un test de reconstitution[165] a montré qu'un conducteur alerte, avec le système de Volvo actif, aurait pu limiter la vitesse d'impact au moment de la collision à 15 km/h. Ce chiffre a par ailleurs été confirmée par une simulation menée par Volvo[166].

Le système de freinage d'urgence d'Uber – censé remplacer celui de Volvo – était quant à lui encore dans un stade expérimental : notamment, il présentait un nombre important de faux positifs, faisant dès lors parfois plus de mal que de bien. Pour diminuer les coups de frein impromptus, Uber avait implémenté une mesure dite « de suppression d'action »[167]. Si le système de freinage détecte une urgence, celle-ci est ignorée pendant une seconde, sans donner aucun signal au conducteur. Après cette seconde, si l'urgence persiste, la voiture entame un freinage progressif. Le conducteur n'est averti que lorsqu'une collision devient inévitable : ce qui explique que même si la voiture a reconnu une situation d'urgence 1,6 secondes avant la collision, la suppression d'action a été initiée 1,3 secondes avant et le véhicule n'a sollicité la conductrice de reprendre la main que 0,28 secondes avant[168]. Autant dire – et nous nous départons ici volontairement du ton très neutre du rapport du NTSB – que dans de telles conditions, la catastrophe était pour ainsi dire *programmée*, courue d'avance !

---

[164] NTSB, *Human Performance Group Chairman's Factual Report*, pp. 11-12. Ces licenciements étaient surtout dus à des dénonciations de la part de collègues, attitude qu'Uber encourageait formellement.
[165] Voir le rapport de NTSB, *Volvo XC90 Testing by Thatcham Research*.
[166] Voir NTSB, *Highway Accident Report*, 2019, pp. 21-22. Pour être précis, la simulation prédit un évitement dans 17 cas sur 20, et une réduction de la vitesse dans les trois autres cas. Hâtons-nous cependant d'ajouter qu'aucune précision technique quant à la nature de cette simulation ne figure dans les rapports : une attitude critique est donc de mise.
[167] La suppression d'action est décrite dans NTSB, *Highway Accident Report*, 2019, pp. 13-14.
[168] Cf. NTSB, *Human Performance Group Chairman's Factual Report*, p. 14.

### *3.4.3.2. L'automatisation et l'excès de confiance*

Pour clore ce rapide parcours des facteurs humains et institutionnels, il faut nous concentrer maintenant sur la conductrice. Après tout, elle reluquait fréquemment son téléphone portable avant l'accident, malgré des instructions strictes d'Uber à ce sujet. C'est là un manquement grave, sans doute – pour lequel il faut croire que le tribunal ne fera preuve d'aucune indulgence[169]. Il faut malgré tout regarder plus loin, ne pas nous contenter du premier homme (ou femme) que nous pouvons accabler du lourd fardeau de la culpabilité. Aurions-nous, à sa place, avec plus de bonne volonté peut-être, fait autre chose ? C'est loin d'être sûr, comme nous allons le voir maintenant.

Le problème majeur que pose l'automatisation de la conduite automobile semble bien être, en premier lieu, lié au psychisme humain : il s'agit d'un phénomène connu sous le nom d'excès de confiance induit par l'automatisation (*automation-induced complacency*). Tout au long du trajet, la conductrice n'a cessé de reluquer son téléphone afin de regarder une émission de télévision proposée par la plate-fome de diffusion Hulu. Avant la collision, pendant environ 5,3 secondes, l'opératrice n'a pas levé les yeux une seule fois sur la route[170]. Or dans une reconstitution de l'accident, la police de Tempe a constaté qu'un conducteur normalement attentif aurait dû détecter la piétonne entre 4 à 2 secondes avant la collision[171]. Comme le temps nécessaire à un évitement de l'accident s'élevait à 1,25 secondes, il est possible de comprendre que la cause première de l'accident, selon le rapport du NTSB, est l'inattention de la conductrice induite par un excès de confiance dans les capacités de la voiture à tenir la route. En cela, le cas n'est pas sans rappeler le premier accident mortel impliquant une voiture autonome, où a péri le propriétaire d'une voiture Tesla qui utilisait la fonction autopilote en dehors d'une situation officiellement prise en charge[172].

Ce biais de la cognition humaine est connu et documenté depuis les années '90 du siècle dernier au moins dans d'autres domaines automatisés. Dans une communication du groupe Volvo[173] – fournisseur d'Uber – nous apprenons que les constructeurs de voitures autonomes assumaient généralement que ce biais ne s'appliquait pas à la surveillance routière que leurs conducteurs sont censés exercer, étant donné que leur propre sécurité était directement mise en jeu. Or une expérience menée par Volvo tend à prouver le contraire ; dans cette expérience, un certain nombre

---

[169] La conductrice a entre-temps été inculpée pour homicide involontaire.

[170] NTSB, *Highway Accident Report*, 2019, p. 42.

[171] La reconstruction correspond assez bien au matériel fourni par les caméras de la Volvo, même si elle est assez optimiste : sur les caméras, la piétonne apparaît deux secondes avant la collision (NTSB, *Human Performance Group Chairman's Factual Report*, pp. 7-8).

[172] Voir, là encore, le *Highway Accident Report* qu'a dressé le NTSB en 2017. Les deux cas se distinguent cependant dans la mesure où la voiture commercialisée par Tesla n'ambitionnait nullement d'être pleinement autonome. Sur l'échelle d'automatisation automobile utilisée assez couramment dans l'industrie (SAE, disponible sur https://www.sae.org/standards/content/j3016_201806/), la voiture Tesla ne revendiquait qu'une autonomie partielle, de niveau 2, alors que la voiture d'Uber est beaucoup plus ambitieuse, ciblant le niveau 4 (sur 5), soit un niveau d'automatisation élevée. Cependant, comme le souligne le rapport (NTSB, *Highway Accident report*, 2019, p. 60), il faut toujours garder à l'esprit qu'il s'agit ici d'une norme industrielle et que la plupart de ces voitures, quelles qu'elles soient, sont commercialisées comme des voitures « autonomes », ce qui réduit considérablement l'incidence qu'une telle échelle peut avoir sur la conscience des consommateurs.

[173] Voir l'article que le groupe a envoyé au NTSB suite à l'accident de Tempe : VOLVO CARS, *Submission to the National Transportation Safety Board (N.T.S.B.) for the Tempe Accident involving an UBER test vehicle based on a Volvo XC90 MY2017*.

de sujets sont invités à tester une voiture autonome de Volvo. Après une demi-heure, la voiture était programmée pour perdre le contrôle de la route ; l'expérience visait surtout à déterminer l'influence sur l'attention des conducteurs des deux facteurs suivants : primo, la fréquence de rappels à la vigilance ; secundo, les instructions données au préalable aux conducteurs. Ce qui est intéressant dans les résultats, ce n'est pas tant que la fréquence des rappels était très efficace pour inciter les conducteurs à tenir les mains sur le volant et les yeux sur la route : c'était là un résultat, somme toute, attendu. Beaucoup plus intéressant en revanche est le fait qu'un nombre significatif de conducteurs avaient développé, à l'issue des 30 minutes, une confiance excessive dans la performance de la voiture : ils échouaient par conséquent à rattraper le couac programmé de la voiture et ce, *indifféremment* des avertissements reçus au préalable concernant la fiabilité de la voiture. Plus encore, la confiance excessive était même présente chez des conducteurs recevant fréquemment des rappels, ce qui suggère que le seul fait d'avoir le regard dirigé vers la route et les mains reposant sur le volant ne sont pas des indices fiables pour mesurer l'attention des conducteurs.

Les résultats à tirer de cette expérience sont assez clairs : l'inattention par excès de confiance s'installe également chez les conducteurs dont la vie dépend pourtant de la qualité de leur propre supervision. L'expérience confirme que l'être humain est un très piètre exécutant de ce type de tâches non seulement répétitives, mais surtout passives, d'autant plus lorsqu'il est seul. Les recommandations de Volvo tombent dès lors sous le sens : il faut trouver des moyens d'impliquer plus activement le conducteur ; il faut privilégier deux surveillants dans la voiture plutôt qu'un seul ; il faut assurer des rotations (*shifts*) fréquentes, de préférences toutes les 2 à 3 heures. Les auteurs de la communication n'hésitent d'ailleurs pas à tirer une conclusion assez radicale de l'expérience : si l'être humain est si médiocre pour surveiller la machine, il convient de soulever la question de l'utilité même d'une telle supervision :

> *These results illustrate that the concept of a fall-back ready driver needs to be questioned, and that more robust ways of ensuring test fleet safety likely need to be pursued.*[174]

Le NTSB a d'ailleurs suivi l'argument de Volvo, puisque nous lisons dans le rapport final ce qui suit :

> *When it comes to the human capacity to monitor an automation system for its failures, research findings are consistent — humans are very poor at this task. The NTSB concludes that the vehicle operator's prolonged visual distraction, a typical effect of automation complacency, led to her failure to detect the pedestrian in time to avoid the collision. The NTSB further concludes that the Uber ATG did not adequately recognize the risk of automation complacency and develop effective countermeasures to control the risk of vehicle operator disengagement, which contributed to the crash.*[175]

Insistons: l'hypothèse – assez répandue[176] – comme quoi la présence d'un surveillant dans le véhicule constitue une pièce plus ou moins maîtresse dans le dispositif de sécurité est tout simplement

[174] VOLVO CARS, *loc. cit.*, p. 6.
[175] NTSB, *Highway Accident Report*, 2019, p. 56.
[176] À titre d'exemple, nous la retrouvons dans les politiques de la Californie et de la Pennsylvanie en matière de conduite autonome. Ces politiques, même si elles sont par ailleurs assez différentes l'une de l'autre, ont en commun non seulement d'être nettement plus contraignantes que celle de l'Arizona, mais aussi de réserver une place privilégiée à la supervision humaine directe, à l'intérieur du véhicule (NTSB, *Highway Accident Report*, 2019, p. 65).

*erronée* ! Dès le moment où nous savons qu'aujourd'hui, la plupart des accidents de la route n'impliquent pas de défaillances techniques d'aucune sorte, mais sont dus à la négligence des conducteurs humains[177] ; qu'en outre ce même être humain est encore moins performant lorsqu'il s'agit de surveiller passivement – plutôt que de conduire activement –, il est pour le moins paradoxal de vouloir ériger la présence humaine comme une garantie de sécurité significative. Pour terminer ce paragraphe, posons-nous la question : si « excès de confiance » il y a, s'agit-il d'un excès de confiance en la machine… ou en l'homme ? Péchons-nous par excès d'optimisme vis-à-vis de la voiture, ou à l'égard de la faculté de l'homme de lui servir de guide ?

#### *3.4.3.3. La défaillance logicielle*

Les facteurs humains et organisationnels ont été la cause première de l'accident ; il n'en demeure pas moins qu'il y a eu, également, une faille technique, purement logicielle – raison pour laquelle, d'ailleurs, cet accident peut figurer dans un mémoire d'informatique : que s'est-il donc passé à ce niveau-là[178] ? Afin d'aborder la faille, exposons d'abord brièvement le fonctionnement du pilote automatique. Lorsque celui-ci est activé, il scrute continûment l'environnement à la recherche de nouveaux objets. Pour ce faire, la voiture dispose de trois outils principaux : un radar, un lidar[179], ainsi qu'un ensemble de 10 caméras. Une fois perçu, le logiciel va classifier l'objet selon une ontologie adaptée au trafic routier : l'objet peut être un véhicule, un piéton, un cycliste, ou encore tomber dans la catégorie fourre-tout « autre ».

Une fois l'objet perçu et classifié, un objectif lui est assigné : ainsi lorsque l'objet détecté est classifié comme un véhicule et que celui-ci se trouve sur une voie, le logiciel s'attendra à ce qu'il circule le long de la voie. Le logiciel va alors calculer des trajectoires possibles pour l'objet. Il se base sur la classification, l'objectif, mais aussi sur le parcours que l'objet a suivi jusque-là (*track history*). Ces projections de trajectoire (*path predictions*) sont continuellement mises à jour avec la dernière position connue de l'objet. Toutefois, lorsqu'un objet change de classification, il n'hérite pas de l'historique de parcours. Cette logique donne lieu au raisonnement hésitant de l'autopilote dans les quelques secondes qui vont de la première détection de la piétonne jusqu'à la collision, comme il ressort du tableau ci-dessous :

---

[177] J. K. GURNEY, *Imputing Driverhood*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, p. 53.

[178] Cette section se base principalement sur NTSB, *Vehicle Automation Report*.

[179] Un lidar émet des faisceaux lumineux qui sont réfléchis par les objets qu'ils croisent ; le temps nécessaire à la lumière réfléchie pour faire le retour jusqu'au lidar permettra à ce dernier de calculer la distance entre la voiture et l'objet. En Belgique, nous avons une expérience très concrète de cette technologie grâce à l'utilisation qui en est faite depuis 2013 pour détecter les excès de vitesse sur les routes wallonnes et bruxelloises.

| Sec. av. collision | Classification (source) | Projection de trajectoire |
|---|---|---|
| **5,6** | Véhicule (radar) | Le radar détecte un véhicule et en détermine la vitesse. Ce véhicule n'est cependant pas considéré comme étant sur le chemin de la Volvo. |
| **5,2** | Autre (lidar) | Le lidar détecte un objet inconnu, statique : ni parcours ni vitesse ne sont calculés. |
| **4,2** | Véhicule (lidar) | Le lidar reclassifie l'objet comme véhicule (statique) : perte de l'historique du parcours. |
| **3,9** | Véhicule (lidar) | L'autopilote projette une trajectoire du véhicule comme circulant le long de la voie à gauche de la Volvo. |
| **3,8 – 2,7** | Alternance entre « autre » et « véhicule » (lidar) | L'objet est considéré tantôt comme statique, tantôt comme circulant dans la voie à gauche de la Volvo. |
| **2,6** | Vélo (lidar) | Le lidar reclassifie l'objet comme vélo (statique) : perte de l'historique du parcours. |
| **2,5** | Vélo (lidar) | Le vélo est désormais considéré comme circulant le long de la voie à gauche de la Volvo. |
| **1,5** | Inconnu (lidar) | Le lidar détecte un objet inconnu, sans historique de parcours. Aucun objectif ne lui est assigné. Comme l'objet est partiellement sur le chemin de la Volvo, l'autopilote génère un plan pour le contourner. |
| **1,2** | Vélo (lidar) | Le lidar détecte un vélo qui est pleinement sur le chemin de la Volvo. Comme le plan de contournement n'est plus réalisable, une situation hasardeuse est détectée. La suppression d'action est initiée. |
| **0,2** | Vélo (lidar) | Fin de la suppression d'action. Comme la situation est toujours hasardeuse, le freinage progressif commence et la conductrice en est avertie. |

*Tableau 1 Historique des évènements enregistrés*[180]

Ce tableau appelle deux remarques. Premièrement, à aucun moment, la piétonne n'a été reconnue telle quelle, c'est-à-dire comme une piétonne, un vélo à la main. Le logiciel était tout simplement incapable d'une telle classification, étant donné qu'il ne lui était pas possible d'assigner cette catégorie à un objet qui ne se trouvait pas sur un trottoir ou sur un passage pour piétons :

> *According to Uber ATG, the SDS [le logiciel autopilote] did not have the capability to classify an object as a pedestrian unless that object was near a crosswalk. Since the pedestrian crossed in the middle of the street away from a sidewalk, the SDS initially classified her as an unknown object, then as a vehicle, then as a vehicle or a bicycle, and finally as a bicycle. Additionally,*

[180] Adapté d'après un tableau similaire dans NTSB, *Vehicle Automation Report*, pp. 10-11.

*because the SDS was unable to correctly classify the pedestrian, it was also unable to predict her path and speed on the roadway. Under both the vehicle and bicycle classifications, the SDS predicted that the object would stay in its travel lane, which was the lane to the left of the SDV [la voiture autonome].*[181]

Le rapport du NTSB ne mentionne pas que cette faille a été corrigée après la collision : un piéton « qui sort du cadre » ne sera donc toujours pas reconnu comme un piéton. Ce fait étonnant doit nous faire réfléchir sur le bien-fondé de la norme comme contrainte, comme nous l'avons vu dans la section consacrée à l'épreuve de la norme (§ 3.3.2.1) : pouvons-nous utiliser la norme comme règle constitutive de l'identité d'un agent ? Voilà un important problème que cet accident soulève et qui fait se détacher la face sombre de l'agentification : celle-ci, n'étant pas fondée sur des principes rigoureux mais *ad hoc*, en fonction du domaine applicatif, comporte toujours le risque de receler une part plus ou moins grande de *violence* : si autrui ne se conforme pas aux règles que nous avons fixées pour lui, il est pour ainsi dire jeté hors des murs de la cité, sans voie de retour possible.

Deuxième remarque, à chaque reclassification, l'autopilote « perdait la mémoire » : ainsi, de précieuses informations pour le calcul correct des trajectoires des objets détectés se sont effacées en cours de route. Selon Uber – et contrairement à la première faille discutée ci-dessus – cette faille-ci a été corrigée après l'accident : désormais, la génération des trajectoires possibles se base à la fois sur la catégorie de l'objet (et l'objectif correspondant) et sur toutes les positions antérieures connues. Nous devons cependant avouer que nous ne voyons pas très bien comment cela pourrait se faire. Précisions la source de notre perplexité : déjà à un niveau très bas, la voiture a échoué à discerner une unité d'agir, selon la terminologie de notre premier chapitre (§ 1.7.2.1) : comment assigner des positions antérieures à un objet si le logiciel ne peut établir une source de mouvement ? S'il échoue – et nous grossissons à peine le trait – à faire la distinction entre un arbre (catégorie « autre ») et un piéton qui a en main une canne à pêche (catégorie « piéton »), il pourrait désormais détecter un piéton en mouvement là où en réalité il n'y a qu'une rangée d'arbres le long de la route[182].

Quoi qu'il en soit, les failles du logiciel autopilote sont pour beaucoup dans l'accident de Tempe, d'autant plus regrettable qu'il aurait pu être évité. Soulignons que cette situation, dans laquelle une erreur logicielle est un facteur important, est encore nouvelle. Cela explique peut-être pourquoi les

---

[181] NTSB, *Human Performance Group Chairman's Factual Report*, p. 15.

[182] Soulignons que ce problème n'est en rien lié à la distinction entre procédures relevant de l'apprentissage automatique et celles recourant au raisonnement logique, comme le voudrait faire croire l'article signé par M. GIANCOLA, S. BRINGSJORD, N. S. GOVINDARAJULU et J. LICATO, *Adjudication of Symbolic & Connectionist Arguments in Autonomous-Driving AI* : dans cet article, les auteurs argument que l'accident de Tempe aurait pu être évité en recourant à des raisonneurs logiques multiples, dont la sémantique ne repose pas sur une distinction bivalente entre le vrai et le faux, mais porte la promesse de gérer des situations de grande incertitude en adoptant un système de non moins de 13 degrés de certitude, allant de -6 (certitude absolue en défaveur d'une proposition) à 6 (certitude absolue en faveur de la proposition), en passant par 0 (aucune croyance ni en faveur, ni en défaveur de la proposition). Si deux raisonneurs obtiennent des résultats différents, par exemple un raisonneur qui perd la mémoire à chaque recatégorisation, et l'autre non, un arbitre (*adjudicator*) est appelé à la rescousse pour trancher. Et les auteurs d'appuyer leurs thèses par la démonstration faite par un démonstrateur de théorèmes (*theorem prover*). À notre avis, c'est là méconnaître la nature du problème : comment « nourrir », à l'aide des données sensorielles brutes, le deuxième raisonneur, celui qui doit se débrouiller sans information catégorielle ?

rapports du NTSB – par ailleurs si consciencieux – font pourtant l'impasse sur les aspects proprement logiciels du problème. Après avoir lu tous les rapports dédiés à l'accident, nous n'en savons toujours pas très long sur l'implémentation de l'autopilote. De même, lorsque Volvo ou Uber disent avoir fait « des simulations » – qui pour démontrer la performance de son dispositif de freinage, qui pour démontrer l'efficacité de ses mesures correctrices entreprises après l'accident[183] – les enquêteurs ne se sont pas apparemment pas demandé comment ces simulations ont été faites. En guise de recommandation au NTSB, nous ne pouvons qu'appeler de nos vœux un aiguisement de leur regard critique pour ces aspects de l'accidentologie, qui gagneront sans doute plus d'importance encore dans les années à venir.

#### *3.4.3.4. L'éclairage de la SBA*

Il est tentant, pour analyser cet accident, et d'autres cas semblables, de recourir à un modèle de type « dilemme du tramway » (*trolley problem*), comme dans l'exemple suivant :

> *Suppose a large autonomous vehicle is going to crash […] and that it is on its way to hitting a minivan with five passengers head on. If it hits the minivan head on, it will kill all five passengers. However, the autonomous vehicle recognizes that since it is approaching an intersection, on the way to colliding with the minivan it can swerve in such a way that if first collides into a small roadster, thus lessening the impact on the minivan. This would spare the minivan's five passengers, but it would unfortunately kill the one person in the roadster. Should the autonomous vehicle be programmed to first crash into the roadster?*[184]

Or il est plus que douteux qu'un tel éclairage soit le meilleur point de départ pour méditer sur ce type de situations. La pierre d'achoppement, selon nous, c'est *l'hypothèse de l'information parfaite* qui sous-tend ce modèle : presque à la manière d'un dieu omniscient, la voiture – par quelque miracle – aurait la science de savoir combien de morts elle fera en virant à gauche, et combien elle en fera en virant à droite[185].

Une telle hypothèse est problématique. Tout d'abord, parce que si elle était vraie, nous voyons mal comment la voiture pourrait se retrouver dans des situations où une collision serait inévitable : il ne peut s'agir que de cas où « l'autre » serait manifestement en tort ; il faut en outre penser que ces cas seraient de toute façon trop rares pour que nous construisions nos raisonnements sur eux. Si, par extraordinaire, ce type de cas était sinon fréquent, du moins réaliste, l'hypothèse de l'information parfaite serait toujours malencontreuse : elle fait, en effet, passer à côté de ce qui fait *la nouveauté*

---

[183] Voir NTSB, *Highway Accident Report*, 2019, pour Volvo à la page 33 et pour Uber pp. 42-43 ; dans les deux cas, il est fait mention de « simulation » sans aucune autre précision : le lecteur critique des logiciels que nous sommes, reste donc sur sa faim.

[184] V. Bhargava et T. Wan Kim, *Autonomous vehicles and moral uncertainty*, dans P. Lin, R. Jenkins et K. Abney, *Robot Ethics 2.0*, p. 5.

[185] En relisant ce paragraphe, nous nous sommes rendu compte qu'il exite une deuxième raison – plus profonde – à l'insuffisance d'un tel « modèle », à savoir qu'il est discursif : de par sa mise en forme narrative, il incite son lecteur à accepter l'hypothèse de l'information parfaite. C'est un écueil que la SBA – n'étant pas discursive mais technique – peut éviter, à condition de ne pas nous laisser fasciner par l'accompagnement métaphorique qui la soutient.

*radicale* de ce type de situation et qui sollicite, avant toute autre considération, la réflexion éthique. En effet, la situation où il faut prendre une décision éthique en – disons – l'espace d'une seconde, est *absolument inédite* jusqu'à ce jour. L'être humain, dans ce type de situation, ne peut qu'esquisser un geste *réflexe*, avec plus ou moins de bonheur ; en l'occurrence, notre appareil cognitif n'a pas la vitesse requise pour élaborer un raisonnement approfondi.

Si donc la voiture était amenée à produire ce type de raisonnements, il s'agirait d'un élargissement remarquable du champ de la réflexion éthique, qui doit dès lors doit porter sur deux problèmes que l'hypothèse de l'information parfaite ne peut qu'escamoter : premièrement, dans quelle mesure est-il légitime qu'un logiciel s'érige en tribunal donnant la vie et la mort ? Deuxièmement, une fois traité le premier problème, quelles informations est-il opportun qu'une voiture collecte afin d'instruire son « procès » ? Citons ici l'exemple d'un article où il faut choisir de donner la mort à l'un des deux motocyclistes : l'un porte un casque, l'autre non[186]. Et l'auteur de se demander s'il faut foncer dans le motocycliste sans casque, qui mérite d'être « puni » en quelque sorte pour son irresponsabilité, ou dans le motocycliste avec casque, qui a plus de chance de s'en sortir vivant ? Ici encore, le dilemme du tramway nous joue des tours, nous faisant penser que l'exemple donne la liste exhaustive des attributs pertinents. Or il s'agit d'une question entièrement ouverte, et il n'y aucune limitation, en principe, à l'information sur laquelle la voiture – ou son logiciel – pourrait se baser : le motocycliste à gauche est peut-être un prix Nobel de la Paix ? Celui à droite n'a pas de casier judiciaire vierge, ou la voiture détecte le port d'une arme, qui après une consultation rapide du registre national s'avère être illégale ?

La SBA, au contraire, fait davantage justice à la réalité : de par sa nature stochastique, elle nous prémunit contre le mirage d'une fausse certitude : personne ne saurait dire qui est prédestiné à mourir dans une collision ; elle représente adéquatement la perception limitée qu'ont les agents – voitures et piétons confondus – de la réalité. Le « piéton » a la vue courte par rapport à la voiture munie d'un radar et d'un lidar, voit moins loin de nuit que de jour, voit encore moins s'il est sous l'influence d'un psychotrope, etc. Filons l'usage qui peut être fait ici d'une représentation multi-agents : nous sommes en présence de deux agents usagers de la route, qui doivent se partager un même *environnement* : qu'en est-il de l'état de la route ? Des passages pour piétons ? Des conditions climatiques ?

Rappelons-nous, pour finir, les représentations intra-agents : si appel est fait à un module BDI, il est également possible de rendre compte de valeurs et croyances différenciées par agent : tel agent *désire* avec la plus haute priorité se rendre à sa destination, tel autre préfère s'y rendre plus prudemment, tel autre encore le désire avec autant d'ardeur, mais *croit* qu'il y parviendra mieux en train, eu égard aux embouteillages dont il a entendu parler à la radio. La SBA, ainsi, non seulement permet d'accueillir certaines questions primordiales, mais aussi d'étendre le questionnement à la *finalité* même de la route, à savoir la *mobilité*[187]. La SBA nous invite donc à changer de regard, étant

[186] J. MILLAR, *Ethics settings for autonomous vehicles*, dans P. LIN, R. JENKINS et K. ABNEY, *Robot Ethics 2.0*, p. 21.

[187] Dans un article dont nous avons pris connaissance tardivement (U. LOTZMANN et M. MÖHRING, *Simulating Normative Behavior and Norm Formation Processes*), la question des accidents de la route est ainsi abordée sous l'angle de la mobilité : les agents s'y voient assignés des lieux à atteindre dans un certain délai, faute de quoi ils sont pénalisés. Les agents peuvent être en voiture ou à pied ; dans le dernier cas, ils peuvent choisir de traverser la route immédiatement ou de faire le détour par un passage pour piétons ; dans le premier, le choix porte sur ce qu'il convient de faire lorsqu'on

donné qu'elle permet à l'automobiliste de devenir piéton, ou usager des transports en commun[188] : le *choix des agents*, insistons une fois de plus sur ce point, n'est pas neutre mais dépend des questions que le modélisateur souhaite poser à son modèle.

Si un robot ne fonctionne pas bien sur nos routes, il n'y a là aucune raison de nous enorgueillir, pas plus d'ailleurs de la difficile cohabitation, à l'heure actuelle, des automobilistes et des piétons dans ce même environnement qu'ils sont pourtant condamnés à partager. Ne faut-il pas plutôt y voir une faiblesse de nos routes (le milieu associé de la voiture), qu'on a laissé dégénérer en jungle ? Pour que la technologie de la voiture autonome puisse devenir vraiment intéressante, il faudra sans doute repenser radicalement la mobilité, le réseau routier en premier lieu. Nous aborderons brièvement cette piste dans le paragraphe suivant.

#### *3.4.3.5. Pour une innovation responsable*

Nous écrivons ces lignes deux ans après la publication du rapport final du NTSB. Entre-temps, le 27 août 2020 très exactement, Rafael(a) Vasquez a été inculpé(e) pour l'homicide involontaire d'Elaine Marie Herzberg. Il ne nous appartient pas de commenter cette inculpation ; notre problème n'est pas de nature juridique. Or le problème éthique que pose la voiture autonome ne se cantonne nullement aux accidents qu'elle peut causer, sur la responsabilité à imputer afin de déterminer, en gros, qui va payer les pots cassés. Comme nous l'avons vu au premier chapitre (§ 1.7.2.6), la responsabilité est une notion qui regarde autant vers l'avenir que vers le passé : rétrospectivement, elle se fonde sur une imputation causale pour attribuer la paternité d'un effet constaté au présent ; prospectivement, elle projette nos actions sur l'horizon des valeurs possibles aujourd'hui et demain.

C'est pourquoi nous ne voudrions pas clore cette section dédiée à l'usage de la voiture autonome sans dire un mot sur l'horizon éthique plus large dans lequel celle-ci s'inscrit. En effet, pour qu'une innovation comme la voiture autonome soit éthique, il ne suffit pas que les tests se déroulent sans heurts. L'évaluation d'une innovation technique requiert que soient problématisées quatre questions[189], dont celle du *comment* : très prégnant en éthique biomédicale (consentement éclairé, etc.), cet aspect revient, en l'occurrence, à s'assurer que la voiture autonome ne soit pas

---

rencontre un piéton sur la route : accélérer, ralentir, ou même s'arrêter. Des collisions sur des passages pour piétons sont très pénalisantes pour l'automobiliste, qui fait donc plus attention à ces endroits, ce qui stimule en retour les piétons de davantage recourir aux passages. Pour simpliste que soit l'approche, soulignons que nous ne sommes pas ici en présence d'un mécanisme d'optimisation du trafic de façon à minimiser les collisions, mais qu'il s'agit d'une simulation où il y a un réel apprentissage des normes, basé sur la plate-forme EMIL que nous avons déjà rencontrée en abordant la motivation intrapersonnelle à l'égard de la norme (§ 3.3.1).

[188] Par exemple, ce cas d'utilisation est pris en charge par la plate-forme de simulation GAMA par sa fonctionnalité de classes d'agents « imbriqués », permettant de représenter une même « entité » comme différents types d'agents : « The multi-level architecture is often used in order to represent an entity through different types of agent. For example, an agent "bee" can have a behavior when it is alone, but when the agent is near from a lot of agents, he can changes his type to "bee_in_swarm", defined as a micro-species agent of a macro-species "swarm" agent. Another example: an agent "pedestrian" can have a certain behavior when walking on the street, and then change his type to "pedestrian_in_building" when he is in a macro-agent "building". » (source : https://gama-platform.github.io/wiki/MultiLevelArchitecture).

[189] Nous nous basons ici sur la contribution de B. STAHL, Gr. EDEN et M. JIROTKA, *Responsible Research and Innovation in Information and Communication Technology*, dans R. OWEN, J. BESSANT et M. HEINTZ, *Responsible Innovation*, pp. 202-213.

commercialisée avant d'être suffisamment éprouvée, grâce à des tests régis par des protocoles de sécurité exigeants. Une deuxième question concerne le *quoi* de l'innovation : de quoi s'agit-il exactement ? Dans le cas d'une voiture autonome, il s'agit alors avant tout de bien définir le degré d'autonomie (comme ça a été fait dans la norme SAE cité dans le paragraphe 3.4.3.2). Une troisième question est celle de savoir *qui* innove ? Il faut problématiser les intentions, les intérêts et les enjeux des chercheurs/concepteurs, des bailleurs de fonds et des entités qui déploient (ou commercialisent) le produit.

La dernière question est peut-être la plus importante, c'est celle du *pourquoi* de l'innovation, de sa *finalité* : le but de la recherche est-il justifiable ? De quel *projet de société* participe-t-elle ? Alors que cet aspect va pour ainsi dire de soi en éthique biomédicale (qui constitue encore la référence en matière éthique), il devient autrement plus problématique dans le cas de l'innovation technique qui nous préoccupe ici. La question est pourtant essentielle, se trouve même à la base même de la notion d'innovation responsable : que voulons-nous que fassent la science et l'innovation[190] ? Or pour prendre cette question au sérieux nous devons nous poser la question de l'avenir : nous voyons ainsi réaffirmée l'importance de *l'anticipation*, exercice au cœur tant de la SF que du développement durable. Anticiper l'avenir relève cependant d'une attitude qui requiert de dépasser l'antagonisme entre *valeur économique* d'une part, et évaluation étroite en termes de *risques et règlementations*, d'autre part[191]. Réfléchir sur un projet de société est un exercice qui, tout en dépassant l'interrogation éthique, y fait tout de même appel, en tant que la réflexion porte sur la justice, comme nous l'avons vu avec Ricœur (§ 3.1). Tant que nos sociétés ne seront pas capables de telles réflexions, il faut craindre que l'innovation reste marquée du sceau de « l'irresponsabilité organisée »[192], où le seul intérêt sérieusement pris en compte soit celui des investisseurs[193].

L'évaluation du projet de mobilité dans lequel pourrait s'inscrire la voiture autonome dépasse de très loin les limites de ce mémoire, certes. Qui plus est, définir les effets d'une technologie est une tâche ardue, à laquelle ne suffisent ni la bonne volonté, ni le simple exercice de la pensée critique[194]. Ces réserves étant faites, nous pouvons cependant déjà soulever quelques points d'interrogation quant aux bénéfices attendus de la voiture autonome. Parmi ceux-ci, nous trouvons cités notamment la sécurité – des routes plus sûres – la réduction de la pollution sous la forme d'une diminution d'émissions de gaz à effet de serre, enfin un renforcement de l'autonomie des personnes à mobilité réduite[195]. Sans être expert en la matière, il semble évident que les personnes à mobilité réduite

---

[190] L'importance de la finalité de l'innovation a été fortement soulignée par R. OWEN, J. BESSANT et M. HEINTZ dans leur préambule (*Preface*) à l'ouvrage collectif *Responsible Innovation*, p. xx.

[191] R. OWEN, J. STILGOE, Ph. MACNAGHTEN ET A., *A Framework for Responsible Innovation*, dans R. OWEN, J. BESSANT et M. HEINTZ, *op. cit.*, p. 29.

[192] Jugement rapporté par J. STILGOE dans son avant-propos (*Foreword*) à *Responsible Innovation*, p. xii.

[193] Malheureusement, même dans des ouvrages consacrés aux questions éthiques soulevées par l'innovation technique, le centre d'intérêt principal n'est que trop souvent cet amalgame entre éthique d'une part et responsabilité juridique et compensation financière des conséquences non-voulues, d'autre part. Citons – à seul titre d'illustration – le propos suivant : « If courts and legislatures do not adequately resolve the compensation issue [for harm caused by robots], robot producers may incur inexpected and excessive costs, which would disincentivize investment » (J. K. GURNEY, *Imputing Driverhood*, dans P. LIN, R. JENKINS et K. ABNEY, *op. cit.*, p. 51).

[194] Cf. R. VON SCHOMBERG, *A Vision of Responsible Research and Innovation*, dans R. OWEN, J. BESSANT et M. HEINTZ, *op. cit.*, pp. 54-57.

[195] J. K. GURNEY, *Imputing Driverhood*, dans P. LIN, R. JENKINS et K. ABNEY, *op. cit.*, p. 51.

peuvent bénéficier de toute une série d'aménagements mineurs qui ne nécessitent pas une voiture autonome[196]. De surcroît, vu le prix de ce type de véhicule, il semble prudent d'avancer que ces personnes pourraient également bénéficier des services d'un simple taxi. En revanche, la voiture autonome risque d'augmenter le nombre de véhicules sur la route : toute personne n'ayant aujourd'hui pas de permis ou ne se sentant pas en sécurité (les jeunes, certaines catégories de personnes âgées…) pourrait à l'avenir se laisser conduire. Si cette hypothèse devait se réaliser, elle s'opposerait frontalement à l'objectif de la diminution de la pollution, d'autant plus que les coûts de fabrication d'une telle voiture – notamment le besoin en matières premières de toutes les composantes électroniques sophistiquées – grèvent déjà ce bénéfice très hypothétique. Quant à la sécurité routière, dans le paragraphe précédent nous avons déjà émis l'hypothèse que l'amélioration de celle-ci ne constitue probablement pas un problème que pourrait résoudre une innovation technique, mais devra faire l'objet d'une refonte plutôt ambitieuse de notre manière de concevoir la mobilité, ainsi que la reconfiguration en profondeur du réseau routier.

Il n'est dès lors pas sûr que la voiture autonome soit autre chose qu'un gadget. Ou plutôt, autre chose qu'une *fuite en avant* technologique, dont a si éloquemment parlé Jacques Ellul : face à un problème induit par la technologie, notre société ne semble pas vraiment capable de répondre autrement que par un surcroît de la même technologie, sans se poser la question des fins des innovations techniques. Face aux problèmes de mobilité, de routes congestionnées, de pollution créée par l'usage immodéré de la voiture, une telle analyse semble pourtant plus que jamais d'actualité.

[196] Un relecteur nous a cependant fait observer que la voiture autonome pourrait sensiblement renforcement l'autonomie de certaines personnes présentant un handicap physique sévère.

# Conclusion

Conclure un travail dont la gestation a duré non loin de trois ans n'est pas chose facile. Il le faut pourtant, fournir un dernier effort afin de ressaisir les lignes de force qui font converger la matière – protéiforme – de l'ensemble des chapitres sur un horizon commun.

Nous avons vu l'éthique des machines s'émanciper de l'éthique de l'informatique traditionnelle. Tandis que cette dernière est avant tout un domaine d'application de l'éthique appliquée, la première a des ambitions plus grandes. Certes, son souci principal est d'ordre pratique : offrir des garanties de fiabilité et de sécurité alors que le champ d'intervention de la machine s'étend comme une tache d'huile. Ceci n'empêche cependant nullement l'éthique des machines de se concevoir également comme une *contribution originale* à l'éthique.

Nous avons vu comment l'éthique des machines se veut une interrogation *fonctionnelle* de l'éthique : l'éthique est ainsi appréhendée à travers les effets que nos comportements produisent dans le réel. L'idée d'équivalence fonctionnelle vient ici en appui de l'ambition de l'éthique des machines à éclairer l'éthique, ambition qui repose sur la *calculabilité* du comportement. L'idée de la calculabilité du comportement éthique est en effet décisive pour comprendre la nouvelle discipline. Ainsi l'hypothèse de la calculabilité favorise-t-elle l'approche des représentations explicites de l'éthique et ce, contrairement au courant majoritaire en robotique qui privilégie les solutions précâblées. La calculabilité du comportement encourage également l'idée d'une procédure de calcul indépendante par rapport au contexte du comportement et, par là, privilégie des vues descendantes sur l'éthique.

Rappelons encore que le primat de la calculabilité du comportement est en principe neutre par rapport à la question de savoir quel type de comportement doit être privilégié comme objet d'étude. Aussi deux tendances se dessinent-elles à l'intérieur de l'EM : une tendance prescriptive et une autre, descriptive. Avec les projets du conseiller éthique et de MoralDM, nous avons vu un bel échantillon de chacune. Le conseiller éthique formalise et généralise les règles jugées idéales par son utilisateur au moyen d'un moteur de logique inductive, alors que MoralDM cherche à implémenter l'équivalent fonctionnel d'un moteur de raisonnement éthique humain, mettant en œuvre trois type de raisonnements : déontologique, utilitariste, casuistique. Quoique d'inspiration et d'implémentation sensiblement différentes, les deux entreprises butent sur la même difficulté qui est celle de comparer des options éthiques qualitativement différentes ; en d'autres termes, le problème abordé au troisième chapitre – qui est celui de la téléologie – est ainsi pressenti avec acuité.

Par ailleurs, le premier chapitre peut se lire comme un dépouillement progressif : suivant en ceci l'intuition fonctionnaliste, nous avons cherché à trouver une définition éthique minimale du comportement humain. Or nous avons vu que, pour garder à une éthique, même définie en termes fonctionnels minimaux, un sens, la notion de *justification d'un comportement* doit faire partie intégrante des « livrables » de l'éthique des machines. Tout comme le comportement auquel elle

préside, la justification doit satisfaire l'exigence de calculabilité[1]. De fait, l'importance de la justification est souvent mise en relief, bien qu'elle soit rarement thématisée : nous avons cherché à combler cette lacune par un emprunt – illustratif – à la sociologie : les justifications y sont portées par un nombre limité de valeurs, soit autant de façons de voir le monde. Si ces valeurs sont en général partagées entre interactants, leur application à telle ou telle situation peut faire l'objet de débats, de négociation. Aussi le recours à Thévènot et Boltanski a-t-il permis de clarifier notre propre concept de valeur : celle-ci permet de subsumer la situation concrète sous un cadre plus général, cadre qui est en même temps une clef, puisque loin de cadenasser la réalité, en ouvrant une certaine porte, il permet de la voir d'une certaine façon. La valeur se manifeste ainsi non seulement comme un principe que nous pourrions qualifier d'herméneutique, comme un certain éclairage permettant de mieux appréhender l'un ou l'autre aspect de notre vivre-ensemble ; non, dans la mesure où tel éclairage est nécessaire à notre compréhension du monde, la valeur se voit par là même dotée d'une vertu structurante de notre vécu : elle fait « objectivement » partie de notre subjectivité, s'il nous est permis de recourir à ces notions usées d'avoir trop servi.

Ce travail, qui appartient à l'horizon disciplinaire de la sociologie, parle essentiellement de la justification dans la vie de tous les jours. C'est pourquoi sa transposition dans le cadre de l'éthique a été un geste cohérent avec la définition initiale que nous avions de cette dernière, car l'un et l'autre sont de nature topique. L'exigence topique ne va pas, cependant, sans soulever des difficultés, voire un paradoxe : par définition, une justification doit être interprétable par un être humain ; en termes informatiques, cela revient à exiger une approche logico-symbolique de la justification. Or un mouvement topique requiert plutôt une approche de type apprentissage automatique, probabiliste. De cette double exigence – logique d'un côté, probabiliste de l'autre – il résulte une tension que l'informatique a beaucoup de peine à résorber. L'éthique des machines, avec son intérêt marqué pour les conditions de possibilité cognitives d'une vie éthique dans l'être humain, se doit alors de recourir à des architectures d'une vertigineuse complexité pour accommoder une telle exigence : rappelons-nous l'architecture LIDA, mettant en œuvre un mixte entre intelligence symbolique et mécanisme attentionnel probabiliste.

Le recours à la justification de Boltanski et Thèvenot nous a aussi amené à nous interroger sur la portée du questionnement de la valeur en éthique des machines. La question de la valeur, si elle n'est pas absente, n'est pourtant jamais présentée comme une question de choix : l'éthicien « sait » quelles valeurs appliquer. Le plus souvent d'ailleurs, l'éthique des machines conçoit essentiellement l'éthique comme un conformisme social (le cas échéant, à une société idéalisée) d'agents anthropomorphes. Un tel biais provient sans doute du parti pris de la plupart des auteurs, qui s'inscrivent le plus souvent d'emblée dans une apologie de l'entreprise robotique.

C'est à ce propos que nous avons cru bon d'ouvrir le débat en rappelant que l'éthique n'est pas une affaire de calcul de l'individu – savant ou non – dans sa tour d'ivoire : tout se renégocie, les règles,

[1] Ce mémoire était déjà terminé lorsque nous avons eu connaissance d'une idée de Jean Nabert, selon qui *l'injustifiable* est à la racine du Mal (idée rapportée par Éric BLONDEL, *Le problème moral*, p. 178). La calculabilité de la justification revêtirait, dans une telle perspective, une importance décisive. En effet, si l'absence de possibilité de justification est une figure du Mal, deux questions s'imposent à nous : celle d'abord de la *validité* du calcul – une justification, par le fait même d'être calculable, évite-t-elle le Mal ? – ; ensuite la question de la *complétude* du calcul : toute justification valide peut-elle être calculée ?

les jugements et les valeurs, jusques et y compris la valeur première, l'homme et son image. À l'intérieur de la négociation, la règle cède souvent le pas au jugement : une règle éthique est toujours de second ordre, (rappelons-nous l'exemple, « voler est mauvais ») et reconnaître l'applicabilité de telle ou telle règle fait partie de la négociation, de la création de sens. Loin d'être l'antichambre du relativisme, le rappel de la négociation fait écho à la pluralité des manières d'habiter le monde. C'est dire qu'essayer de comprendre en quoi l'éthique des machines peut enrichir la réflexion éthique soulève au préalable une question plus fondamentale : que peut enseigner l'entreprise informatique sur l'homme ? Formuler une réponse à la première question présuppose une réponse à la deuxième. Plus fondamentalement encore, c'est le sens de l'informatique elle-même qu'il faudrait d'abord interroger.

Nous avons ouvert le deuxième chapitre par une présentation du paradigme multi-agents. Tout au long de cette présentation, nous nous sommes efforcé de prendre au sérieux l'unité du paradigme, quand bien même celui-ci paraît de prime abord éclaté, applicable à des technologies si éloignées les unes des autres que l'emploi d'une seule notion pour les regrouper toutes ne paraît pas d'emblée évident. Ces différentes technologies ont chacune leur histoire, plus ou moins complexes : alors que dans le cas de la robotique et de la programmation orientée agents, les sources et les dettes se laissent aisément retracer, l'histoire de la simulation à base d'agents est plus embrouillée, faite d'emprunts divers : à la théorie des automates cellulaires, aux systèmes experts, aux métaheuristiques, etc. Aussi avons-nous préféré éviter de trop nous engager sur la chronique de cette émanation du paradigme, chronique qui mériterait à elle seule de faire l'objet d'un mémoire entier.

Toujours est-il que le seul relevé des technologies se réclamant du paradigme nous a permis d'en éprouver les concepts fondamentaux : l'agent, l'environnement, le système, le temps, l'espace et la contingence. La première technologie étudiée a été le BDI, technologie qui permet d'implémenter des comportements complexes en dotant ses agents d'un répertoire de croyances, d'intentions et de désirs. Nous avons vu les ramifications de cette technologie, ainsi un mécanisme motivationnel qui permet de générer des intentions, inaugurant par là un début réel d'autonomie à l'agent. Le BDI peut ainsi se faire l'écho des préoccupations individuelles de l'éthique des machines.

Pourtant, la formalisation que donne le paradigme multi-agents de la notion d'agent a ceci de particulier qu'elle ne définit pas l'agent pour lui-même : il est toujours déjà aux prises avec son environnement. S'il est donc nécessairement source d'effets dans le réel, une unité d'agir, nous avons aussi insisté sur la dissociation toujours possible entre potentialité d'action et identité. La dissociation est importante, car nous avons vu par ailleurs l'importance de l'identité dans la constitution de l'individu éthique. N'avons-nous pas relevé que l'intérêt éthique d'un robot pourrait résider dans son manque d'identité, de son altérité radicale à l'égard de l'homme ?

C'est dire que dans le paradigme multi-agents, le « multi » doit être pris au sérieux : l'agent individuel n'a d'intérêt que dans le jeu d'interactions, dans l'échange d'informations, qu'il établit avec ses semblables, en vue de produire des effets dans le « réel », que celui-ci soit physique ou informationnel. La décentralisation de l'intelligence, du flux de contrôle, en SMA est en effet apparue comme un nœud névralgique : consentie à regret, les tentatives de parer l'atomisation ne manquent

pas. Le renoncement à l'intelligence centrale implique des remaniements importants dans la façon de voir une architecture logicielle, qui désormais s'accompagne d'un nouvel ordre symbolique (FIPA ACL), de nouvelles interfaces pour s'ouvrir au monde (l'usage d'artéfacts et d'institutions)… Ces tentatives trouvent leur point culminant dans la SBA où, en vertu des héritages multiples, la réification de l'environnement finit le plus souvent par réintroduire une indéniable intelligence centrale, lui offrir un lieu où elle pourra survivre, à l'intérieur même du paradigme. Même si une telle intelligence centrale ne se confond nullement avec une intelligence collective – de même que l'environnement en SBA se distingue de l'ensemble des agents qui se meuvent à sa surface – elle peut recevoir une interprétation matérialiste ; n'est-ce pas ce que vise Simon lorsqu'il attribue la complexité du tracé de la fourmi à la morphologie sinueuse du terrain que celle-ci doit parcourir ? Dans cette perspective, il peut s'avérer intéressant de creuser si, dans les travaux en SMA, les limites de cette base matérialiste sont toujours respectées : tous les effets attribués à l'intelligence centrale peuvent-ils l'être, « en réalité », au milieu physique des agents ?

L'environnement se montre de fait un facteur différenciateur entre les différentes approches en SMA. Comme nous l'avons dit, l'environnement devient même primordial dans le cas de la SBA. Mettant fortement l'accent sur la formalisation de l'espace, la simulation à base d'agents implique une autre façon d'appréhender le monde, où la mise en situation est préférée à l'abstraction disciplinaire. Sa grande capacité d'intégration de formalismes et la place faite à la contingence la font alors parfois prendre des allures de boîte noire. L'épaisseur ainsi gagnée la fait apparaître comme une empirie de second ordre, qui peut à son tour être pris comme objet d'étude et d'observation. Cette caractéristique de la SBA fait qu'elle est le plus souvent considérée comme une « quasi-expérience ». Or nous avons suggéré, en nous appuyant sur la réflexion d'Isabelle Stengers sur les sciences du terrain, que le paradigme expérimental n'est pas la seule façon d'éclairer la rationalité propre de la simulation à base d'agents. En effet, tout aussi pertinent pour comprendre cette dernière, se trouve être le paradigme *historique*, qui vise la compréhension d'un devenir toujours singulier en lui adressant des questions, en lui appliquant des méthodes inspirées par les autres disciplines scientifiques qui, elles, peuvent vouloir viser l'universel.

Plus un système devient opaque, se montre capable de nous surprendre, plus sa justification risque de nous échapper, d'où le problème de l'exigence de la justification. Or la SBA s'épanouit dans une vue systémique du monde, où la réalité est vue comme une imbrication de sous-systèmes. La validité de la connaissance qu'elle produit est alors affaire du rattachement au « bon niveau » de description. Nous avons vu, en effet, que chaque système un tant soit peu complexe se compose de plusieurs couches « cognitives », si nous pouvons risquer cette métaphore mentaliste, et qu'une justification du comportement visible de l'extérieur n'aura pas à prendre en compte toutes les couches de la même manière, même si toutes les couches, jusqu'à la plus primitive, contribuent au résultat comportemental visé.

Il est vrai que pour celui qui étudie la cognition humaine, le BDI ne peut être que simulacre. De même, l'observateur scrupuleux des institutions humaines ne saurait voir qu'artifice dans les institutions virtuelles des systèmes multi-agents. Or celui qui s'intéresse à un comportement en société sera ravi de découvrir – dans ce même BDI – une justification individuelle ; et pour celui qui s'intéresse à la justice, les effets produits par certaines institutions plutôt que telles autres seront une source très

riche de réflexion. L'accusation toujours prompte à surgir du simulacre nous a amené à nous interroger sur la portée de la métaphore dans la connaissance produite par la simulation à base d'agents : qu'est-ce qu'un dispositif si hautement technique, technicisé, porteur d'une volonté de contrôle si l'on veut, peut nous apprendre – sur une couche métaphorique suffisamment épaisse pour ne pas être transparente – sur le réel ?

Nous avons vu que démarche fonctionnelle et démarche métaphorique ne s'excluent pas. Disons d'emblée que le terme de « métaphore », tout au long du présent travail, a été utilisé dans un sens assez lâche (le terme plus général d'analogie eût peut-être mieux convenu). De fait, nous avons rencontré la métaphore à différents niveaux, nous l'avons vu s'appliquer à différentes sphères de l'activité humaine. Ainsi le paradigme multi-agents métaphorise-t-il l'espace – support d'intégration d'autres savoirs –, le temps (temps du simulat, temps de la simulation), les états mentaux de l'agent et jusqu'à la métaphorisation de l'ordre symbolique lui-même au travers des ontologies et du protocole de communication FIPA ACL.

Loin d'être un défaut, l'usage de la métaphore est précisément ce qui permet aux SMA de briller comme support de justification, d'expliquer les comportements dans des termes valables dans le domaine de référence. La métaphore, en effet, se pose comme condition d'intelligibilité, situation que nous retrouvons certes en informatique mais qui n'est pas étrangère à la science. Or la métaphore n'est pas seule à conférer à la SBA sa faculté de signifier : l'intégration spatiale y contribue tout autant. Nous pouvons donc dire que la métaphore de l'agent et l'intégration spatiale des connaissances permettent à la SBA de s'exprimer dans le domaine de connaissance visé. Ce point est essentiel à l'élaboration de justifications adéquates. La possibilité de formalisation de la SBA ne doit pas se substituer à la métaphore, elles contribuent toutes deux à l'intelligibilité de la réalité qu'elles décrivent.

Nous avons ouvert le troisième chapitre par la présentation des liens entre éthique et morale de Paul Ricœur. Faisant dialoguer Aristote et Kant, cet auteur réactualise la téléologie tout en tirant profit de l'universalité du formalisme normatif. Il va cependant plus loin encore, en cela que son traitement de l'interdépendance entre téléologie et déontologie se présente, pour ainsi dire, en double triptyque, car il développe patiemment l'interdépendance des deux pôles à trois niveaux : intrapersonnel, interpersonnel, impersonnel.

Dans un premier temps, nous nous sommes attaché à suivre ce mouvement. Ainsi avons-nous vu, au niveau intrapersonnel, les travaux en SMA d'inspiration téléologique se pencher sur la question de la motivation à l'origine du comportement éthique. Ensuite, nous avons abordé le niveau interpersonnel en examinant la formalisation du jugement éthique : un agent peut juger son propre comportement, juger de l'adéquation du comportement d'un autre agent par rapport à sa propre éthique, évaluer l'adéquation de cet autre agent par rapport à une éthique qu'il a publiquement déclarée, etc. Le jugement peut non seulement porter sur l'adéquation entre une éthique et un comportement, mais aussi sur la compatibilité de deux éthiques entre elles. De ce jeu entre différents jugements peut surgir une réflexion sur la réputation et la confiance, sur un ethos projeté et un pathos interpersonnel, qui tous deux permettent de fonder l'échange et l'entraide. Toujours sur le plan interpersonnel, nous avons vu à l'œuvre la négociation argumentative : sur fond d'une vision du

monde partagée entre agents, celle-ci permet de coordonner l'action en fonction de certains buts. L'approche nous a semblé prometteuse pour aborder la négociation de la valeur, ainsi que la prise en compte de la pluralité qui règne en ce domaine. Finalement, sur le plan impersonnel, nous avons vu des structures institutionnelles, des coalitions, se former et se défaire au gré de préférences éthiques.

Ce que tous ces travaux ont en commun, malgré la spécificité de chaque approche, ayant toutes leurs forces et faiblesses, c'est de placer la valeur et la finalité au cœur même de l'activité qu'ils cherchent à modéliser, plutôt que de les considérer comme des constructions adventices : elles y sont le moteur de l'action, principe même d'une vie qui sans elles s'essoufflerait, faute de sens. Nous avons également pu observer que la SBA n'oblige pas à présupposer la valeur : la modélisation peut l'engager, l'intégrer dans la représentation. Cependant, même ici, la question du rôle de la valeur dans la justification est en général présupposée plutôt qu'assumée. Nous en avons vu d'ailleurs un exemple frappant dans le cas du processus qui aboutit à un jugement éthique : l'importance de la reconnaissance des situations, cette fonction qui génère les états mentaux correspondant à une situation donnée, y est certes stipulée, mais la fonction elle-même n'est pas implémentée. La valeur, là encore, est donc toujours déjà donnée.

Nous avons parcouru ensuite les travaux consacrés au pôle déontologique de l'éthique, ceux qui étudient l'épreuve de la norme et les conditions de son émergence et de son efficacité. Ici encore, nous avons remonté les trois niveaux en commençant par le niveau intrapersonnel : la norme peut être d'autant plus efficace que l'agent l'a intimement internalisée. Nous avons vu ensuite les différents modes sur lesquels la norme peut exiger des comportements de la part des agents qui se situent sous son influence. Nous y avons retrouvé la norme comme contrainte, au sens informatique de ce terme, c'est-à-dire constitutive du comportement à exhiber. Nous avons également vu la norme comme contrat, où l'agent s'expose à des pénalités en cas de rupture, et comme indication d'ordonnancement, où la conformité à la norme devient comme un optimum à calibrer.

À côté de cet intérêt pour l'efficacité de la norme comme épreuve, nous avons également étudié l'origine de la norme, dans l'espoir de pouvoir y faire le lien avec la valeur dont elle procède. Nous y avons vu le pouvoir de l'imitation, l'apprentissage cognitif et le contrôle social dans la promotion des normes, or nulle part la valeur n'est ici apparue telle quelle. Autant dire que la conformité à la norme y a été érigée en but à poursuivre pour lui-même. Cette absolutisation de la norme se comprend cependant dès lors que nous la replaçons dans son contexte : l'ingénieur, ayant renoncé à regret à un flux de contrôle central, cherche à conjurer l'anarchie. L'intérêt des travaux sur la norme se situe peut-être même dans leur radicalité : en poussant à bout l'approche déontologique, ils sont aussi appelés à en explorer les limites.

Sur le plan impersonnel, nous avons vu l'importance de la vie des institutions pour structurer celle des individus : les rôles et les responsabilités liées à certaines charges s'y sont révélées comme des sources de nécessité intarissables. De façon surprenante peut-être, le paradigme multi-agents se montre tout autant sinon plus prometteur sur le plan impersonnel que sur le plan interpersonnel. En effet, c'est surtout au niveau impersonnel – rôles, institutions… – que la SBA pourrait rendre des services, ce niveau parfois négligé en éthique, où la tentation de se cantonner au niveau

(inter)personnel n'est jamais absente ; cette tentation – nous l'avons vu au premier chapitre – est d'ailleurs assez prononcée dans les présuppositions de l'éthique des machines elle-même. Aussi le paradigme multi-agents a-t-il des atouts majeurs pour porter l'idée de la calculabilité dans des sphères qui, sous cet angle au moins, n'ont été jusqu'ici que peu explorées.

Ainsi, le paradigme multi-agents aide à penser, à replacer les individus dans leur société, sans laquelle ils ne sauraient vivre. Il semble suggérer qu'un individu n'existe que grâce aux faisceaux d'échanges dont il participe. Si le créneau spécifique du paradigme se trouve probablement à la jonction des niveaux interpersonnel et impersonnel, il n'est pas sans avoir un atout sur le niveau personnel également, si l'on veut bien considérer le BDI comme une approche mentaliste fonctionnaliste, où les états déclaratifs sont capables de décrire le comportement qu'ils font émerger dans le jeu de leurs interactions avec le monde.

Nous avons terminé le troisième chapitre en passant en revue quelques cas sinon pratiques, du moins aux prises avec le réel, effectif ou, pour ce qui est du premier cas, le réel à venir, possible. Ce premier cas, en effet, s'intéressait de près aux lumières que peut apporter la science-fiction à la réflexion éthique. Nous y avons donné une lecture – certes sommaire – de deux œuvres qui, chacune à sa manière, ont apporté un éclairage sur un thème majeur du paradigme multi-agents, à savoir le renoncement à l'intelligence centrale. Ce thème a été introduit, presque à notre corps défendant, dès l'introduction, au travers de l'interrogation maeterlinckienne de la société ingénieuse organisée en fourmilière. Dans la nouvelle de Marcel Thiry, dans la mise en scène des Secs réunis autour du Vase, nous avons entrevu une figure, éphémère tout autant qu'étrange, d'un nouveau mode de vie rendu possible grâce au sacrifice partiel de l'individualité. Dans le projet « Fils de l'Homme », décrit dans le roman de Gilbert Hottois, nous avons assisté – non sans effroi – à la collecte cynique de matière cérébrale, réassemblée à des composantes cybernétiques comme s'il s'agissait de pièces détachées interchangeables. La question de l'intelligence centrale peut ainsi être conçue comme un degré de liberté – pour le meilleur ou pour le pire ; bref comme un problème éthique, engageant l'homme et son image.

La science-fiction, toutefois, n'est pas intéressante seulement en ce qu'elle apporte une thématique qui lui est propre. Elle nous interpelle également en vertu de sa méthode, l'anticipation, c'est-à-dire l'exploration de l'espace des possibles afin d'en éprouver les implications, la cohérence interne, les effets sur les modes de vie. Nous avons pris la mesure de l'importance de cet exercice, dont nous avons vu avec Hottois (mais la philosophie d'un Hans Jonas, son insistance sur une éthique du futur, aurait également pu être appelée à la barre) la pertinence éthique.

La SBA, elle aussi, peut être utilisée à des fins d'anticipation, mais elle s'en distingue tout de même nettement en ceci qu'elle ne perd jamais de vue le calcul de ce qui est *probable*. Quand les probabilités dominent l'exercice, la SBA fait œuvre de prévision, non d'anticipation. La frontière entre l'une et l'autre mériterait certainement de plus amples réflexions. Nous nous sommes contenté d'en donner deux illustrations dans le deuxième cas pratique. Avec COMOKIT, nous avons vu se déployer un modèle multi-agents destiné à prévoir des évolutions possibles de la pandémie Covid-19 à une échelle locale. Avec la modélisation d'accompagnement, nous avons vu s'appliquer, dans des situations concrètes, une représentation multi-agents comme grille de lecture dans des conflits du

vivre-ensemble, comme un conflit sur l'usage des ressources d'eau douce sur l'atoll de Tarawa. De COMOKIT à la modélisation d'accompagnement, la distribution du possible et du probable s'équilibre diversement. De façon tout aussi significative, le rapport à la vérité, la validité du savoir obtenu grâce à la modélisation, se construit sur des bases différentes : dans COMOKIT, le fondement de la validité est à chercher dans l'adéquation, de nature statistique, du modèle par rapport au monde observable ; dans la modélisation d'accompagnement, en revanche, prime l'adéquation du modèle à une vision partagée du monde. Au besoin, cette vision partagée sera même créée si elle n'existe pas encore préalablement à l'exercice de modélisation, et le modélisateur pourra être tout autant – sinon davantage – heureux du résultat obtenu.

Dans les deux formes de modélisation, cependant, le paradigme multi-agents structure la démarche et la pensée selon des modalités qui lui sont propres. Son apport à la prise de décision – qui était le point de départ du deuxième cas – est de nous inviter à penser un devenir et une dynamique inscrits dans l'espace. Elle nous force en quelque sorte à nous interroger sur l'intelligence inscrite dans le terroir, ou de façon moins romantique peut-être, sur l'intelligence détenue collectivement par un groupement *situé* d'hommes en interaction. C'est ainsi que nous est apparue la nécessité d'une philosophie de l'espace, à côté d'une réflexion sur le temps. Nous sommes même allé jusqu'à y déceler – sans toutefois creuser cette piste – un point de départ potentiellement fécond pour une réflexion éthique nouvelle, attentive au contexte, à la situation des êtres plutôt qu'à leur essence.

Le dernier cas pratique avait pour objet l'accident de la route où une piétonne a trouvé la mort suite à la collision avec une voiture autonome, censée être surveillée par une opératrice humaine. Or celle-ci, pour des raisons que nous avons amplement commentées, était plus préoccupée par une émission projetée par son téléphone portable que sur le drame réel dont son environnement immédiat était à ce moment le théâtre. Grâce aux rapports de l'administration états-unienne, nous avons d'abord pu examiner cette mésaventure sous toutes ses coutures : la négligence d'Uber, la frivolité de l'État d'Arizona, le phénomène d'excès de confiance induite par l'automatisation nous sont apparus tout aussi importants, sinon davantage, que la défaillance logicielle qui était immédiatement à l'origine de l'accident. Ensuite, nous avons appliqué une représentation multi-agents à ce type de cas : elle s'est révélée non seulement génératrice de points de vue nouveaux, mais elle a aussi apporté une certaine hygiène intellectuelle, nous renvoyant sans concession aucune au concret du réel. Enfin, nous nous sommes appuyé sur ce cas pour aborder les critères d'une innovation technique responsable, problématique qui est, somme toute, voisine de celle de l'éthique des machines.

Voilà, en résumé, le contenu de ce mémoire : nous avons développé l'ambition de l'éthique des machines à calculer le comportement et sa justification. Les systèmes multi-agents sont à même de fournir des justifications, à comprendre comme un mécanisme déclaratif de prise de décision, aux comportements des agents qui les peuplent, s'inscrivant dans un temps et un espace particuliers : le temps peut être fonction du possible, d'ordre anticipatoire, être fonction aussi du probable, d'ordre prévisionnel. L'espace, quant à lui, peut être réticulé ou en forme de damier, dynamique ou non, l'essentiel étant toujours que l'espace permette de faire communiquer, en un seul langage, des savoirs plus ou moins formalisés, entre eux. La prise de décision éthique se joue sur trois plans, intrapersonnel, interpersonnel, impersonnel, où la force majeure du paradigme multi-agents semble

se situer au carrefour des deux niveaux supérieurs. C'est dire que la justification ne doit pas être individuelle (comme c'est le cas avec le BDI) ; cependant elle doit toujours chercher à lier normes et fins ; les premières étant *gage d'universalité*, les deuxièmes de *pertinence par rapport au contexte* dans lequel s'inscrit la décision à prendre. La prise en compte de la téléologie, l'ancrage spatial d'agents aux comportements divers, le tout traduisible en vertu de la métaphore agent vers le domaine de référence visé, voilà la force du paradigme multi-agents. Il incombe dès lors aux ingénieurs qui s'en servent d'en tirer le meilleur parti, que ce soit à des fins de connaissance empirique, d'exploration de mondes possibles ou, plus prosaïquement, de venir à bout d'un problème de coordination ardu.

Tout ceci a une incidence majeure sur le type de justification que nous pouvons attendre d'un système multi-agents. Concrètement, nous savons désormais que celle-ci sera ancrée dans un temps et un lieu particulier, récusant le déterminisme. Son ancrage dans le temps se manifeste dans son intérêt pour la question du devenir d'une situation particulière, ainsi que dans son orientation vers l'avenir, que ce soit sur un mode provisionnel ou anticipatoire. Son aversion du déterminisme tient à ce que, même s'il privilégie les solutions probables, obtenues par l'exécution répétée pour se faire une bonne idée de la normalité de la distribution, des exécutions isolées peuvent donner à voir un possible à la limite du probable.

Nous avons vu aussi, cependant, que la technique ne nous offre que très peu d'aide à sélectionner les agents les plus adéquats ; c'est là une tâche qui reste dévolue à la pensée critique, qui peut se faire aider de méthodologies tierces, dont nous avons approfondi un exemple sous la forme de la modélisation d'accompagnement, ou appeler à la rescousse les ressources de disciplines scientifiques diverses. Toutefois, en raison même de son indétermination, le choix des agents peut à l'occasion se révéler insidieux. Nous en avons vu un exemple lorsque nous avons commenté l'accident de la route de Tempe : faute d'avoir suivi le code de la route, une piétonne s'est vu refuser le statut d'agent, refus dont nous savons les conséquences.

Il faut avouer que le contenu de ce mémoire déborde de loin le plan que nous avons initialement présenté à notre directrice de mémoire. Même si notre lecteur en trouvera probablement encore d'autres, nous avons cependant pleinement conscience d'au moins quelques-unes de ses faiblesses : victime d'une documentation trop riche, notre argumentation s'est parfois perdue en digressions, laissant plus d'une fois sa proie pour suivre son ombre. Trop souvent aussi, nous nous sommes contenté de juxtaposer nos arguments, sans prendre suffisamment de recul et de temps pour les articuler ensemble. Nous devons en convenir : le fil conducteur de ce mémoire a tendance à s'emmêler, jusqu'à – en de rares occasions – s'enrouler sur lui-même. Nous osons cependant espérer quelque indulgence de la part de notre lecteur. Ce travail, sur une matière si vaste, ne s'est-il pas étalé sur trois ans ? Ainsi, au moment de finir la rédaction du troisième chapitre, notre relecture du premier nous a fait redécouvrir maint détail que nous avions déjà oublié !

Si donc ce travail ne brille point par sa concision, en revanche, nous croyons qu'il est le témoin exemplaire d'un certain parcours d'apprentissage. Sur le plan technique, d'abord : nous avions des systèmes multi-agents une idée assez vague avant d'entamer ce mémoire ; nous avons eu l'occasion de compulser une littérature foisonnante à son sujet. Mais on mesure très certainement aussi

l'évolution dans notre conception de l'éthique : sous l'influence d'une culture exclusivement littéraire peut-être, elle ressemblait au début davantage à une esthétique du comportement, voire à une poétique de l'action ; elle était purement téléologique et peu soucieuse d'universalité. Progressivement elle s'est ouverte à la norme, au vivre-ensemble que celle-ci rend possible. Nous avons également appris que toute vie en société se construit nécessairement sur un socle éthique.

Qu'il nous soit permis d'illustrer ce dernier apprentissage par une anecdote personnelle : au début de notre carrière professionnelle, dans une grande administration bancaire à Bruxelles, nous nous trouvâmes souvent devant la machine à café de l'étage. Comme toute machine à café, celle-ci offrait un petit choix de cafés divers accessibles à travers autant de boutons : café expresso, au lait, au chocolat, et ainsi de suite. Il suffisait d'introduire sa carte, d'appuyer sur le bouton correspondant au choix, et voilà que la machine éjectait promptement un gobelet dans lequel coulait, sans attendre, un flot généreux d'or noir. Cependant, parmi cet éventail de choix, il en fut un qui nous a toujours laissé un peu perplexe : au même tarif, en appuyant sur un bouton un centimètre plus bas, nous pouvions nous flatter de boire un café « éthique », estampillé, si notre souvenir est bon, Max Havelaar. Drôle de choix, en vérité ! Si ce café-là était éthique, cela voulait-il dire que les autres, eux, ne l'étaient pas ? D'un côté, le café équitable, promettant un revenu juste aux agriculteurs et une culture respectueuse de l'environnement ; alors, nous demandions-nous, quel choix sera le nôtre si nous prenons un autre café ? Des salaires de misère, ou pire des enfants dans les champs, des cultures riches en pesticides ne faisant d'heureux que parmi les actionnaires ? Nous avons fini par esquiver le problème en descendant au sous-sol, l'étage du réfectoire, où une machine à pression dispensait un café de grain fraîchement moulu ; sans choix ni promesses, cet appareil était dépourvu de la troublante ambiguïté de l'autre, qui officiait quelques étages plus haut.

La problématique de ce mémoire nous a fait repenser à cet épisode : une société juste ne devrait-elle pas refuser certains choix aux individus qui la composent ? Après tout, nous ne parlerons jamais à l'agriculteur qui s'éreinte sur un champ qui ne lui appartient pas, nous ne jouerons jamais avec l'enfant qui a confectionné nos vêtements bon marché. Pourtant, un beau matin au retour du printemps, alors que nous prendrons comme d'habitude une tasse de café au jardin, quand la dernière abeille viendra s'échouer dans un parterre sans fleurs, nous saurons que leur détresse est aussi la nôtre !

C'est dire que les défis éthiques du XXIe siècle dépassent l'individu ; leur fardeau semble si formidable et si lourd qu'ils paralysent les meilleures volontés, font reculer les esprits les plus intrépides. Sur le chantier où œuvrent les bâtisseurs d'un monde plus juste, quelle place pour l'éthique des machines ? Tel est, en dernière analyse, le problème que la nouvelle discipline doit affronter si elle veut tenir ses promesses. Sa perspective, certes, est limitée : pour reprendre l'exemple de la machine à café, elle n'aurait probablement rien trouvé à y redire. Mais quelle perspective peut se vanter d'embrasser la totalité des points de vue possibles ? L'éthique est essentiellement plurielle ; aussi les regards que nous portons sur elle doivent-ils être aussi divers que le faisceau de rayons lumineux qu'un prisme réfracte et disperse. La luminescence particulière de l'éthique des machines ne peut être que modeste ; nous aurons tort cependant de lui préférer une zone d'ombre, si fuyante soit-elle.

Ce n'est donc pas l'éthique des machines qui pourra décider s'il faut remplacer les infirmières humaines par leurs consœurs robotiques ; les abeilles par des mini-drones ou lâcher des tueurs automatisés sur les champs de bataille. En revanche, nous pouvons nous adresser à elle pour évaluer le risque, le potentiel d'une telle technique, et surtout, nous pouvons lui demander quels garde-fous *internes à la technique* peuvent accompagner l'innovation technique. Celle-ci ne s'arrêtera pas… et après tout, la technophobie de principe est tout autant négatrice de l'homme qu'une technolâtrie irréfléchie, car si l'homme a créé la technique, celle-ci a également créé l'homme. Elle continuera d'ailleurs à le faire, à contribuer à « l'hominescence » de l'homme, pour reprendre l'heureuse trouvaille terminologique de Michel Serres : elle obligera probablement l'homme à trouver sa dignité autre part que dans le travail ; elle nous forcera peut-être à revoir de fond en comble les questions de sécurité, de richesse et de paix à l'échelle de la planète. Robots humanoïdes, automatisation de larges pans de l'activité humaine, cybersécurité, propriété intellectuelle, etc., voici des questions dont l'importance ne cessera de croître et qui dès aujourd'hui contribuent à façonner nos valeurs et nos regards sur l'homme.

# Liste des abréviations

| | |
|---|---|
| AAII | *Australian Artificial Intelligence Institute* |
| ACL | *Agent Communication Language* |
| AR | *Analogical Reasoning* |
| ASP | *Answer Set Programming* |
| BDI | *Beliefs, Desires, Intentions* |
| COMOKIT | *COVID-19 Modeling Kit* |
| CTL | *Computational Tree Logic* |
| DALMAS | *Deontic Action-Logic Multi-Agent Systems* |
| DARPA | *Defense Advanced Research Project Agency* |
| DD | développement durable |
| EA NLU | *Explanation Agent Natural Language Understanding* |
| EASI | *Environment as Active Support for Interaction* |
| EM | éthique des machines |
| EMIL | *Emergence In the Loop* |
| EPSRC | *Engineering and Physical Sciences Research Council* |
| FIPA | *Foundation for Intelligent Physical Agents* |
| FPR | *First-Principles Reasoning* |
| IA | intelligence artificielle |
| IDD | Institut pour un Développement Durable |
| IRD | Institut de Recherche pour le Développement |
| JADE | *Java Agent Development Framework* |
| Lisp | *List Processing* |
| LTL | *Linear-time Temporal Logic* |
| MaNEA | *Magentix2 Norm-Enforcing Architecture* |
| MEESTAR | *Model for the Ethical Evaluation of Socio-Technical Arrangements* |
| NAC | *Norm Acquisition Context* |
| n-BDI | *normative BDI* |
| NCC | *Norm Compliance Context* |
| NTSB | *National Transportation Safety Board* |

| | |
|---|---|
| OCC | *Ortony, Clore et Collins* |
| O.D.D. | *Overview, Design concepts, Details* |
| OMR | *Orders of Magnitude Reasoning* |
| OMS | *Organization Management System* |
| PRS | *Procedural Reasoning System* |
| SAE | *Society of Automotive Engineers* |
| SBA | simulation(s) à base d'agents |
| SEIR | *Susceptible*, *Exposed*, *Infectious*, *Recovered* |
| SeSAm | *Shell for Simulated Agent Systems* |
| SF | science-fiction |
| SIG | système(s) d'information géographique |
| SMA | système(s) multi-agents |
| SOAP | *Simple Object Access Protocol* |
| SPP | service public de programmation |
| THOMAS | *Teams and Hierarchies for Open Multi-Agent Systems* |
| UML | *Unified Modeling Language* |

# Bibliographie

AMGOUD (Leila), BELABBES (Sihem) et PRADE (Henri), *Towards a Formal Framework for the Search of a Consensus between Autonomous Agents*, dans PARSONS (Simon), MAUDET (Nicolas), MORAITIS (Pavlos) et RAHWAN (Iyad), coord., *Argumentation in Multi-Agent Systems. Second International Workshop*, Heidelberg, Springer, 2005, coll. « Lecture Notes in Artificial Intelligence », vol. 4049, pp. 264-278.

ANDERSON (Michael) et ANDERSON (Susan Leigh), coord., *Machine Ethics*, Cambridge, Cambridge University Press, 2011.

AMGOUD (Leila), BELABBES (Sihem) et PRADE (Henri), *Towards a formal framework for the search of a consensus between autonomous agents*, dans PECHOUCEK (Michel), STEINER (Donald) et THOMPSON (Simon), coord., *AAMAS '05: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, Association for Computing Machinery, 2005, pp. 537-543.

ANDLER (Daniel), FAGOT-LARGEAULT (Anne) et SAINT-SERNIN (Bertrand), *Philosophie des sciences I*, Paris, Gallimard, 2002, coll. « Folio/Essais ».

IID., *Philosophie des sciences II*, Paris, Gallimard, 2002, coll. « Folio/Essais ».

ANDRIGHETTO (Giulia), Campennì (Marco), CONTE (Rosaria) et PAOLUCCI (Marco), *On the Immergence of Norms: a Normative Agent Architecture*, dans TRAJKOVSKI (Goran P.) et COLLINS (Samuel G.), coord., *Emergent Agents and Socialities. Social and Organizational Aspects of Intelligence*, 2007, coll. « AAAI Fall Symposium – Technical Report », pp. 11-18.

EAD., VILLATORO (Daniel) et CONTE (Rosaria), *Norm internalization in artificial societies*, dans *AI Communications*, vol. 23, n° 4, 2010, pp. 325-339.

ARKIN (Ronald C.), *Governing Lethal Behavior in Autonomous Robots*, Boca Raton [États-Unis], CRC Press, 2009.

AXELROD (Robert Marshall), *The Complexity of Cooperation. Agent-Based Models of Competition and Collaboration*, Princeton, Princeton University Press, 1997, coll. « Princeton Studies in Complexity ».

ID. et HAMILTON (William D.), *The Evolution of Cooperation*, dans *Science*, vol. 211, n° 4489, 1981, pp. 1390-1396.

BADARIOTTI (Dominique), BANOS (Arnaud) et MORENO (Diego), *Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine : le modèle Remus*, dans *Cybergeo. Revue européenne de géographie, Dossiers*, article 403, 2007.

BADDOURA (Ritta), *Le potentiel thérapeutique de l'interaction homme-robot*, dans *La Lettre du Psychiatre*, vol. 12, n° 1, 2016, pp. 8-11.

BADIOU (Alain), *Le concept de modèle. Introduction à une épistémologie matérialiste des mathématiques*, Paris, Fayard, 2007, coll. « Ouvertures ».

BAULT (Nadège), CHAMBON (Valérian), MAÏONCHI-PINO (Norbert), PÉNICAUD (François-Xavier), PUTOIS (Benjamin) et ROY (Jean-Michel), sous la dir. de, *Peut-on se passer de représentations en sciences cognitives ?*, Bruxelles, De Boeck, 2011, coll. « Neurosciences et cognition ».

BERSINI (Hugues), *Le Tamagotchi de Mme Yen et autres histoires*, Paris, Le Pommier, 2012, coll. « Plumes de science ».

ID., *Qu'est-ce que l'émergence ?*, Paris, Ellipses, 2007.

ID., *Quételet, l'invention de l'homme moyen par un homme tout sauf moyen*, conférence donnée au Collège Belgique en date du 27 novembre 2019.

BLONDEL (Éric), *Le problème moral*, Paris, Presses Universitaires de France, 2000, coll. « Philosopher ».

BOISSIER (Olivier), BALBO (Flavien) et BADEIG (Fabien), *Controlling multi-party interaction within normative multi-agent organizations*, dans DE VOS (Marina), FORNARA (Nicoletta), PITT (Jeremy V.) et VOUROS (George), coord., *Coordination, Organizations, Institutions, and Norms in Agent Systems VI. COIN 2010 International Workshops*, Heidelberg, Springer, 2010, coll. « Lecture Notes in Artificial Intelligence », vol. 6541, pp. 17-32.

BOLTANSKI (Luc) et THÉVENOT (Laurent), *De la justification. Les économies de la grandeur*, Paris, Gallimard, 1991, coll. « NRF Essais ».

BORDINI (Rafael H.) et HÜBNER (J. F.), *An Overview of Jason*, dans *The ALP newsletter*, vol. 19, n° 3, août 2006, disponible à l'adresse https://dtai.cs.kuleuven.be/projects/ALP/newsletter/aug06/.

ID. et MOREIRA (A. F.), *Proving BDI properties of agent-oriented programming languages. The asymmetry thesis principles in AgentSpeak(L)*, dans *Annals of Mathematics and Artificial Intelligence*, n° 42, 2004, pp. 197-226.

BOSSE (Tibor), SHARPANSKYKH (Alexei) et TREUR (Jan), *Integrating Agent Models and Dynamical Systems*, dans BALDONI (Matteo), SON (Tran Cao), RIEMSDIJK (M. Birna van) et WINIKOFF (Michael), coord., *Declarative Agent Languages and Technologies V. 5th International Workshop DALT*, Heidelberg, Springer, coll. « Lecture Notes in Artificial Intelligence », vol. 4897, pp. 50-68.

BOSTROM (Nick), *Superintelligence. Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2017.

BOULANGER (Paul-Marie) et BRÉCHET (Thierry), *Modélisation et aide à la décision pour un développement durable : état de l'art et perspectives. Rapport final au SPP Politique Scientifique (SPP-PS). Action de support AS/F5/16*, Ottignies, Institut pour un Développement Durable, 2003.

BOURETZ (Pierre), *D'un ton guerrier en philosophie. Habermas, Derrida & Co*, Paris, Gallimard, 2010, coll. « NRF Essais ».

BOURGAIS (Mathieu), TAILLANDIER (Patrick) et VERCOUTER (Laurent), *Cognition, émotions et relations sociales pour la simulation multi-agent*, dans GARBAY (Catherine) et BONNET (Grégory), *JFSMA 2017. Systèmes multi-agents. Cohésion : fondement ou propriété émergente*, Toulouse, Cépaduès, 2017, coll. « JFSMA », pp. 127-136.

BRATMAN (Michael E.), *Intention, Practical Rationality, and Self-Governance*, dans *Ethics*, n° 119, avril 2009, pp. 411–443.

BRUGIÈRE (Arthur), CHAPUIS (Kevin), CHOISY (Marc), DROGOUL (Alexis), GAUDOU (Benoît), HUYNH (Quang Nghi), NGUYEN (Ngoc Doanh), LARMANDE (Pierre), PHILIPPON (Damien) et TAILLANDIER (Patrick), *O.D.D. description of the COMOKIT model*, le 22 juin 2020, disponible à l'adresse https://comokit.org/ressources/ODD-COMOKIT_v1.0.1.pdf.

CHAPOUTHIER (Georges) et KAPLAN (Frédéric), *L'homme, l'animal et la machine*, Paris, CNRS Éditions, 2013, coll. « Biblis ».

CHAZAL (Gérard), *Le miroir automate. Introduction à une philosophie de l'informatique*, Seyssel, Champ Vallon, 1995, coll. « milieux ».

ID., *Les réseaux du sens. De l'informatique aux neurosciences*, Seyssel, Champ Vallon, 2000, coll. « milieux ».

COINTE (Nicolas), BONNET (Grégory) et BOISSIER (Olivier), *Coopération entre agents autonomes fondée sur l'éthique*, dans GARBAY (Catherine) et BONNET (Grégory), *JFSMA 2017. Systèmes multi-agents. Cohésion : fondement ou propriété émergente*, Toulouse, Cépaduès, 2017, coll. « JFSMA », pp. 9-18.

IID., *Éthique collective dans les systèmes multi-agents*, dans *Revue d'Intelligence Artificielle*, vol. 31, n° 1-2, 2017, pp. 71-96.

IID., *Jugement éthique dans le processus de décision d'un agent BDI*, dans *Revue d'Intelligence Artificielle*, vol. 31, n° 4, 2017, pp. 471-499.

IID., *Jugement éthique dans les systèmes multi-agents*, dans MICHEL (Fabien) et SAUNIER (Julien), sous la dir. de, *JFSMA 2016. Systèmes Multi-Agents et simulation*, Toulouse, Cépaduès, 2016, coll. « JFSMA », pp. 149-158.

IID., *Multi-Agent Based Ethical Asset Management*, dans BONNET (Grégory), HARBERS (Maaike), HINDRIKS (Koen), KATELL (Mike) et TESSIER (Catherine), coord., *Ethics in the Design of Intelligent Agents. Proceedings of the 1st Workshop on Ethics in the Design of Intelligent Agents*, 2016, pp. 52-57.

COMTE-SPONVILLE (André), *Petit traité des grandes vertus*, Paris, Éditions du Seuil, 1995, coll. « Points ».

CORAPI (Domenico), DE VOS (Marina), PADGET (Julian), RUSSO (Alessandra) et SATOH (Ken), *Norm Refinement and Design through Inductive Learning*, dans FORNARA (Nicoletta) et VOUROS (George), coord., *Coordination, Organizations, Institutions, and Norms in Agent Systems VI. COIN 2010 International Workshops*, Heidelberg, Springer, 2010, coll. « Lecture Notes in Artificial Intelligence », vol. 6541, pp. 33-48.

CRIADO (Natalia), ARGENTE (Estefania), NORIEGA (Pablo) et BOTTI (V.), *Towards a Normative BDI Architecture for Norm Compliance*, dans FORNARA (Nicoletta) et VOUROS (George), coord., *Coordination, Organizations, Institutions, and Norms in Agent Systems VI. COIN 2010 International Workshops*, Heidelberg, Springer, 2010, coll. « Lecture Notes in Artificial Intelligence », vol. 6541, pp. 65-81.

EAD., *Using Norms to Control Open Multi-Agent Systems. Programa de doctorado de reconocimiento de formas e inteligencia artificial*, Valence, Universidad Politécnica de Valencia, 2012.

DAMASIO (Antonio R.), *L'erreur de Descartes. La raison des émotions*, trad. de l'anglais ,Paris, Odile Jacob, 2021, coll. « sciences ».

DANIELSON (Peter A.), *Artificial Morality. Virtuous robots for virtual games*, Londres-New York, Routledge, 1992.

ID., coord., *Modeling Rationality, Morality, and Evolution*, Oxford, Oxford University Press, 1998, coll. « Vancouver Studies in Cognitive Science ».

DELCOURT (Marie), *In Memoriam. Jean Hubaux (Marcinelle 1894 - Liège 1959)*, dans *L'antiquité classique*, tome 28, fasc. 1, 1959, pp. 4-8.

DIGNUM (Frank), *Autonomous agents with norms*, dans *Artificial Intelligence and Law*, vol. 7, n° 1, 1999, pp. 69-79.

D'INVERNO (Mark), LUCK (Michael) et GEORGEFF (Michael), *The dMARS Architecture. A Specification of the Distributed Multi-Agent Reasoning System*, dans *Autonomous Agents and Multi-Agents Systems*, n° 9, 2004, pp. 5-53.

DJERROUD (Halim) et CHERIF (Arab Ali), *Environment Engine for Situated MAS*, dans *ICAART 2019. Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, vol. 1, Setúbal, SciTePress, pp. 129-137.

DOMINICY (Marc) et FRÉDÉRIC (Madeleine), sous la dir. de, *La mise en scène des valeurs. La rhétorique de l'éloge et du blâme*, Lausanne, Delachaux et Niestlé, 2001, coll. « Textes de base en Sciences du Discours ».

DRAY (Anne), PEREZ (Pascal), JONES (Natalie), LE PAGE (Christophe), D'AQUINO (Patrick), WHITE (Ian) et AUATABU (Titeem), *The AtollGame Experience: from Knowledge Engineering to a Computer-Assisted Role Playing Game*, in *Journal of Artificial Societies and Social Simulation*, 2006, vol. 9, n° 1.

EAD., PEREZ (Pascal), LE PAGE (Christophe), D'AQUINO (Patrick) et WHITE (Ian), *Who wants to terminate the game? The role of vested interests and metaplayers in the AtollGame experience*, in *Simulation & Gaming*, 2007, vol. 38, n° 4, pp. 494-511.

DROGOUL (Alexis), TAILLANDIER (Patrick), GAUDOU (Benoît), CHOISY (Marc), CHAPUIS (Kevin), HUYNH (Nghi Quang), NGUYEN (Ngoc Doanh), PHILIPPON (Damien), BRUGIÈRE (Arthur) et LARMANDE (Pierre), *Designing social simulation to (seriously) support decision-making: COMOKIT, an agent-based modelling toolkit to analyze and compare the impacts of public health interventions against COVID-19*, dans *Review of Artificial Societies and Social Simulation*, le 27 avril 2020.

DUBOIS (Michel J. F.), *La métaphore et l'improbable. Émergence de l'esprit post-scientifique ?*, Paris, L'Harmattan, 2015, coll. « Ouverture philosophique ».

DUYCKAERTS (Éric), *Expérience imaginaire et Intelligence Artificielle*, dans *Quaderni*, n° 1, printemps 1987, pp. 47-63.

EITER (Thomas), IANNI (Giovambattista) et KRENNWALLNER (Thomas), *Answer Set Programming : A Primer*, dans TESSARIS (Sergio), FRANCONI (Enrico), EITER (Thomas), GUTERRIEZ (Claudio), HANDSCHUH (Siegfried), ROUSSET (Marie-Christine) et SCHMIDT (Renate A.), coord., *Reasoning Web. Semantic Technologies for Information Systems*, Heidelberg, Springer, 2009, pp. 40-110.

ELLUL (Jacques), *Le Système technicien*, Paris, le Cherche-Midi, 2012, coll. « Documents/Société ».

ID., *Pour qui, pour quoi travaillons-nous ? Textes choisis, présentés et annotés par* M. HOURCADE, J.-P. JÉZÉQUEL *et* G. PAUL, Paris, La Table Ronde, 2018, coll. « la petite vermillon ».

ÉTIENNE (Michel), coord., *La modélisation d'accompagnement. Une démarche participative en appui au développement durable*, Paris, Éditions Quæ, 2010, coll. « Update Science & Technologies ».

EVERTSZ (Rick), FLETCHER (Martyn), JONES (Richard), JARVIS (Jacquie), BRUSEY (James) et DANCE (Sandy), *Implementing Industrial Multi-agent Systems Using JACK™*, dans DASTANI (Mehdi), DIX (Jürgen) et SEGHROUCHNI (Amal El Fallah), coord., *Programming Multi-Agent Systems. First International Workshop, ProMAS 2003*, Heidelberg, Springer, 2003, coll. « Lecture Notes in Artificial Intelligence », vol. 3067, pp. 18-48.

FIPA, *FIPA ACL Message Structure Specification*, Genève, Foundation for Intelligent Physical Agents, 2002, disponible à l'adresse : http://www.fipa.org/specs/fipa00061/SC00061G.pdf.

ID., *FIPA Communicative Act Library Specification*, Genève, Foundation for Intelligent Physical Agents, 2002, disponible à l'adresse : http://www.fipa.org/specs/fipa00037/SC00037J.pdf.

FISHER (Michael), BORDINI (Rafael H.), HIRSCH (Benjamin) et TORRONI (Paolo), *Computational Logics and Agents. A Road Map of Current Technologies and Future Trends*, dans *Computational Intelligence*, vol. 23, n° 1, 2007, pp. 61-90.

FIX (Julia), SCHEVE (Christian von) et MOLDT (Daniel), *Emotion-based norm enforcement and maintenance in multi-agent systems: foundations and petri net modeling*, dans *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2006, pp. 105-107.

FLORIDI (Luciano), *Information ethics. On the philosophical foundation of computer ethics*, dans *Ethics and Information Technology*, n° 1, 1999, pp. 37–56.

FODOR (Jerry A.) et PYLYSHYN (Zenon W.), *Connectionism and Cognitive Architecture. A Critical Analysis*, dans *Cognition*, vol. 28, n° 1-2, 1988, pp. 3-71.

FOLLON (Jacques), *Réflexions sur la théorie aristotélicienne des quatre causes*, dans *Revue Philosophique de Louvain*, tome 86, n°71, 1988, pp. 317-353.

FOUCAULT (Michel), *Histoire de la sexualité I. La volonté de savoir*, Paris, Gallimard, 1976, coll. « Tel ».

ID., *Histoire de la sexualité II. L'usage des plaisirs*, Paris, Gallimard, 1984, coll. « Tel ».

FRISCH (Max), *Homo faber. Ein bericht*, Frankfurt am Main, Suhrkamp Verlag, 1977.

GANASCIA (Jean-Gabriel), *Modelling ethical rules of lying with Answer Set Programming*, dans *Ethics and Information Technology*, n° 9, 2007, pp. 39-47.

GAUDOU (Benoît), HUYNH (Nghi Quang), PHILIPPON (Damien), BRUGIÈRE (Arthur), CHAPUIS (Kevin) TAILLANDIER (Patrick), LARMANDE (Pierre) et DROGOUL (Alexis), *COMOKIT: A Modeling Kit to Understand, Analyze, and Compare the Impacts of Mitigation Policies Against the COVID-19 Epidemic at the Scale of a City,* dans *Frontiers in Public Health*, le 24 septembre 2020.

GELFOND (Michael), *Answer Sets*, dans HARMELEN (Frank van), LIFSCHITZ (Vladimir) et PORTER (Bruce), coord., *Handbook of Knowledge Representation*, Amsterdam, Elsevier, 2008, coll. « Foundations of Artificial Intelligence », pp. 285-316.

GIANCOLA (Michael), BRINGSJORD (Selmer), GOVINDARAJULU (Naveen Sundar) et LICATO (John), *Adjudication of Symbolic & Connectionist Arguments in Autonomous-Driving AI*, dans *EPiC Series in Computing*, vol. 72, 2020, pp. 28-33.

GIRAUD (Gaël), *La théorie des jeux*, Paris, Flammarion, 2009, coll. « Champs/essais ».

GOFFI (Jean-Yves), *Les transhumanismes, la technique, la terre et l'espace*, conférence donnée au Collègue Belgique en date du 9 octobre 2019.

GRAZIANI (Romain), *L'Usage du vide. Essai sur l'intelligence de l'action, de l'Europe à la Chine*, Paris, Gallimard, 2019, coll. « Bibliothèque des Idées ».

GRESS (Thibaut) et MIRAULT (Paul), *La philosophie au risque de l'intelligence extraterrestre*, Paris, Vrin, 2016, coll. « Pour Demain ».

GUNKEL (David J.), *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*, Cambridge, MIT Press, 2012.

HARDY (Sam A.) et CARLO (Gustavo), *Identity as a Source of Moral Motivation*, dans *Human Development*, n° 48, 2005, pp. 232-256.

HJELMBLOM (Magnus), *Deontic Action-Logic Multi-Agent Systems in Prolog*, Gävle, Högskolan i Gävle, 2008.

ID. et ODELSTAD (Jan), jDALMAS: *A Java/Prolog Framework for Deontic Action-Logic Multi-Agent Systems*, dans HÅKANSSON (Anne), NGUYEN (Ngoc Thanh), HARTUNG (Ronald L.), HOWLETT (Robert J.) et JAIN (Lakhmi C.), coord., *Agent and Multi-Agent Systems: Technologies and Applications. Proceedings of the Third KES International Symposium*, Heidelberg, Springer, 2009, coll. «Lecture Notes in Artificial Intelligence », vol. 5559, pp. 110-119.

HOTTOIS (Gilbert), *Généalogies philosophique, politique et imaginaire de la technoscience*, Paris, Vrin, 2013, coll. « Pour demain ».

ID., *La technoscience : de l'origine du mot à ses usages actuels*, dans *Recherche en soins infirmiers*, vol. 3, n° 86, 2006, pp. 24-32.

ID., coord., *Philosophie et science-fiction*, Paris, Vrin, 2000, coll. « Annales de l'institut de philosophie de l'Université de Bruxelles ».

ID., *Species Technica. Suivi d'un dialogue philosophique autour de* Species Technica *vingt ans plus tard*, Paris, Vrin, 2002, coll. « Pour demain ».

HÜBNER (Jomi Fred), BOISSIER (Olivier),·KITIO (Rosine)·et RICCI (Alessandro), *Instrumenting multi-agent organisations with organisational artifacts and agents. "Giving the organisational power back to the agents"*, dans *Autonomous Agents and Multi-Agent Systems*, vol. 20, n° 3, 2010, pp. 369-400.

ID., SICHMAN (Jaime Simão) et BOISSIER (Olivier), *Developing Organised Multi-Agent Systems Using the Moise+ Model: Programming Issues at the System and Agent Levels*, dans *International Journal of Agent-Oriented Software Engineering*, vol. 1, n° 3-4, 2007, pp. 370-395.

ID., VERCOUTER (Laurent) et BOISSIER (Olivier), *Instrumenting Multi-Agent Organisations with Reputation Artifacts*, dans DIGNUM (Virginia) et MATSON (Eric), coord., *Coordination, Organization, Institutions and Norms in Agent Systems*, 2008, coll. « AAAI Workshop – Technical Report », pp. 17-24.

HUI (Yuk), *On the Existence of Digital Objects*, Minneapolis-Londres, University of Minnesota Press, 2016.

ID., *Simondon et la question de l'information*, dans BARTHÉLÉMY (J.-H.), sous la dir. de, *Cahiers Simondon*, n° 6, Paris, L'Harmattan, 2015, pp. 29-47.

JACKSON (Gabrielle B.), *Skill and the Critique of Descartes in Gilbert Ryle and Maurice Merleau-Ponty*, dans SEMONOVITCH (Kascha) et DEROO (Neal), coord., *Merleau-Ponty at the Limits of Art, Religion and Perception*, New York, Continuum, 2010, pp. 63-78.

JAFFRO (Laurent) et LABRUNE (Monique), sous la dir. de, *Gradus philosophique. Un répertoire d'introductions méthodiques à la lecture des œuvres*, Paris, GF Flammarion, 1996.

JONAS (Hans), *Le Principe responsabilité. Une éthique pour la civilisation technologique*, trad. de l'allemand, Paris, Flammarion, 1995, coll. « Champs essais ».

JORAY (Pierre), sous la dir. de, *La quantification dans la logique moderne*, Paris, L'Harmattan, 2005, coll. « Épistémologie et Philosophie des Sciences ».

KALLINIKOS (Jannis), AALTONEN (Aleksi) et MARTON (Attila), *A theory of digital objects*, dans *First Monday*, vol. 15, n° 6, 7 juin 2010.

KAZAKOV (Dimitar) et KUDENKO (Daniel), *Machine Learning and Inductive Logic Programming for Multi-Agent Systems*, dans LUCK (Michael), MAŘÍK (Vladimír), ŠTĚPÁNKOVÁ (Olga) et TRAPPL (Robert), coord., *Multi-Agent Systems and Applications*, Heidelberg, Springer, 2001, coll. « Lecture Notes in Artificial Intelligence », vol. 2086, pp. 246-270.

KERMISCH (Céline) et HOTTOIS (Gilbert), coord., *Techniques et philosophies des risques*, Paris, Vrin, 2007, coll. « Pour demain ».

KIRN (Stefan), HERZOG (Otthein), LOCKEMANN (Peter) et SPANIOL (Otto), coord., *Multiagent Engineering. Theory and Applications in Enterprises*, Heidelberg, Springer, 2006, coll. « International Handbooks on Information Systems ».

KOLLINGBAUM (Martin J.) et NORMAN (Timothy J.), *Norm Adoption and Consistency in the NoA Agent Architecture*, dans DASTANI (Mehdi), DIX (Jürgen) et SEGHROUCHNI (Amal El Fallah), coord., *Programming Multi-Agent Systems. First International Workshop, ProMAS 2003*, Heidelberg, Springer, 2003, coll. « Lecture Notes in Artificial Intelligence », vol. 3067, pp. 169-186.

ID., VASCONCELOS (Wamberto), GARCÍA-CAMINO (Andres) et NORMAN (Timothy J.), *Conflict Resolution in Norm-Regulated Environments Via Unification and Constraints*, dans BALDONI (Matteo), SON (Tran Cao), RIEMSDIJK (M. Birna van) et WINIKOFF (Michael), coord., *Declarative Agent Languages and Technologies V. 5th International Workshop DALT*, Heidelberg, Springer, 2008, coll. « Lecture Notes in Artificial Intelligence », vol. 4897, pp. 158-174.

LADRIÈRE (Jean), *Mathématiques et formalisme*, dans *Revue des Questions Scientifiques*, tome 188, n° 4, 2017, pp. 456-492, fac-similé de l'article paru en octobre 1955.

LAVENDHOMME (René), *Introduction à la théorie des catégories*, dans *Revue des Questions Scientifiques*, tome 188, n° 4, 2017, pp. 493-512, conférence donnée à la réunion de la Société Scientifique de Bruxelles, le 12 avril 1967.

LE NY (Jean-François), *Comment l'esprit fait du sens. Notions et résultats des sciences cognitives*, Paris, Odile Jacob, 2005.

LEVINAS (Emmanuel), *Autrement qu'être ou au-delà de l'essence*, Paris, Le Livre de Poche, 1987, coll. « Biblio essais ».

LIN (Patrick), JENKINS (Ryan) et ABNEY (Keith), coord., *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*, Oxford, Oxford University Press, 2017.

LOTZMANN (Ulf) et MÖHRING (Michael), *Simulating Normative Behavior and Norm Formation Processes*, dans OTAMENDI (Javier), BARGIELA (Andrzej), MONTES (Jose Luis) et DONCEL (Luis), coord., *European Conference on Modelling and Simulation 2009 Proceedings*, European Council for Modeling and Simulation, 2009, pp. 187-193

LUCK (Michael), GRIFFITHS (Nathan) et D'INVERNO (Mark), *From Agent Theory to Agent Construction. A Case Study*, dans MÜLLER (Jörg P.), WOOLDRIDGE (Michael J.) et JENNINGS (Nicholas R.), coord., *Intelligent Agents III. Agent Theories, Architectures, and Languages*, Heidelberg, Springer, 1997, coll. « Lecture Notes in Artificial Intelligence », vol. 1193, pp. 49-63.

MAETERLINCK (Maurice), *La vie des fourmis*, Paris, Le Livre de Poche, 1964.

MASCARDI (Viviana), DEMERGASSO (Daniela) et ANCONA (Davide), *Languages for Programming BDI-style Agents : an Overview*, dans DI NAPOLI (Claudia), ROSSI (Silvia) et STAFFA (Mariacarla), coord., *From Objects to Agents. Proceedings of the 16th Workshop "From Objects to Agents"*, Bologne, Pitagora Editrice, 2005, pp. 9-15.

MATHIVET (Virginie), *L'intelligence artificielle pour les développeurs. Concepts et implémentations en Java*, Saint Herblain, ENI, 2015, coll. « DataPro ».

MEYER (Michel), *Principia Moralia*, Paris, Fayard, 2013.

ID., *Principia Rhetorica. Une théorie générale de l'argumentation*, Paris, Fayard, 2008, coll. « Ouvertures ».

MICHAUX (Henri), *Poteaux d'angle*, Paris, Gallimard, 1981, coll. « Poésie ».

MISSELHORN (Catrin), *Grundfragen der Machinenethik*, Stuttgart, Philipp Reclam, 2019, coll. « Reclams Universal-Bibliothek ».

MULDER (Laetitia B.), JORDAN (Jennifer) et RINK (Floor), *The effect of specific and general rules on ethical decisions*, dans *Organizational Behavior and Human Decision Processes*, vol. 126, 2015, pp. 115-129.

NAGEL (Ernest), NEWMAN (James R.), GÖDEL (Kurt) et GIRARD (Jean-Yves), *Le théorème de Gödel*, trad. de l'allemand et de l'anglais, Paris, Éditions du Seuil, 1989, coll. « Points/Sciences ».

NORMAN (Timothy J.) et LONG (Derek), *Goal Creation in Motivated Agents*, dans WOOLDRIDGE (Michael J.) et JENNINGS (Nicholas R.), coord., *Intelligent Agents. ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, Heidelberg, Springer, 1994, coll. « Lecture Notes in Artificial Intelligence », vol. 890, pp. 277-290.

NTSB, *Highway Accident Report: Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida May 7, 2016*, Washington D.C, National Transportation Safety Board, 2017, disponible à l'adresse : https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf.

ID., *Highway Accident Report: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018*, Washington D.C, National Transportation Safety Board, 2019, disponible à l'adresse : https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

ID., *Human Performance Group Chairman's Factual Report*, Washington D.C, National Transportation Safety Board, 2019, disponible à l'adresse : https://data.ntsb.gov/Docket/Document/docBLOB?ID=40477743&FileExtension=.PDF&FileName=Human%20Performance%20Factual-Master.PDF.

ID., *Vehicle Automation Report*, Washington D.C, National Transportation Safety Board, 2019, disponible à l'adresse : https://data.ntsb.gov/Docket/Document/docBLOB?ID=40477717&FileExtension=.PDF&FileName=Vehicle%20Automation%20Report-Master.PDF.

ID., *Volvo XC90 Testing by Thatcham Research*, Washington D.C, National Transportation Safety Board, 2019, disponible à l'adresse : https://data.ntsb.gov/Docket/Document/docBLOB?ID=40477755&FileExtension=.PDF&FileName=Volvo%20XC90%20Testing%20by%20Thatcham%20Research-Master.PDF.

OJEDA (Almerindo E.), *A Computational Introduction to Linguistics. Describing Language in plain Prolog*, Stanford, CSLI Publications, 2013.

OWEN (Richard), BESSANT (John) et HEINTZ (Maggy), coord., *Responsible Innovation. Managing the responsible émergence of science and innovation in society*, Chichester [Royaume-Uni], John Wiley & Sons, 2013.

PARROCHIA (Daniel), *L'homme volant. Philosophie de l'aéronautique et des techniques de navigation*, Seyssel, Champ Vallon, 2003, coll. « milieux ».

ID. et TIRLONI (Valentina), sous la dir. de, *Formes, systèmes et milieux techniques après Simondon*, Lyon, Jacques André, 2012, coll. « Thériaka, remèdes et rationalités ».

PERELMAN (Chaïm), *L'empire rhétorique. Rhétorique et argumentation*, Paris, Vrin, 2002, coll. « Bibliothèque d'histoire de la philosophie ».

PHAM (Duc Quang), HARLAND (James) et WINIKOFF (Michael), *Modeling Agents' Choices in Temporal Linear Logic*, dans BALDONI (Matteo), SON (Tran Cao), RIEMSDIJK (M. Birna van) et WINIKOFF (Michael), coord., *Declarative Agent Languages and Technologies V. 5th International Workshop DALT*, Heidelberg, Springer, 2007, coll. « Lecture Notes in Artificial Intelligence », vol. 4897, pp. 140-157.

PIGNARRE (Philippe) et STENGERS (Isabelle), *La sorcellerie capitaliste. Pratiques de désenvoûtement*, Paris, La Découverte, 2007.

POKAHR (Alexander), BRAUBACH (Lars) et LAMERSDORF (Winfried), *Jadex: Implementing a BDI-Infrastructure for JADE Agents*, dans *In Search of Innovation*, vol. 3, n° 3, septembre 2003, pp. 76-85.

PUTNAM (Hilary), *Le Réalisme à visage humain*, trad. de l'anglais, Paris, Gallimard, 2011, coll. « Tel ».

RAMGE (Thomas), *Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern*, Stuttgart, Philipp Reclam, 2019, coll. « Reclams Universal-Bibliothek [Was bedeutet das alles?] ».

RAYNAUD (Philippe) et RIALS (Stéphane), sous la dir. de, *Dictionnaire de philosophie politique*, Paris, Presses Universitaires de France, 1996, coll. « Quadrige ».

RICŒUR (Paul), *La métaphore vive*, Paris, Éditions du Seuil, 1975, coll. « Points/Essais ».

ID., *Le mal. Un défi à la philosophie et à la théologie*, Genève, Labor et Fides, 2004.

ID., *Soi-même comme un autre*, Paris, Éditions du Seuil, 1990, coll. « Points/Essais ».

ID., *Temps et récit 1. L'intrigue et le récit historique*, Paris, Éditions du Seuil, 1983, coll. « Points/Essais ».

ID., *Temps et récit 3. Le temps raconté*, Paris, Éditions du Seuil, 1985, coll. « Points/Essais ».

SALANSKIS (Jean-Michel), *Le monde du computationnel*, Paris, Les Belles Lettres, 2011, coll. « À présent ».

SARTRE (Jean-Paul), *Esquisse d'une théorie des émotions*, Paris, Le Livre de Poche, 1995, coll. « Références ».

SCHALANSKY (Judith), *Der Hals der Giraffe. Bildungsroman*, Berlin, Suhrkamp Verlag, 2011.

SCHEVE (Christian von), MOLDT (Daniel), FIX (Julia) et LÜDE (Rolf von), *My Agents Love to Conform: Emotions, Norms, and Social Control in Natural and Artificial Societies*, dans *Social Intelligence and Interaction in Animals, Robots and Agents. Proceedings of the Symposium on Normative Multi-Agent Systems*, Claverton Down, AISB, 2005, pp. 73-84.

SCHWARTZ (Shalom H.), *Les valeur de base de la personne. Théorie, mesures et applications*, dans *Revue française de sociologie*, 2006, vol. 47, n° 4, pp. 929-968.

SHOHAM (Yoav), *Agent Oriented Programming. An overview of the framework and summary of recent research*, dans NASA, Lyndon B. Johnson Space Center, *The Sixth Annual Workshop on Space Operations Applications and Research*, 1993, pp. 296-304.

SIEKMANN (Jörg H.), sous la dir. de, *Computational Logic*, Amsterdam, Elsevier, 2014, coll. « Handbook of the History of Logic ».

SIMON (Herbert A.), *Les sciences de l'artificiel*, trad. de l'anglais, Paris, Gallimard, 2004, coll. « Folio/Essais ».

SIMONDON (Gilbert), *Du mode d'existence des objets techniques*, Paris, Aubier, 2012, coll. « Philosophie ».

SLOTERDIJK (Peter), *Du mußt dein Leben ändern*, Frankfurt am Main, Suhrkamp, 2011.

STARHAWK, *Quel monde voulons-nous ?, trad. de l'anglais et présenté par Isabelle Stengers*, Paris, Éditions Cambourakis, 2019, coll. « sorcières ».

STENGERS (Isabelle), *Au temps des catastrophes. Résister à la barbarie qui vient*, Paris, La Découverte, 2013.

EAD., *Cosmopolitiques I. La guerre des sciences. L'invention de la mécanique : pouvoir et raison. Thermodynamique : la réalité physique en crise*, Paris, La Découverte, 2003.

EAD., *Cosmopolitiques II. Mécanique quantique : la fin du rêve. Au nom de la flèche du temps : le défi de Prigogine. La vie et l'artifice : visages de l'émergence. Pour en finir avec la tolérance*, Paris, La Découverte, 2003.

EAD., sous la dir. de, *D'une science à l'autre. Des concepts nomades*, Paris, Éditions du Seuil, 1987, coll. « Science ouverte ».

EAD., *Sciences et pouvoirs. Faut-il en avoir peur ?*, Bruxelles, Éditions Labor, 1997, coll. « Quartier libre ».

EAD., *Souviens-toi que je suis Médée. Medea nunc sum*, Le Plessis-Robinson, Synthélabo, 1993, coll. « Les Empêcheurs de penser en rond ».

EAD., *Une autre science est possible ! Manifeste pour un ralentissement des sciences*, Paris, La Découverte, 2013, coll. « Les Empêcheurs de penser en rond ».

EAD. et SCHLANGER (Judith), *Les concepts scientifiques. Invention et pouvoir*, Paris, Gallimard, 1991, coll. « folio/essais ».

TESSIER (Catherine), *Autonomie des robots : enjeux techniques et perspectives*, dans DOARÉ (Ronan), DANET (Didier) et BOISBOISSEL (Gérard de), *Drones et* killer robots*. Faut-il les interdire ?*, Rennes, Presses Universitaires de Rennes, coll. « L'univers des normes », pp. 65-77.

THIRY (Marcel), *Nouvelles du grand possible. Lecture de Pascal DURAND*, Bruxelles, Éditions Labor, 1987, coll. « Espace Nord ».

ID., *Œuvres poétiques complètes. Tome II 1950-1969*, Bruxelles, Académie royale de langue et de littérature françaises de Belgique, 1997.

TICKOO (Asha), *On assertion without free speech*, dans *Journal of Pragmatics*, n° 42, 2010, pp. 1577-1594.

TIDHAR (Gil), *Flying Together. Modelling Air Mission Teams*, dans *Applied Intelligence*, n° 8, 1998, pp. 195-218.

TINLAND (Frank), sous la dir. de, *Ordre biologique ordre technologique*, Seyssel, Champ Vallon, 1994, coll. « milieux ».

TRAPPL (Robert), coord., *A Construction Manual for Robots' Ethical Systems. Requirements, Methods, Implementations*, Heidelberg, Springer, 2015.

TREUIL (Jean-Pierre), DROGOUL (Alexis) et ZUCKER (Jean-Daniel), *Modélisation et simulation à base d'agents. Exemples commentés, outils informatiques et questions théoriques*, Paris, Dunod, 2008.

TURING (Alan) et GIRARD (Jean-Yves), *La machine de Turing*, trad. de l'anglais, Paris, Éditions du Seuil, 1995, coll. « Points/Sciences ».

VARENNE (Frank), *Les notions de métaphore et d'analogie dans les épistémologies des modèles et des simulations*, Paris, Pétra, 2006, coll. « ACTA STOÏCA ».

VERBEEK (Peter-Paul), *Obstetric Ultrasound and the Technological Mediation of Morality : A Postphenomenological Analysis*, dans *Human Studies*, n° 31, 2008, pp. 11-26.

VERNANT (Denis), *Introduction à la philosophie contemporaine du langage. Du langage à l'action*, Paris, Armand Colin, 2010.

VIEIRA (Renata), MOREIRA (Álvaro F.), BORDINI (Rafael H.) et HÜBNER (Jomi), *An Agent-Oriented Programming Language for Computing in Context*, dans DEBENHAM (John), coord., *Professional Practice in Artificial Intelligence*, 2006, coll. « IFIP International Federation for Information Processing, », vol. 218, pp. 61-70.

VISSER (Wietske), HINDRIKS (Koen) et JONKER (Catholijn), *Argumentation-Based Preference Modelling with Incomplete Information*, dans DIX (Jürgen), FISHER (Michael) et NOVÁK (Peter), coord., *Proceedings of the 10th International Workshop on Computational Logic in Multi-Agent Systems*, 2009, Clausthal-Zellerfeld, Technische Universität Clausthal, coll. « IfI Technical Report Series », pp. 156-171.

VOLVO CARS, *Submission to the National Transportation Safety Board (N.T.S.B.) for the Tempe Accident involving an UBER test vehicle based on a Volvo XC90 MY2017*, Gothenburg, 2019, disponible à l'adresse: https://data.ntsb.gov/Docket/Document/docBLOB?ID=40477754&FileExtension=.PDF&FileName=Volvo%20Cars%20Party%20Submission-Master.PDF.

WALLACH (Wendell) et ALLEN (Colin), *Moral Machines. Teaching Robots Right from Wrong*, Oxford, Oxford University Press, 2009.

WEISS (Gerhard), coord., *Multiagent Systems*, Cambridge, MIT Press, 2013.

WIEGEL (Vincent), VAN DEN HOVEN (M. Jeroen) et LOKHORST (G. J. C.), *Privacy, deontic epistemic action logic and software agents. An executable approach to modeling moral constraints in complex informational relationships*, dans *Ethics and Information Technology*, n° 7, 2005, pp. 251-264.

WOERTHER (Frédérique), *Aux origines de la notion rhétorique d'èthos*, dans *Revue des Études Grecques*, tome 118, janvier-juin 2005, pp. 79-116.

WOOLDRIDGE (Michael), *An Introduction to MultiAgent Systems*, Chichester, John Wiley & Sons Ltd, 2009.

ZDENEK (Sean), *Artificial intelligence as a discursive practice : the case of embodied software agent systems*, dans *AI & Society*, vol. 17, n° 3-4, 2003, pp. 340-363.

# Table des matières

Résumé ..... 2
Abstract ..... 3
Remerciements ..... 4
Introduction ..... 6
i) Vous avez dit « éthique » ? ..... 7
ii) Quelle éthique pour une machine ? ..... 9
iii) Les systèmes multi-agents ..... 10
iv) L'éthique des systèmes ..... 14
Chapitre Ier. L'éthique des machines ..... 17
1.1. Éthique des machines et éthique de l'informatique ..... 17
1.2. Une approche fonctionnelle ..... 18
1.3. Comportements éthiques implicites et explicites ..... 22
1.4. Démarches descriptives et prescriptives ..... 24
1.4.1. MoralDM ..... 25
1.4.2. Le conseiller éthique ..... 27
1.5. Démarches ascendantes et descendantes ..... 31
1.6. La justification ..... 34
1.7. L'agentivité fonctionnelle ..... 37
1.7.1. Approches classiques ..... 37
1.7.2. Approche fonctionnelle ..... 40
1.7.2.1. Unité d'agir ..... 40
1.7.2.2. Identité ..... 44
1.7.2.3. Autonomie ..... 46
1.7.2.4. Intentionnalité ..... 49
1.7.2.5. Liberté ..... 51
1.7.2.6. Responsabilité ..... 52
1.7.2.7. Interactivité ..... 55
1.7.2.8. États internes ..... 57
1.7.2.9. Conscience de soi ..... 59
1.7.2.10. Adaptabilité et apprentissage ..... 61
1.7.2.11. Souffrance ..... 63
1.7.2.12. Émotions ..... 66
1.7.2.13. Résumons… ..... 67
1.8. La valeur fonctionnelle ..... 69

1.8.1. Un système de valeurs ..... 71
1.8.2. Les valeurs en action : le blâme et la reconnaissance ..... 74
1.8.3 Les valeurs en mutation : la négociation ..... 77
1.8.3.1. Négociation et image de soi ..... 78
1.8.3.2. L'obéissance à la règle reste la norme ..... 80
1.8.3.3. Négociation et langage ..... 82
Chapitre II. Le paradigme multi-agents ..... 85
2.1. L'agent comme métaphore ..... 86
2.1.1. L'agent en quête d'identité ? ..... 87
2.1.2. L'agent intentionnel, ou la métaphore mentaliste ..... 90
2.1.2.1. L'agent raisonneur : PRS et ses descendants ..... 91
2.1.2.2. Les logiques BDI ..... 93
2.1.2.3. Raisonnement BDI : limites et perspectives ..... 97
2.1.3. L'agent autonome ..... 100
2.1.3.1. L'agent qui voulait atteindre ses cibles ..... 100
2.1.3.2. L'agent évolutif ..... 104
2.2. L'agent et le système ..... 107
2.2.1. Le système en philosophie des techniques ..... 108
2.2.2. Le système comme environnement ..... 110
2.2.3. De l'environnement à la simulation ..... 114
2.3. La simulation et la connaissance ..... 118
2.3.1. Interfaces entre théorie et expérience ..... 120
2.3.2. Émergence de sens ou simulacre ? ..... 124
2.3.3. Métaphore, formalisme et efficacité ..... 135
2.3.4. La SBA dans le temps et l'espace ..... 139
Chapitre III. Les systèmes multi-agents et l'éthique ..... 146
3.1. Éthique et moralité sous le signe de l'altérité ..... 146
3.2. La téléologie en SMA ..... 149
3.2.1. La question de la motivation ..... 149
3.2.2. Le jugement éthique ..... 152
3.2.3. Réputation et confiance ..... 157
3.2.4. SMA et négociation ..... 160
3.2.5. Les organisations et leur éthique ..... 162
3.3. La déontologie en SMA ..... 163
3.3.1. L'internalisation de la norme ..... 164
3.3.2. L'épreuve de la norme ..... 168

3.3.2.1. La norme comme contrainte ..... 172
3.3.2.2. La norme comme contrat : sanctions et répercussions ..... 175
3.3.2.3. La norme comme indication d'ordonnancement ..... 177
3.3.3. Création de la norme ..... 179
3.3.3.1. Création statique ..... 180
3.3.3.2. Création dynamique descendante ..... 181
3.3.3.3. Création dynamique ascendante ..... 182
3.3.4. Diffusion de la norme ..... 184
3.3.5. Les organisations et leurs rôles ..... 186
3.4. Études de cas ..... 188
3.4.1. La tentation de l'intelligence centrale ..... 188
3.4.2. Simulation à base d'agents et prise de décision ..... 200
3.4.2.1. Processus décisionnel ..... 200
3.4.2.2. Critères pour une modélisation cognitive ..... 201
3.4.2.3. Critères pour une modélisation axiologique ..... 204
3.4.2.4. Classes d'outils de modélisation cognitive ..... 204
3.4.2.5. La SBA contre la pandémie du Covid-19 ..... 207
3.4.2.6. La SBA à l'œuvre dans la modélisation d'accompagnement ..... 211
3.4.2.7. Forces et limites de la SBA comme outil d'aide à la décision ..... 217
3.4.3. Les agents sur la route : un cas d'innovation irresponsable ? ..... 220
3.4.3.1 Un accident de la route évitable ..... 220
3.4.3.2. L'automatisation et l'excès de confiance ..... 223
3.4.3.3. La défaillance logicielle ..... 225
3.4.3.4. L'éclairage de la SBA ..... 228
3.4.3.5. Pour une innovation responsable ..... 230
Conclusion ..... 233
Liste des abréviations ..... 244
Bibliographie ..... 246
Table des matières ..... 260