Sous la direction de
Jean-Pierre Béland
et **Georges A. Legault**

Asimov
et l'acceptabilité des robots

À propos

Asimov et l'acceptabilité des robots

Jean-Pierre Béland et Georges A. Legault

Asimov et l'acceptabilité des robots

Nous remercions le Conseil des arts du Canada de son soutien.
We acknowledge the support of the Canada Council for the Arts.

SODEC Québec

Financé par le gouvernement du Canada
Funded by the Government of Canada

Canada

Conseil des arts du Canada

Canada Council for the Arts

Les Presses de l'Université Laval reçoivent chaque année de la Société de développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

Le présent ouvrage a été réalisé pour un projet de recherche sous la direction de Johane Patenaude (chercheuse principale) intitulé *Développement d'un cadre de référence interdisciplinaire de l'analyse d'impact des nanotechnologies et de leur acceptabilité sociale*. Ce projet de recherche est financé par les Instituts de recherche en santé du Canada (43854).

Mise en pages : Diane Trottier

Maquette de couverture : Laurie Patry

Dépôt légal 4e trimestre 2019

ISBN 978-2-7637-4668-5

Les Presses de l'Université Laval
www.pulaval.com

Table des matières

Préface

Jean-Pierre Béland

Tout le questionnement que soulève notre ouvrage *Asimov et l'acceptabilité des robots* me semble toujours d'actualité étant donné que la littérature (nouvelles et romans sur le développement des robots moraux) de ce maître de la science-fiction américaine constitue une remarquable anticipation philosophique des répercussions sur nous du développement de la robotique. Même si les robots humanoïdes actuels ne sont pas encore vraiment au point (intelligence artificielle, motricité) et restent très loin des robots moraux au cerveau positronique imaginés par Asimov, il n'en reste pas moins que sa littérature nous a permis de montrer la complexité des problèmes moraux à travers notre processus d'analyse d'impact et d'acceptabilité des robots. Le lecteur découvrira que notre processus d'analyse a fait ressortir toutes les questions pertinentes au sujet des répercussions de l'humanisation des robots et de la robotisation de l'humain sur les enjeux économiques, environnementaux, éthiques, légaux et sociaux (E3LS) et sur l'identité de la personne humaine. Il découvrira aussi que la Bible des robots d'Asimov, à travers ses récits, nous invite à dépasser les trois lois morales de la robotique en les inscrivant dans une approche éthique pour réfléchir sur des dilemmes auxquels nous devons faire face aujourd'hui. C'est en situant les lois morales de la robotique en fonction des valeurs que nous voulons atteindre dans nos vies individuelles et collectives

que la pensée d'Asimov rejoint tout le courant de l'éthique appliquée qui s'est développé depuis plus de cinquante ans. Ce courant, qui prend racine dans la réflexion sur la responsabilité sociale des chercheurs à la suite de la création de la bombe nucléaire et des divers scandales en recherche sur les humains, pose les bases du développement responsable des technologies issues de la recherche. Asimov, contrairement à d'autres romans de science-fiction, ne nous fait pas la morale. Il illustre les conflits de valeurs (par exemple, celui entre la sécurité et l'autonomie du robot). Et il n'a pas de solutions toutes faites. Le lecteur se verra ainsi invité à dépasser l'approche morale traditionnelle et à développer des choix responsables en contexte. D'autant plus que la faisabilité des robots moraux soumis aux lois de la robotique d'Asimov semble une utopie. Le robot moral doté d'une intelligence artificielle forte, comme Byerley (candidat à la présidence mondiale en 2044), serait la représentation d'un humain idéal sans défauts. Est-il un projet irréalisable? Le lecteur découvrira tout au long de cet ouvrage que l'intelligence artificielle (faible ou forte) ne constitue pas vraiment une solution, de sorte qu'il n'y aura peut-être jamais de robot autonome parfaitement sécuritaire.

Introduction

Jean-Pierre Béland
Professeur en éthique,
Université du Québec à Chicoutimi

Georges A. Legault
Professeur en éthique, Université de Sherbrooke

Isaac Asimov (1920-1992), écrivain américain qui a marqué la science-fiction, est mort depuis plus de vingt ans. Pourtant l'intérêt pour son œuvre est toujours présent, comme en témoignent les activités qui signaleront ce vingtième anniversaire. La science-fiction possède un pouvoir particulier, celui de nous projeter dans le temps et, par l'imaginaire, de tracer le développement technologique ainsi que les réactions humaines à son égard. Avec le temps, la science-fiction peut paraître très déphasée ou, au contraire, être demeurée pertinente pour penser une autre époque. L'œuvre d'Asimov met en scène, à travers diverses nouvelles et plusieurs romans, des robots dont certains ne sont que des machines complexes alors que d'autres sont des humanoïdes. Ces robots qui interagissent avec les humains sur la terre ainsi que dans les colonies permettent à Asimov de soulever divers enjeux éthiques, économiques, environnementaux, légaux et sociaux connus sous l'acronyme E^3LS. Que peut-on tirer d'une lecture d'Asimov aujourd'hui ? Comment cet auteur pensait-il ces enjeux ? Comment en évaluait-il les risques et les impacts ? De plus, en imaginant une morale des robots pour les rendre plus acceptables socialement,

comment voyait-il les enjeux du vivre-ensemble ? Enfin, on peut se demander aujourd'hui, alors que le développement de la robotique et des implants avance à grands pas, si les robots d'Asimov sont ou seront un jour réalisables. Voilà les questions qui amènent deux philosophes et deux physiciens à se rencontrer pour penser l'acceptabilité des robots dans l'œuvre de science-fiction d'Asimov.

Comment cela s'explique-t-il ? C'est simple : au départ, c'est parce que nous faisons partie d'une équipe de recherche (Groupe de recherche interdisciplinaire InterNE^{3}LS[1]) sur les enjeux Nano-E^{3}LS pour l'amélioration humaine. Nos réflexions portent précisément sur la façon de développer une approche interdisciplinaire dans l'analyse globale des impacts sur ces enjeux permettant de penser l'acceptabilité du développement d'un nanocapteur en santé. Mais elles permettent également de porter un regard d'ensemble sur les impacts du processus technoscientifique de l'humanisation du robot, qui constitue un premier pas vers la transformation des humains en organismes cybernétiques ou « cyborgs » pour vaincre la maladie, la vieillesse et la mort[2].

> L'informaticien Ray Kurzweil, théoricien du « transhumanisme » a prédit, avant 2050, l'invention des robots microscopiques pour prolonger la durée de vie des humains en intervenant à l'échelle moléculaire. Il estime également qu'à terme du développement technologique de la robotique l'homme sera pourvu d'implants le reliant en permanence au réseau et pourra télécharger des informations directement dans son cerveau. Le « transhumanisme » anticipe ainsi la création d'une intelligence artificielle supérieure aux hommes, semblable à ce que décrivait la science-fiction d'Isaac Asimov dès 1940. Ce dernier prophétisait, dans son livre *I robot* (*Les robots*), l'arrivée de ce robot intelligent, d'abord comme esclave de l'homme, puis son égal et enfin son maître[3].

La science-fiction d'Asimov sur les robots nous rapproche en quelque sorte de notre contexte de l'éthique du développement des nanotechnologies d'aujourd'hui. Ce développement nous fera entrer dans l'ère des « robots humanisés » (robots dotés d'intelligence artificielle) et des « cyborgs » (humains ayant reçu des puces et des prothèses robotisés). Et il nous invite, dès maintenant, à aller au bout des questions essentielles qu'il suscite au sujet de l'acceptabilité éthique pour y répondre, comme l'annonce le rapport de l'US National Science Foundation, *Ethics of Human Enhancement : 25 Questions & Answers*[4], en 2009. Mais il ne peut y avoir de réflexion éthique au sujet de l'acceptabilité du développement des robots, des cyborgs et des prothèses, sans réflexion sur ce que nous sommes de notre point de vue. Que nous réserve l'avenir des avancées significatives en matière d'intelligence artificielle ?

> Robots, cyborgs, prothèses neuro-électroniques, tout un monde de substituts du corps se développe, se met en place, s'offre à nous. S'imposant à nous, car il faudra apprendre, s'adapter, accepter, « s'y faire », les uns à leurs prothèses, les autres à leurs compagnons robots, d'autres – le plus grand nombre – à ce qui sera implanté dans leur corps et dont ils sauront parfois peu de chose. Et ces machines qui nous croiseront, que nous croiserons, auront des émotions, accentuant le leurre de cette non-frontière entre l'humain et l'artificiel[5].

Le problème des impacts sur nous comme enjeu éthique peut ainsi prendre la forme de plusieurs questions philosophiques pour penser l'acceptabilité de ces robots comme développement technologique à venir. Qu'adviendra-t-il des humains demeurés naturels au milieu des robots humanoïdes et des cyborgs ? Les humains seront-ils condamnés à un état d'esclavage dans une société fondée sur des dominations graduées, mais inflexibles : les robots en bas et les cyborgs en

haut de l'échelle ? Et ces robots humanoïdes devenus autonomes, qu'aurons-nous à espérer ou à redouter d'eux ?

La science-fiction d'Asimov est une projection sur le futur d'une panoplie de questions semblables en face des robots intelligents qui font réfléchir sur leur acceptabilité éthique et sociale. Cette projection nous renseigne à la fois sur les défis éthiques et sociaux à venir de l'humanisation des robots et de la robotisation de l'humain, mais aussi sur la question de savoir comment on pense pouvoir ou ne pas pouvoir y répondre à partir des ressources scientifiques et technologiques que nous avons. Asimov nous semble être un bon pédagogue à travers sa science-fiction pour nous aider à vulgariser nos travaux sur ces questions de l'acceptabilité des robots.

N'avez-vous jamais imaginé, par exemple, que l'humanité n'est plus seule, sans amis ? Maintenant l'homme cherche à disposer d'une panoplie de robots (robot bonne d'enfants, robot-docteur, robot-artiste, etc.) pour l'aider ; des créatures plus robustes que lui-même, plus fidèles, plus utiles et qui lui sont purement dévouées ! N'avez-vous jamais envisagé l'acceptabilité des robots sous ce jour ?

Imaginez toutefois qu'un ordinateur bipède et capable de parler, comme un « robot-Descartes[6] », commence à douter en ayant dans son intelligence artificielle la question « Qui suis-je ? » Mais, à vos yeux, un robot est un robot. Ce robot l'accepterait-il ? Et vous, accepteriez-vous qu'il vous contredise en disant qu'il pense mieux qu'un humain sur cette question ?

La science-fiction d'Asimov nous devance en nous faisant envisager une telle situation du vivre-ensemble avec des robots humanisés qui deviennent de plus en plus intelligents en raison des progrès de la science et de la technologie. Mais quelles réponses pourrions-nous y apporter ? Ne faut-il pas

trouver un équilibre entre les réponses qui tombent parfois dans un excès d'optimisme scientiste et d'autres qui expriment toutes les craintes, voire les phobies possibles qui tombent dans l'excès du catastrophisme ?

Dans sa préface à *David Starr, justicier de l'espace* (1978), Asimov a ainsi défini son genre de « science-fiction comme la branche de la littérature qui se soucie des réponses de l'être humain aux progrès de la science et de la technologie[7] ».

Les robots issus du progrès de la science et de la technologie sont-ils acceptables sur terre ? Les accepterons-nous ? La science-fiction d'Asimov se préoccupe des réponses par le fait qu'elles peuvent être multiples et contradictoires. Asimov n'était pas en ce sens un activiste nécessairement convaincu que la science-fiction devrait prendre des positions pro-amélioration de l'humain par l'humanisation des robots en les rendant toujours plus intelligents (il s'agit des positions transhumanistes aujourd'hui). Car que dit la nouvelle *L'homme bicentenaire* ? Le robot Andrew ne voulait-il pas se dé-robotiser en voulant devenir aussi libre et mortel qu'un humain ?

Mais Asimov refuse en même temps les positions des prophètes de malheur qui tombent dans l'excès du catastrophisme. Pour lui, ils sont les « victimes d'un certain complexe de Frankenstein[8] », en voulant à tout prix considérer les robots comme des créatures sans âme et mortellement dangereuses, comme moralement inacceptables dans la pièce de Capek, d'où vient d'ailleurs le mot *robot* :

> Il s'agissait de la pièce R.U.R., de l'auteur dramatique tchèque Karel Capek. Écrite en 1921, elle fut traduite en anglais en 1923. R.U.R. signifiait Rossum Universal Robots (Robots universels de Rossum). Comme Frankenstein, Rossum avait découvert le secret de fabriquer des hommes artificiels. On les appelait « robots », d'après un mot tchèque signifiant travailleur.

> Les robots étaient conçus, comme l'indique leur nom, pour servir de travailleurs, mais tout se gâte. L'humanité ayant perdu ses motivations cesse de se reproduire. Les hommes d'État apprennent à se servir des robots pour la guerre. Les robots eux-mêmes se révoltent, détruisent ce qui subsiste de l'humanité et s'emparent du monde[9].

La position d'Asimov n'est pas simple sur cette question de l'acceptabilité des robots. N'est-elle pas plus complexe qu'on le pense ? Elle cherche un juste milieu. Car, d'une part, la science-fiction d'Asimov n'était pas contre la science. Elle en découle autant qu'Asimov s'intéressait à la science (lui-même diplômé en biochimie, il obtient son doctorat en 1948). Mais, d'autre part, il s'inscrit parmi les écrivains humanistes qui réfléchissaient constamment sur les rapports entre science et littérature, science comme question et culture comme réponse[10]. Puisque sa science-fiction est humaniste, Asimov se préoccupe des risques et des impacts possibles que produisent les progrès scientifiques et technologiques, depuis la Deuxième Guerre mondiale. C'est pourquoi sa science-fiction sur les robots dépend du développement de la science et de la technologie (propre à la compagnie U.S. Robots) qui s'intéresse surtout à montrer l'impact positif des robots. Mais, puisqu'elle montre qu'ils ne sont pas sans risques (le risque zéro n'existe pas !), Asimov cherche à trouver une piste de solution (remède) :

> Le savoir a ses dangers, sans doute, mais faut-il pour autant fuir la connaissance ? Sommes-nous prêts à remonter jusqu'à l'anthropoïde ancestral et à renier l'essence même de l'humanité ? La connaissance doit-elle être au contraire utilisée comme une barrière contre le danger qu'elle suscite[11] ?

Quel est le remède ? Asimov, en tant que scientifique et humaniste dans sa science-fiction, se définissait lui-même comme le « bon docteur » (*Good doctor*), surnom qui sera souvent repris pour le désigner. Sa science-fiction propose

en quelque sorte comme réponse un remède moral à ce danger que représente l'avenir du vivre-ensemble avec des robots qui risquent toujours de se tourner (consciemment ou non) contre leurs créateurs : c'est le concept des « Trois Lois de la robotique » qui réifient toujours le comportement de ces robots en des créatures dociles et entièrement au service des humains, comme Asimov l'explique lui-même dans la préface au recueil *Le robot qui rêvait* :

> J'ai commencé à faire des robots les héros de mes nouvelles en 1939 ; j'avais dix-neuf ans et, dès le début, je les ai imaginés comme des appareils soigneusement construits par des ingénieurs, dotés de systèmes de survie spécifiques que j'ai appelés « Les Trois Lois de la robotique ». Incidemment, je fus le premier à employer ce mot, dans le numéro d'*Astounding Science Fiction* de mars 1942[12].

Exposées pour la première fois dans sa troisième nouvelle, *Cercle vicieux* (*Runaround*, 1942), mais annoncées dans quelques histoires précédentes, ces Trois Lois de la robotique sont :

1) Un robot ne peut porter atteinte à un être humain, ni, restant passif, laisser cet être humain exposé au danger.
2) Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres sont en contradiction avec la Première Loi.
3) Un robot doit protéger son existence dans la mesure où cette protection n'est pas en contradiction avec la Première et/ou la Deuxième Loi[13].

Au cours du cycle des robots, une Loi Zéro, qui prendra une importance plus grande que la Première Loi, sera instituée par deux robots humanoïdes, R. Giskard Reventlov et R. Daneel Olivaw, dans *Les robots et l'empire* (1986). Cette Loi Zéro est autoproduite (non programmée) puisqu'elle est déduite par le robot R. Giskard Reventlov. Elle placera ou

tentera de placer l'humanité avant celle d'un individu. Cette nouvelle Loi s'énonce de la façon suivante :

0) « Un robot ne doit causer aucun mal à l'humanité ou, faute d'intervenir, permettre que l'humanité souffre d'un mal. »

Ce concept de la morale robotique comme réponse forme un principe d'organisation et un thème unifiant de toute son œuvre de science-fiction[14], apparaissant dans son cycle des robots (*1. Les robots*, *2. Un défilé de robots*, *3. Les cavernes d'acier*, *4. Face aux feux du soleil*, *5. Les robots de l'aube*, *6. Les robots et l'empire*) et d'autres recueils comme *L'homme bicentenaire* et *Le robot qui rêvait*.

Mais Asimov nous prévient en disant que ce concept de la morale des Trois Lois de la robotique ne flotte pas non plus dans la pure abstraction littéraire : « la direction du mouvement [de la robotique] est nette[15] » :

> Les robots primitifs qu'on utilise ne sont pas les monstres de Frankenstein de la vieille science-fiction primitive. Ils n'aspirent pas à la vie humaine (bien que des accidents puissent se produire avec des robots, tout comme avec des automobiles ou d'autres machines électriques). Il s'agit plutôt d'appareils conçus avec soin pour éviter aux humains des tâches ardues, monotones, dangereuses et ingrates ; et, donc, par l'intention et la philosophie, ils constituent les premiers pas en direction de mes robots imaginaires.
>
> L'avenir devrait nous permettre d'aller plus loin dans la direction que j'ai indiquée[16].

Ainsi les robots dotés d'une intelligence morale sont devenus aujourd'hui un domaine d'étude reconnu, et c'est le néologisme qu'Asimov a inventé pour forger son concept en 1942 et qui désigne la « robotique ». La robotique au sens asimovien peut se définir comme l'ensemble des études scientifiques et des techniques permettant l'élaboration

d'automatismes du robot moral pouvant se substituer à l'homme pour effectuer certaines opérations, et capables d'en modifier lui-même le « cycle » et d'exercer un certain choix éthiquement acceptable. Par exemple, au Japon, le robot nommé **Asimo** en hommage à Asimov est un robot humanoïde bipède en développement par Honda. Des chercheurs de la Waseda's Graduate School of Advanced Science and Engineering de Tokyo ont aussi produit un robot humanoïde (nommé **Kobian**) qui peut exprimer des simulacres d'émotions humaines, comme le plaisir, la surprise, la tristesse et l'aversion. Selon les concepteurs de **Kobian**, l'expressivité de ce robot le rend à même d'interagir avec les humains et de les assister dans le quotidien[17]. Ces deux robots remettent à l'ordre du jour la science-fiction d'Asimov. Est-ce un mythe ou une réalité ?

En somme, son concept des Trois Lois de la robotique n'est pas une pure vue de l'esprit déconnecté de la réalité scientifique, puisqu'il oriente le mouvement de la robotique actuel vers un idéal moral à venir. Mais une grave difficulté scientifique ressort de cette morale conceptuelle dans la science-fiction d'Asimov. Et c'est le défi actuel du développement de la science et de la technologie de surmonter cette grave difficulté, comme le signifie Georges Vignaux, dans *La chirurgie moderne ou l'ivresse des métamorphoses. La chirurgie esthétique. La chirurgie réparatrice. Les prothèses et les robots*, en 2010 :

> La difficulté majeure, c'est que si l'on veut des robots capables de faire des choix éthiques, il faudra leur fournir une sorte d'« âme », les humaniser davantage. Et même si on y parvient, d'autres questions surgiront. Un robot peut-il avoir la capacité de désobéir aux ordres d'un contrôleur humain s'il décide de ne pas envoyer un missile sur une maison parce que son analyse lui fait conclure que le nombre de civils à l'intérieur dépasse largement le nombre de combattants ennemis ? Un

être moralement autonome est souvent imprévisible en effet : il estime lui-même ce qui lui paraît juste[18].

Voilà pourquoi, dans cet essai intitulé *Asimov et l'acceptabilité des robots*, ce qui nous intéresse en tant que chercheurs (philosophes et physiciens), c'est de voir comment Asimov met en scène, dans la somme de ses nouvelles et romans, le questionnement philosophique et éthique sur le développement de la technologie du robot moralement acceptable, d'abord comme esclave de l'homme (selon les Trois Lois), mais qui devient son égal et enfin son maître (selon la Loi Zéro), pour sensibiliser à l'acceptabilité des robots.

QUELS SONT NOS OBJECTIFS ?

Dans cet essai, nous utiliserons le cadre de référence des enjeux E^{3}LS (enjeux économiques, environnementaux, éthiques, légaux et sociaux) qui sont reconnus en éthique du développement des nanotechnologies au niveau provincial[19], et même mondial. Ce cadre nous permettra d'atteindre les trois objectifs suivants qui correspondent aux trois axes de questions au sujet de l'acceptabilité des robots chez Asimov :

1) Montrer en quoi la science-fiction d'Asimov permet de situer les impacts sur les enjeux E^{3}LS du vivre-ensemble avec des robots que les progrès technologiques d'aujourd'hui laissent présager quant à leur acceptabilité.
2) Montrer en quoi la morale de la robotique d'Asimov permet de comprendre les défis de construire la raison pratique d'un robot et aussi de clarifier nos propres jugements sur l'acceptabilité éthique du développement technologique.
3) Montrer ce que l'on peut retenir de la faisabilité du robot qui applique cette morale pour penser l'acceptabilité du développement technologique aujourd'hui.

PLAN

Pour les enjeux E^{3}LS, nous procéderons à l'analyse et à la clarification conceptuelle des impacts et de l'acceptabilité des robots comme développement technologique en suivant l'un à la suite de l'autre les trois grands axes de questions constituant la problématique chez Asimov : 1) la question du vivre-ensemble avec des robots ; 2) la question de la moralité des robots ; 3) la question de la faisabilité du robot moral pour aujourd'hui. D'où la constitution des trois chapitres suivants.

Le premier chapitre, Vivre-ensemble avec des robots, soulève la question complexe de l'acceptation ou de l'acceptabilité de l'humanisation des robots et de la robotisation de l'humain. Dans un premier temps, nous explicitons les notions d'acceptation et d'acceptabilité des risques et des impacts sur les enjeux. Quelles distinctions faisons-nous entre acceptation, acceptabilité sociale et acceptabilité des impacts sur les enjeux E^{3}LS ? Nous traiterons cette question dans le but de montrer par la suite comment Asimov situe la question de l'acceptabilité des robots en tentant de trouver un équilibre entre, d'une part, la position pessimiste de refus (inacceptation) des fondamentalistes qui tombent dans l'excès du catastrophisme en faisant l'analyse des risques et, d'autre part, la position optimiste de la compagnie U.S. Robots qui veut forcer l'acceptation en mettant l'accent sur les impacts positifs. Nous verrons que la position d'Asimov sur l'acceptabilité cherche un juste milieu. Dans un second temps, nous présentons le processus de l'analyse globale des impacts (impacts positifs et négatifs) sur les enjeux E^{3}LS, dans le but de l'utiliser par la suite pour traiter de la question des impacts et de l'acceptabilité du développement des robots dans la science-fiction d'Asimov. Enfin, dans un troisième temps, nous analyserons les risques qui remettent en question notre identité en tant qu'êtres humains dans le contexte de l'humanisation de la machine (les robots

humanoïdes) et de la mécanisation (robotisation) de l'humain chez Asimov.

Le deuxième chapitre, La morale des robots, soulève la complexité de la problématique morale du robot. Dans un premier temps, nous posons la question « Pourquoi une morale robotique ? » dans le but d'introduire à la morale et aux lois de la morale (énoncé, autorité de l'obligation et application) pour assurer le vivre-ensemble, et puis de faire comprendre, en ce sens, la morale des Trois Lois de la robotique chez Asimov. Dans un second temps, nous présentons la conception de la morale humaine chez Asimov (diversité des principes moraux dans l'œuvre et la conception sociologique de la morale) et les liens entre la morale des robots et la morale humaine : à quels principes moraux renvoient les Quatre Lois : *Primum non nocere*, morale de l'obéissance, altruisme et utilitarisme. Dans un troisième temps, il sera question de la raison pratique (en quoi est-elle importante dans la compréhension de la morale pratique ?), dans le but de montrer quelles difficultés de la raison pratique soulève la morale de la robotique d'Asimov : la difficulté de la Loi 1, la difficulté de la Loi 2, la difficulté de la Loi 3, et la création de la Loi Zéro pour résoudre les insuffisances des trois premières Lois.

Le troisième chapitre, Réaliser des robots éthiques, soulève la complexité de la question de l'acceptabilité des robots moraux pour nous : que peut-on retenir de l'œuvre d'Asimov pour penser le développement technologique aujourd'hui ? La principale question est toujours celle de la faisabilité : est-il seulement possible d'appliquer les Lois de la robotique à une intelligence artificielle ? Les robots moraux, comme Daneel et Giskard, sont ou seront-ils un jour réalisables ? Disposons-nous, actuellement, des moyens pour faire des robots éthiques ?

Dans cet essai, nous écrirons des analyses, en suivant chacun de ces trois grands axes de questionnement lié à l'œuvre de science-fiction d'Asimov, dans le but d'en cerner les enjeux servant à penser le plus finement possible l'acceptabilité ou non des robots humanisés.

Le grand enjeu, pour nous, n'est pas alors de résoudre le problème de l'acceptabilité éthique et sociale des robots dans l'œuvre d'Asimov, mais de découvrir ce qui se passe dans le déroulement de ce processus, plutôt que de défendre une position. C'est, en d'autres termes, d'offrir la possibilité d'un vrai choix libre et éclairé sur la question de l'acceptabilité éthique et sociale du robot.

1
Vivre-ensemble avec des robots
Qu'est-ce que la science-fiction d'Asimov nous raconte sur l'acceptabilité des impacts ?

Jean-Pierre Béland
Professeur en éthique,
Université du Québec à Chicoutimi

Vivre-ensemble avec des robots suppose au départ que nous ayons besoin d'eux dans une société et que les robots seront sécuritaires. La majorité des personnes aujourd'hui ne voient pas de problèmes d'acceptation des robots en voie de développement tels que nous les connaissons actuellement dans l'industrie automobile, dans l'aviation, dans les salles de chirurgie, dans les projets des robots parlants. Cette vision se situe aux antipodes du monde de la science-fiction d'Asimov. Il y a sans doute un grand écart entre notre acceptation générale du développement technologique des robots et la résistance des Terriens chez Asimov. Or, justement, cet écart entre la science-fiction et la réalité peut nous permettre d'entrer dans la complexité des questions au sujet de l'acceptabilité des impacts de ce développement technologique. Vivre-ensemble avec des robots peut-il transformer nos façons d'agir, notamment les relations personnelles avec des robots (amitié, dépendance, etc.) ? Étant donné que les robots deviendront de plus en plus humanoïdes, la

distinction humain-robot aura-t-elle encore un sens ? N'y a-t-il pas un problème d'identité qui s'annonce, en ce sens que la frontière entre l'humain et le robot tend à s'estomper ? Qu'est-ce que l'humain en face du robot humanoïde qui ressemble de plus en plus à l'humain ? Quelles peuvent être les conséquences sociales de l'utilisation de tels robots pour des sociétés qui en dépendent ? Quels clivages peuvent exister entre des sociétés sans robots et des sociétés avec robots ?

Le but du présent chapitre est d'aider le lecteur à mieux comprendre ces questions en nous servant de la science-fiction d'Asimov. Plutôt que de répondre à chacune de ces questions de façon isolée, nous utiliserons le processus d'analyse globale d'impact et d'acceptabilité que nous développons actuellement dans nos travaux InterNE^{3}LS[1], qui portent sur la conception et la fabrication d'un capteur de pression à base de nanotubes de carbone qui sera utilisé pour des soins de santé. Le lecteur pourra ainsi, d'une part, mieux comprendre ce processus d'analyse globale d'impact et d'acceptabilité du développement technologique sur les enjeux E^{3}LS (enjeux économique, environnemental, éthique, légal et social) et, d'autre part, voir comment la science-fiction d'Asimov nous permet de penser l'intégration des robots dans notre société.

Le processus d'analyse globale d'impact et d'acceptabilité suppose une compréhension du concept d'acceptabilité tel que nous le définissons dans nos travaux. Dans les textes, nous trouvons les notions clés d'*acceptation*, d'*acceptabilité* et d'*acceptabilité sociale* qui sont souvent confondues. Une fois que nous aurons clarifié ces notions, nous pourrons voir comment Asimov traite la question de l'acceptation des robots comme développement technologique dans la société et sa position par rapport au problème de l'acceptabilité des impacts et des risques.

1. ACCEPTATION ET ACCEPTABILITÉ

1.1 Explicitation des notions d'*acceptation* et d'*acceptabilité* (risques et impacts)

Comment présenter les notions d'acceptation et d'acceptabilité par rapport aux risques et impacts sur les enjeux du vivre-ensemble avec des robots ? Nous pouvons définir différemment les trois notions suivantes : *acceptation*, *acceptabilité sociale* et *acceptabilité*.

Acceptation

Selon le *Petit Robert* (2012), le terme « acceptation » renvoie au fait d'accepter. Cette définition très générale souligne l'état de fait d'un accord ou d'un consentement que l'on constate. Les gens acceptent ou non, la société tolère ou non. Par exemple, la majorité des personnes vivent avec des ordinateurs et des cellulaires.

Le développement technologique se fait souvent en exagérant l'acceptation des produits offerts au public, sans poser des questions sur les impacts de ces produits lors de leur usage. Par exemple, la question de savoir si le cellulaire peut causer le cancer s'est posée seulement après qu'il eut été accepté et utilisé dans la société.

Plusieurs types de robots dans le monde de l'automobile, le domaine médical et le domaine militaire sont déjà utilisés (acceptés) actuellement. D'autres domaines cependant refusent (n'acceptent pas) ce développement technologique. Lorsque la télévision est apparue dans les années 1950, certains refusaient ce produit technologique en évoquant la question des impacts sur les relations familiales. La même situation d'inacceptation s'est reproduite de nos jours avec le développement technologique des organismes génétiquement modifiés (OGM), le clonage, les cellules souches, les hybrides humain-animal, etc.

Il y aura toujours des gens qui acceptent inconditionnellement le développement technologique et d'autres qui ne l'acceptent pas. Il peut y avoir ainsi un conflit incessant entre les technophiles et les technophobes.

Acceptabilité sociale

En 2009, d'après F. Terrade et ses collaborateurs, dans *L'acceptabilité sociale : la prise en compte des déterminants sociaux dans l'analyse de l'acceptabilité des systèmes technologiques*, les modèles classiques de l'étude des usages (« modèle d'acceptation des technologies ») sont à l'origine de la notion de l'acceptabilité sociale d'une nouvelle technologie ou d'un nouveau procédé qui s'inscrit dans le développement des technologies :

> Ainsi, à croire le ministère délégué à la recherche, « la maîtrise des usages est un enjeu majeur pour l'économie et la société : les technologies ne seront motrices d'un développement économique durable que si l'usage qui en est fait est observé et pris en compte. Comprendre les conditions de l'appropriation des technologies par la société est devenu un facteur essentiel de compétitivité » (www2.enseignementsup-recherche.gouv.fr/technologie/techsociete/index.htm). Il est donc crucial de disposer de quelques modèles pour répondre aux deux questions suivantes :
>
> 1) Qu'est-ce qui fait que nous utilisons une nouvelle technologie ou un nouveau procédé ?
> 2) Comment prédire l'utilisation qui sera faite d'une nouvelle technologie mise à disposition des utilisateurs ?
>
> Dans le domaine de l'utilisation des technologies, ce type d'études relève de ce qu'il est commun d'appeler l'étude des usages. Les enquêtes d'usage ont pour objectif d'appréhender la manière dont les personnes s'approprient et utilisent des produits sur un *continuum* temporel. Dans ce contexte de l'usage d'un produit, d'un service ou d'une technologie, l'étude

de l'acceptabilité renvoie à l'examen des conditions qui rendent ce produit ou service acceptable (ou non) pour l'utilisateur avant son usage réel et effectif (Laurencin, Hoffman, Forest et Ruffieux, 2008). Les études portent ici sur les prédictions qui peuvent être faites concernant l'usage d'un produit avant sa mise en service. Mais l'étude des usages peut aussi s'envisager *a posteriori* pour expliquer l'acceptation d'un système technologique[2].

Ce type d'études sur l'acceptabilité sociale cherche à prédire et à maîtriser les conditions d'acceptation pour mettre en place des stratégies de communication afin d'inciter les gens à acheter et à utiliser les produits technologiques. L'acceptabilité sociale est ce qu'on pourrait nommer l'acceptation future d'un produit technologique.

Acceptabilité éthique

La notion d'acceptabilité telle qu'elle se développe en éthique des technologies aujourd'hui s'est forgée à partir des comités d'éthique qui proposent des avis éthiques sur le choix technologique dans la société. Au Québec, la Commission d'éthique de la science et de la technologie (CEST) joue ce rôle d'éclairer le choix social du développement technologique à partir d'une analyse globale d'impacts et d'acceptabilité. Les avis éthiques ont ce but : donner les raisons éclairantes qui justifieraient sur le plan social l'acceptabilité ou non de l'ensemble des risques et des impacts du développement technologique. Parmi les arguments (raisons) utilisés pour justifier le choix social, il y a toujours des jugements de valeur. Une partie cruciale de l'acceptabilité repose sur les jugements de valeur faits par les membres des comités d'éthique. Par exemple, le débat des arguments moraux sur la question de l'acceptabilité des impacts des technologies pour la fabrication des prothèses et des robots (« nanorobots ») qui ont pour finalité des cyborgs est dans l'impasse[3]. Le défi d'une commission d'éthique de la science

et de la technologique est de dépasser l'impasse en conciliant les diverses perspectives[4].

Ayant maintenant précisé ces trois notions d'acceptation, d'acceptabilité sociale et d'acceptabilité, essayons de voir comment Asimov traite la question de l'acceptation des robots comme développement technologique dans la société et sa position par rapport au problème de l'acceptabilité des impacts et des risques.

1.2 Comment Asimov situe-t-il la question de l'acceptabilité des robots ?

La science-fiction d'Asimov évolue dans un décor composé de prises de positions sociales d'acceptation inconditionnelle par la compagnie U.S. Robots et de refus inconditionnel des robots par des fondamentalistes. L'opposition entre les fondamentalistes et l'Industrie U.S. Robots représente cette lutte entre deux groupes sociaux visant à influencer l'acceptation. Mais aucun des clans ne propose une analyse globale d'impacts et d'acceptabilité des robots issus des progrès de la science et de la technologie (progrès de la robotique). Dans ses récits de science-fiction, Asimov se préoccupe de l'ensemble des risques et des impacts du développement technologique des robots et nous permet ainsi de traiter la question de l'acceptabilité.

Dans son autobiographie, Asimov dit que ses « livres tendent à encenser le triomphe de la technologie plutôt qu'à dénoncer ses effets négatifs[5] ». Cela se comprend puisqu'il veut libérer la pensée du scénario pessimiste de l'apocalypse en proposant comme solution le robot moral. Le « cerveau positronique » (ordinateur) de ce robot est construit selon la programmation des Trois Lois de la robotique pour qu'aucun mal ne soit fait à l'être humain. C'est la réponse progressive et optimiste d'Asimov pour éviter de reculer vers un pessimisme sans fin des fondamentalistes (médiévalistes) face à

la science réellement dangereuse, depuis l'éclatement de la bombe atomique. Mais la position optimiste d'Asimov refuse en même temps de faire place « à une confiance irréfléchie dans le progrès et en l'avènement inévitable d'un royaume d'utopie par la science[6] ». L'essence même de sa science-fiction est comme une position médiane qui dépasse l'impasse de prises de position entre l'optimisme aveugle des promoteurs de la compagnie U.S. Robots qui sont « inconditionnellement pour » et le pessimisme des fondamentalistes qui sont « inconditionnellement contre » l'acceptation. Elle se situe entre ces deux positions extrêmes en posant la question de l'acceptabilité des impacts et des risques de vivre-ensemble avec des robots moraux.

Voyons à partir de quels principaux personnages Asimov, dans sa science-fiction, met en scène les positions pour ou contre l'acceptation en vue de leur dépassement pour traiter la question de l'acceptabilité.

1.2.1 Les principaux personnages des romans d'Asimov

Dans la science-fiction d'Asimov, la compagnie United States (U.S.) Robots, fondée en 1982 par Lawrence Robertson, s'impose au XXIe siècle comme le principal constructeur de robots. Les robots produits par l'U.S. Robots sont moraux parce qu'ils disposent d'un cerveau positronique (ordinateur) dans lequel sont obligatoirement gravées les Trois Lois de la robotique. L'entreprise connaît alors un essor rapide sous l'impulsion de personnes talentueuses : Susan Calvin, robopsychologue, Alfred Lanning, directeur de la recherche, Peter Bogert, mathématicien. D'abord, l'entreprise se spécialise dans le travail extra-terrestre en fournissant des robots utiles dans l'extraction de minerai, la surveillance… L'U.S. Robots gagne ainsi le monopole de cette distribution. Les robots fabriqués dans les usines sur terre sont ensuite testés *in situ* par des techniciens spécialisés, comme Michael Donovan et Gregory Powell. Cependant, les contacts des

humains avec les robots sur terre sont fragiles, et l'entreprise doit sans cesse lutter contre le « complexe de Frankenstein », qui caractérise la peur du créateur envers la rébellion possible de la créature. Ce complexe du robot envahisseur ou aliéné (qui est en contradiction avec la morale robotique) sera toujours présent chez les humains sur terre (les médiévalistes, les fondamentalistes et la Société pour l'humanité) qui nourrissent le pessimisme sous l'effet de l'optimisme aveugle de la compagnie U.S. Robots. Les romans d'Asimov (*Les cavernes d'acier*, *Face aux feux du soleil*, *Les robots de l'aube* et *Les robots et l'empire*) mettent surtout en scène Elijah Baley désigné pour faire des enquêtes spéciales sur des meurtres de créateurs de robots (le D^r^ Sarton, le D^r^ Fastolfe) et des roboticides (Jander Panell) qui risquent de compromettre la fragile paix entre Terriens et Spaciens. D'abord, Elijah Baley (dans les trois premiers romans) et ensuite son descendant, D.G. Baley (dans *Les robots et l'empire*) seront aidés pour cela par un robot humanoïde super intelligent (R. Daneel Olivaw) et évolueront sur fond d'émeutes antirobots. C'est d'abord sur la planète Terre (dans *Les cavernes d'acier*), ensuite sur la planète Solaria (dans *Face aux feux du soleil*) et, enfin, la planète Aurora et la planète Terre (dans *Les robots de l'aube*) qu'Elijah Baley et R. Daneel Olivaw vont exercer leur talent d'enquêteur. Dans *Les robots et l'empire*, la mémoire d'Elijah Baley est alors dans le cerveau du robot R. Giskard Reventlov et dans celui de R. Daneel.

Le problème de Baley, dans ses enquêtes, est que les meurtres et les roboticides jettent le trouble dans la paix fragile entre les Terriens de culture médiévaliste (ou fondamentaliste) et les Spaciens de culture scientifique propre aux savants de la compagnie U.S. Robots. Les Terriens sont des humains qui n'acceptent pas de vivre en relation avec des robots dans les cavernes d'acier (comme New York), tandis que les Spaciens sont des humains robotisés qui acceptent

de vivre en relation avec des robots humanisés sur la planète Aurora ou sur la planète Solaria. Mais il y a aussi des Spaciens et des robots dans la zone réservée de Spacetown (à la frontière de New York) sur la planète Terre. Baley (Terrien), Gladia (Solarienne), Fastolfe (Aurorain de la compagnie U.S. Robots) et R. Daneel Olivaw (robot humanoïde) sont alors les principaux personnages qui vont permettre d'articuler une position mitoyenne entre l'acceptation inconditionnelle et le refus inconditionnel des robots. La transformation de Baley dans sa relation avec les robots est au cœur de la compréhension de la position mitoyenne.

1.2.2 Les positions d'acceptation et d'inacceptation

L'objectif ici est de présenter les positions sociales qui entrent en conflit pour l'acceptation des robots : la position pessimiste des Terriens médiévalistes (celle qui refuse le développement technologique des robots sur terre en mettant l'accent sur les impacts négatifs) et la position optimiste de la compagnie U.S. Robots (celle qui veut forcer l'acceptation des robots sur terre en mettant l'accent sur les impact positifs).

La position pessimiste des médiévalistes : le refus général des robots sur la planète Terre

La position des médiévalistes chez Asimov est pessimiste. Elle consiste à mettre l'accent sur le « risque théorique[7] » grave du développement technologique des robots dangereux qui menacent la vie naturelle des humains. Depuis la grande catastrophe menaçant l'humanité, la catastrophe nucléaire qui aurait été inconcevable sans les développements de la science, il n'y a aucune science pour amener les médiévalistes à conclure que les robots ne sont pas sans danger. Sous l'influence de la littérature apocalyptique de Frankenstein, le robot sans âme (sans conscience morale) qui tue son créateur demeure ainsi le thème clé de la science-fiction d'Asimov. Craignant cette fin apocalyptique, les médiévalistes militent

pour le refus général des robots sur la terre dans des conflits électoraux, des histoires de meurtres, des procès.

Pour clarifier que cette position pessimiste des médiévalistes se fonde sur des raisons pour parvenir à convaincre l'adversaire (l'U.S. Robots), nous allons prendre des exemples dans différents contextes de la science-fiction d'Asimov.

Dans *Les cavernes d'acier*, le commissaire principal de police de New York, Julius Enderby, est le modèle du personnage médiévaliste. Il sert à définir l'appartenance d'un individu à cette prise de position pessimiste qui consiste dans le refus général des robots sur la planète Terre. Asimov révèle par ce modèle qu'un médiévaliste peut devenir assez violent, même s'il semble doux et amical de prime abord. Car Baley découvrira à la fin de l'histoire que c'est Julius Enderby qui a tué le docteur Sarton (Spacien) de l'U.S Robots en pensant que ce n'était qu'un robot. Quel était le motif du meurtre ? Les médiévalistes comme Enderby n'hésitaient pas à recourir à la force et au meurtre, pour « se débarrasser des monstres, c'est-à-dire des robots, et aussi des Spaciens[8] ». Ils voulaient éliminer le risque de l'invasion des robots humanoïdes (comme Daneel) dans les cavernes d'acier comme New York. Quelles sont les raisons d'un tel choix ? Il se fonde sur l'apocalypse du nucléaire : dans le développement technologique de leur histoire[9], l'idée d'un cerveau positronique suppose que « l'énergie exigée pour produire des positrons en quantité et l'énergie dégagée quand ils sont détruits en quantité sont effroyables[10] ». À cela s'ajoute l'incertitude face aux garanties posées par les Trois Lois morales de la robotique : « Si l'on tentait de construire des robots qui ne sont pas basés sur les Trois Lois fondamentales[11]. » La différence entre nature humaine et artifice sert alors pour contester le fait qu'un robot d'apparence humaine (robot humanoïde) peut servir à tromper les humains :

> – Non docteur ce n'est pas une supercherie. Dites-moi maintenant autre chose : en construisant un robot aussi humanoïde que celui-ci [comme Daneel], dans le but bien arrêté de se faire passer pour un homme, n'est-il pas nécessaire de doter son cerveau de facultés presque identiques à celles du cerveau humain ? – Certainement. – Parfait. Alors un tel cerveau humanoïde ne pourrait-il pas ignorer la Première Loi[12] ?

C'est pourquoi les médiévalistes adoptent une position pessimiste de refus des robots sur terre. D'autant plus qu'ils sont sous l'influence d'un très puissant complexe : celui de Frankenstein qui est « le nom du héros d'un roman de l'époque médiévale, qui construisit un robot, lequel se retourna contre son créateur[13] ». Alors une seule trame de l'histoire semble désormais possible : « des robots étaient créés et détruisaient leur créateur ; des robots… etc.[14] »

Dans sa nouvelle « Évidence », Asimov appelle aussi ces médiévalistes des « fondamentalistes » dans le but de faire ressortir la raison fondamentale qu'ils ont pour détester les robots et évoluer dans des courses électorales sur fond d'émeutes antirobots ou dans des procès antirobots. Si l'on cherchait à observer le discours d'un fondamentaliste comme Quinn dans sa lutte électorale contre Byerley pour le poste de maire, on verrait justement que la différence entre humain et robot (nature *vs* artifice) est au cœur de la position fondamentaliste. Donc, en soi, un robot humanoïde qui peut tromper les humains ne peut pas par essence gouverner l'humain. Mais accepter cela, c'est aussi mettre l'artifice (ce robot artificiel qui pourrait tromper les humains) comme équivalent à la nature. Cela aide à expliquer pourquoi le seul fait que Quinn accusait Byerley d'être un robot était pour lui une raison suffisante pour qu'il constitue « un danger pour la grande masse d'autres êtres humains que nous appelons la société[15] ». Quinn agissait-il inconsciemment en mettant l'accent sur le risque que Byerley soit un

robot dangereux ? Non. Car, dans le contexte d'une lutte électorale, il n'y avait qu'un scénario possible, lui « gagnant », Byerley « perdant ». En fait, pour gagner, il s'agissait seulement à Quinn de dire que Byerley est un robot, parce que cela provoquait la réaction antirobot des fondamentalistes qui croyaient qu'un robot met nécessairement en danger la vie naturelle de l'homme. Autrement dit, si le juriste Byerley cache sous sa peau un cerveau positronique, une telle intelligence artificielle contredit la morale du cours naturel des choses et de la vie simple. Il est dangereux pour la culture médiévale humaine, parce qu'un tel être artificiel ne peut pas être l'équivalent d'un homme dans le vivre-ensemble.

Le spectre du monstre chez les médiévalistes et les fondamentalistes a toujours ainsi mobilisé les craintes du risque éventuel de la fin de l'humanité (crime contre la vie naturelle, crime contre l'espèce humaine), ou au contraire l'inventivité normative pour obliger à maintenir le refus général des robots en développement que la compagnie U.S. Robots pouvait offrir sur la terre. Dans la nouvelle « Le correcteur », le professeur Goodfellow a ouvert le procès antirobot en rappelant au docteur Lanning de l'U.S. Robots qu'il existe des lois qui interdisent l'usage des robots à la surface de la terre :

> Comme vous le savez, docteur Lanning, il existe des lois qui interdisent l'usage des robots à la surface de la Terre, remarqua-t-il. – Ces lois ne sont pas simples, professeur Goodfellow. Les robots ne doivent pas être employés dans des lieux ou des édifices publics. Ils ne doivent pas être utilisés sur des terrains ou à l'intérieur d'édifices privés, sauf sous certaines restrictions qui correspondent la plupart du temps à des interdictions pures et simples[16].

En somme, tout cela nous aide à comprendre que cette position pessimiste de l'inacceptation d'un produit (robot

moral) en développement par l'U.S. Robots sur terre repose sur une stratégie qui vise :

1) à montrer que la science du robot rend possible une apocalypse semblable au nucléaire ;
2) à montrer l'incertitude face aux garanties posées par les Trois Lois morales de la robotique ;
3) à montrer que l'élimination de la frontière entre le robot humanoïde (l'artifice) et l'humain (la nature humaine) implique que le robot pourrait tromper les humains en ignorant la Première Loi et que les robots pourraient gouverner l'ensemble des activités humaines.

Quel mot clé pourrions-nous retenir pour signifier de tels risques sur lesquels met l'accent cette position pessimiste du refus général des robots sur terre ? Le nom de Frankenstein est resté comme symbole. Les médiévalistes en gardent un très puissant complexe. Ils sont comme « nombre d'adultes – victimes d'un complexe de Frankenstein » – « voulant à tout prix considérer ces robots comme des créatures mortellement dangereuses[17] ».

La position optimiste de la compagnie U.S. Robots : forcer l'acceptation des robots sur la planète Terre

La compagnie U.S. Robots ne considère jamais l'acceptation des robots sur terre comme un état de fait. Mais comment peut-elle forcer l'acceptation ? Elle met l'accent surtout sur les avantages (impacts positifs) des robots moraux pour l'humain. Essayons de clarifier les raisons à la base de cette position en prenant quelques exemples tirés de la science-fiction d'Asimov.

Dans *Les cavernes d'acier*, le docteur Anthony Gerrigel de Washington explique à l'inspecteur Baley que les plus grands savants de l'*Histoire de la robotique* ont construit des robots humanoïdes en les acceptant pendant toute leur vie comme quelque chose de « normal ». Ils ont pu s'opposer aux

fondamentalistes (médiévalistes) en donnant aux robots une forme humanoïde à l'avantage des humains. Quel est le véritable enjeu (raison) de la compagnie U.S. Robots pour forcer l'acceptation en donnant au robot un caractère plus humain ? C'est l'avantage économique (de la morale coût-bénéfice de l'entreprise) qui prédomine :

> – C'est le point de vue économique qui a prévalu et a inspiré les décisions. Voyons Monsieur Baley ! Supposez que vous avez à exploiter une ferme : auriez-vous envie d'acheter un tracteur à cerveau positronique, une herse, une moissonneuse, un semoir, une machine à traire, une automobile, etc., tous ces engins étant également dotés de cerveau positronique ? Ou bien ne préfériez-vous pas avoir du matériel sans cerveau, et le faire manœuvrer par un seul robot positronique ? Je dois vous prévenir que la seconde solution représente une dépense cinquante ou cent fois moins grande que la première.
>
> – Bon ! Mais pourquoi donner au robot une forme humaine ?
>
> – Parce que la forme humaine est, dans toute la nature, celle qui donne le meilleur rendement. Nous ne sommes pas des animaux spécialisés, monsieur Baley, sauf au point de vue de notre système nerveux, et dans quelques autres domaines. Si vous désirez construire un être mécanique, capable d'accomplir un très grand nombre de mouvements, de gestes et d'actes, sans se tromper, vous ne pouvez mieux faire qu'imiter la forme humaine. Ainsi, par exemple, une automobile est construite de manière à ce que ses organes de contrôle puissent être saisis et manipulés aisément par des pieds et des mains d'homme, d'une certaine dimension, et d'une certaine forme : ces pieds et ces mains sont fixés au corps par des membres d'une longueur déterminée et par des articulations bien définies. Les objets, même les plus simples, comme les chaises, les tables, les couteaux, ou les fourchettes, ont été conçus en fonction des dimensions humaines et pour être maniés le plus facilement possible par l'homme. Il s'ensuit que l'on trouve plus

> pratique de donner aux robots une forme humaine que de réformer radicalement les principes selon lesquels nos objets usuels sont créés[18].

L'U.S. Robots met aussi toujours l'accent sur « la conséquence pratique des Trois Lois de la robotique[19] ». Les Lois garantissent les impacts positifs (selon la Loi 1, un robot ne peut pas porter atteinte à un être humain ni permettre que du mal soit fait à un être humain ; selon la Loi 2, il doit obéir aux ordres donnés par un être humain… ; selon la Loi 3, il doit protéger son existence tant que cette protection n'entre pas en contradiction avec la Première ou la Deuxième Loi). Ces impacts positifs garantis par l'U.S. Robots favorisent le point de vue économique des décisions (le profit de la compagnie). Il peut alors arriver que, face à ses propres employés, la compagnie manque de transparence sur les risques probables des robots dont elle modifie parfois dangereusement les Lois. Dans la nouvelle « Le petit robot perdu », le gouvernement obligera les responsables de la compagnie U.S. Robots à maintenir secrète la perte d'un robot Nestor modifié (non conforme à la morale robotique). Car elle veut éviter la controverse avec des radicaux fondamentalistes sur les risques que ce robot blesse un humain : « La seule défense que le gouvernement pouvait opposer aux radicaux fondamentalistes en l'occurrence, c'est que les robots sont toujours construits en vertu de la Première Loi – ce qui les met dans l'impossibilité absolue de molester des êtres humains en quelque circonstance que ce soit[20]. » (La probabilité du risque semble donc inexistante !) Mais, dans la nouvelle, Susan Calvin permettra d'établir l'hypothèse de la relation entre le robot Nestor modifié et le risque probable qu'il puisse faire du mal à un être humain. Elle met elle-même le robot à l'épreuve selon une expérience pour le démontrer étant donné qu'elle « savait », dit-elle, qu'il accepte mal les ordres donnés par un être humain. Sa démonstration a permis de

conclure que le petit robot perdu était dangereux (seule la Loi 1 le retenait difficilement) lorsqu'il « souffrait d'un complexe de supériorité qui ne cessait de croître et de s'amplifier ». Selon cette démonstration scientifique (le risque est avéré), il fallait donc détruire les autres Nestor modifiés : « Les autres Nestor devront naturellement être détruits, dit le Dr Calvin. – Ils le seront. Nous les remplacerons par des robots normaux[21]. »

La nouvelle « Pour que tu t'y intéresses » correspond davantage à la notion d'acceptabilité sociale à partir de laquelle la compagnie U.S. Robots cherche à prédire et à maîtriser les conditions d'acceptation pour mettre en place des stratégies servant à habituer les gens à acheter et à utiliser les produits technologiques. Dans cette nouvelle, la compagnie demande au robot George Dix de l'aider à trouver une stratégie pour surmonter le problème de l'inacceptation des robots. L'entreprise espère que le robot pourra lui apporter un regard nouveau plus éclairant qu'un regard humain afin de résoudre le problème :

> Nous allons vous et moi créer un monde qui va commencer à accepter les robots quels qu'ils soient. L'homme moyen peut avoir peur d'un robot qui ressemble à un homme et qui semble assez intelligent pour le remplacer, mais il n'aura pas peur d'un robot qui ressemble à un oiseau et qui se borne à manger des insectes pour son bien-être à lui. Puis, en fin de compte, quand il aura perdu l'habitude d'avoir peur de certains robots, il n'aura pas peur d'aucun robot. Il aura tellement l'habitude des oiseaux-robots et des abeilles-robots et des vers-robots, qu'un homme-robot ne lui semblera qu'un prolongement des autres[22].

La solution trouvée par George Dix consiste à créer des robots animaux utiles à l'environnement pour forcer l'acceptation progressive des robots humanoïdes : « Il fallait à tout prix que les George et ceux de la même nature et de la même forme qui suivraient dominent. C'était ce qu'imposaient, et

toute autre action était impossible, les Trois Lois de l'Humanique » pour le plus grand bien de l'humanité[23]. Telle est la stratégie adoptée qui est propre à l'analyse de l'acceptabilité sociale des robots.

Tout cela nous aide à comprendre que l'acceptation d'un produit (robot) en développement par l'U.S. Robots repose sur la maximisation de l'intérêt économique et vise :

1) à montrer les avantages d'un produit ;
2) à montrer qu'il y a peu d'inconvénients (diminuer les irritants) ;
3) à trouver une stratégie pour surmonter la peur du robot humanoïde.

Donc, comme toute bonne compagnie, il faut rendre les robots de plus en plus utiles pour les humains et réduire les irritants en les rendant moraux pour le plus grand bien de l'humanité.

Le dépassement de l'impasse entre la position pessimiste et la position optimiste : l'acceptabilité par l'analyse globale des impacts

Baley est sans doute le personnage principal qui représente la position de ceux qui cherchent à dépasser la position pessimiste des médiévalistes et la position optimiste de la compagnie U.S. Robots. Le but est de montrer ici comment Baley a été amené, dans ses enquêtes et dans sa vie, à dépasser sa position médiévaliste de rejet du robot et de sa vie sans robots pour adopter l'idée de travailler avec des robots et d'aller vers la colonisation du projet de Fastolfe.

Dès la situation de départ, dans *Les cavernes d'acier*, Asimov met en scène la position pessimiste et médiévaliste du rejet du robot dans laquelle se trouve Baley. L'histoire commence avec le personnage Baley qui se durcit et qui est fort irrité par le robot R. Sammy qui le prévient que son patron Enderby le demande : « – J'ai dit : entendu ! répéta

Baley. Fous le camp. R. Sammy pivota sur les talons, et s'en fut vaquer à ses occupations ; et Baley, fort irrité, se demanda, une fois de plus, pourquoi ces occupations-là ne pouvaient pas être confiées à un homme[24]. » Baley se rend au bureau de son patron et lui demande aussitôt de ne plus l'envoyer chercher par R. Sammy. Mais il va se sentir plus mal à l'aise encore en apprenant la raison pour laquelle Enderby l'avait fait venir : Enderby l'oblige, non seulement à trouver l'assassin du docteur Sarton en se rendant à Spacetown (qui relève de New York), mais à prendre, pour mener son enquête, un associé spacien : « Alors, Lije, êtes-vous prêts à accepter de prendre avec vous un associé spacien[25] ? » Baley proteste (« Non, monsieur le commissaire. Inutile ![26]… »), mais, en vain, puisqu'il se trouve mis dans la double obligation, celle d'accepter cet associé indésirable, nommé « Robot Daneel Olivaw », et celle d'accepter de l'inviter à habiter dans sa maison.

Telle est la situation originelle de Baley dans sa position pessimiste d'extrême méfiance à l'égard de Daneel au début de l'enquête. Mais la situation d'arrivée de Baley à la fin de l'enquête est qu'il accepte de travailler avec le robot Daneel. Il parvient même à convaincre Enderby (le modèle du médiévaliste) qu'il ne révélera pas que celui-ci est le meurtrier, seulement s'il choisit de défendre avec lui et Daneel l'idée de travailler avec des robots et d'aller vers la colonisation du projet de Fastolfe. Deux raisons fondent cette décision : « La colonisation de l'espace est l'unique voie sur la Terre[27]. » Et la confiance qu'il a maintenant envers Daneel pour marcher dans cette voie : « Je n'aurais jamais pensé qu'un jour je pourrais dire quelque chose de ce genre à une créature telle que vous, Daneel. Mais voilà : j'ai confiance en vous, et même je vous admire[28]. »

Dans *Les robots et l'empire*, Giskard se rappelle l'ensemble des raisons justifiant l'acceptabilité des robots pour aller dans cette direction de l'expansion et de l'évolution, contrairement à la non-expansion et à la décadence :

> – Mais pour quelle raison souhaitez-vous une telle expansion, Baley ? J'ai le sentiment que, sans expansion d'aucune sorte, l'humanité ne peut progresser. Ce ne doit pas être obligatoirement une expansion géographique, mais c'est là la manière la plus évidente de provoquer d'autres expansions, corrélativement. Si l'on peut se lancer dans l'expansion géographique sans que cela se fasse au détriment d'autres êtres intelligents, s'il existe des espaces vides où s'étendre, alors pourquoi pas. S'opposer à l'expansion dans de telles conditions, c'est assurer la décadence.
>
> – C'est donc l'alternative que vous voyez ? L'expansion et l'évolution ? Ou la non-expansion et la décadence[29].
>
> – Oui, je crois. Si, donc, la Terre refuse l'expansion, les Spaciens doivent l'accepter. L'humanité, qu'elle soit spacienne ou terrienne, doit s'étendre. J'aimerais que ce soient les Terriens qui se chargent de cette tâche, mais, à défaut, mieux vaut une expansion spacienne que pas d'expansion du tout. C'est soit l'un, soit l'autre.
>
> – Et si l'expansion est le fait des uns, mais pas des autres ?
>
> – Dans ce cas, la société à l'origine de l'expansion deviendra de plus en plus forte et l'autre de plus en plus faible.
>
> – En êtes-vous certain ?
>
> – Je crois que ce serait inévitable.
>
> – Je le crois aussi, en effet, dit Fastolfe, hochant la tête. C'est pourquoi je tente de persuader et les Terriens et les Spaciens de s'étendre et d'évoluer. C'est là une troisième solution et, à mon sens, la meilleure[30].

En résumé, nous venons de voir comment Asimov articule à travers sa science-fiction les trois notions d'acceptation, d'acceptabilité sociale et d'acceptabilité. Dans ses nouvelles et ses romans, Asimov met en scène différents impacts reliés au vivre-ensemble avec des robots. Si nous voulons avoir une vue générale des impacts et des risques du vivre-ensemble avec des robots, il nous revient de procéder à une analyse globale d'impact et d'acceptabilité.

2. ANALYSE GLOBALE D'IMPACT ET D'ACCEPTABILITÉ

Que faut-il entendre par analyse globale d'impact et d'acceptabilité ? Nous ferons quelques remarques préliminaires avant de procéder à une telle analyse chez Asimov. Le but de cette section est de montrer quelle analyse globale d'impact et d'acceptabilité des robots permet de faire Asimov dans ses nouvelles et ses romans.

2.1 Explication du processus d'analyse d'impact et d'acceptabilité

Depuis 2012, notre groupe de recherche interdisciplinaire utilise une grille conceptuelle en forme de processus réflexif d'analyse d'impact et d'acceptabilité. Ce processus renvoie à trois moments subséquents que nous présentons dans le tableau suivant.

PROCESSUS D'ANALYSE D'IMPACT ET D'ACCEPTABILITÉ

Moment 1 : La détermination des impacts sur des enjeux

Étape 1 : Identification de la source technologique pouvant avoir un impact sur un enjeu

Étape 2 : Identification d'un enjeu pouvant subir un impact de la source

Étape 3 : Détermination de l'impact réel ou négatif de la source sur l'enjeu

Moment 2 : L'évaluation des impacts à partir des valeurs retenus

Étape 1 : Qualification des impacts sur des enjeux en termes de valeur

Étape 2 : Jugement final d'évaluation d'un impact positif ou négatif

Moment 3 : La pondération des jugements finaux d'évaluation en vue de la décision

Étape 1 : Détermination du type de pondération retenue : acceptabilité des risques ou acceptabilité globale des impacts ?

Étape 2 : Processus de pondération a) selon l'acceptabilité des risques, b) ou selon l'acceptabilité globale des impacts

Voyons comment ce processus d'analyse globale d'impact et d'acceptabilité s'applique au capteur de pression à base de nanotubes de carbone à des fins de santé[31].

MOMENT 1 : La détermination des impacts sur des enjeux

Le but est de réaliser des analyses scientifiques permettant 1) de retracer la source technologique pouvant avoir un impact sur un enjeu et 2) de tenir compte du contexte pour valider que cette source technologique pouvant avoir un impact sur un enjeu s'applique dans le cas particulier.

Étape 1 : Identification de la source technologique pouvant avoir un impact sur un enjeu

Dans le présent cas, il s'agit d'un capteur fait de nanotubes de carbone intégré dans une semelle en polymère. Ce produit peut être analysé de différentes façons selon qu'on le considère comme *produit*, *procédé*, *processus de développement technologique* et *usages*.

Étape 2 : Identification d'un enjeu pouvant subir un impact de la source

Chacun des éléments précédents de la source technologique peut avoir des impacts différents sur ce que nous appelons un enjeu. Par exemple, le fait que le capteur soit fait de nanotubes de carbone soulève la question de la toxicité (impact négatif) pour l'humain qui y serait exposé (enjeu sur la santé). Si le capteur peut entraîner la mort, la question de la toxicité peut se poser au sujet la mort (enjeu vie-mort). Elle peut aussi se poser au sujet de l'environnement (enjeu environnemental).

Si l'on regarde le capteur comme faisant partie du processus de développement technologique, la question de l'impact économique (enjeu économique) se pose, tout comme celle de l'impact sur le statut et le développement de la recherche scientifique (enjeu du statut et du développement de la recherche scientifique).

Lorsqu'on regarde le capteur comme un produit fini, il vise à donner des renseignements pouvant prévenir des plaies de pression pour les diabétiques (enjeu santé). Par contre, les renseignements obtenus permettraient aussi de retracer le porteur des semelles (enjeu liberté de choix et enjeu vie privée). Intégrer le capteur dans les soins de santé aura des impacts dans l'institution sur les plans local et national (vivre-ensemble sur les plans local et national) et sur le plan international au sujet de l'accessibilité des semelles dans les pays en voie de développement

(vivre-ensemble sur le plan international). Mais, si le capteur pouvait être incorporé dans le pied du patient, il aurait des impacts sur les représentations culturelles de l'humain (enjeu des représentations culturelles de l'humain : identité, nature, personne).

Étape 3 : Détermination de l'impact réel de la source sur l'enjeu

Lorsque nous faisons une analyse scientifique, nous devons préciser sur quelle étude se base la détermination de l'impact positif ou négatif sur l'enjeu. Nous devons aussi prendre en compte la situation concrète dans laquelle des personnes peuvent être exposées à la source technologique. Dans le cas d'une situation précise, il s'agit d'établir la probabilité que l'exposition aux nanotubes de carbone se produise. Par exemple, si les nanotubes de carbone sont dans une semelle de polymère, quelle est la probabilité qu'une personne soit exposée à ces nanotubes de carbone ? Et, dans le pire des cas, cette exposition est-elle suffisante pour générer l'impact négatif prévu sur la santé ?

MOMENT 2 : L'évaluation des impacts à partir des valeurs retenues

Le but est de poser un jugement de valeur sur les impacts positifs et négatifs retenus dans la détermination des impacts. L'évaluation consiste en un jugement de valeur qui établit pour chaque impact positif ou négatif le niveau de maximisation ou de minimisation de la valeur qui y est associée. L'évaluation des impacts varie d'une personne à l'autre.

Étape 1 : Qualification des impacts sur des enjeux en termes de valeur

L'identification d'un enjeu dans la détermination des impacts n'est pas neutre. Mais elle indique par définition que la dimension particulière de notre vie qui peut subir un impact est importante pour nous individuellement ou

collectivement ; autrement dit, que cette dimension a de la valeur, si par « valeur » nous entendons ce qui fait l'objet d'une préférence, ce qui est estimé, préféré ou désiré.

On peut associer à chacun des enjeux retenus la valeur qu'on y trouve. Généralement, il s'agit de la qualité. Par exemple, pourquoi l'enjeu de la santé est-il pour nous si important ? Parce que nous cherchons à maximiser la qualité de la santé pour chacun de nous. Par exemple, si la semelle contenant un capteur composé de nanotubes de carbone vise à améliorer la santé des diabétiques, elle cherche donc à maximiser la qualité de la santé pour le diabétique. Par contre, si le diabétique est exposé à la toxicité (impact) des nanotubes de carbone, cela minimise la qualité de la santé humaine.

Donc, lorsque la Commission de l'éthique de la science et de la technologie du Québec se prononce, elle essaie de faire un consensus entre les participants sur les valeurs qui sont associées aux enjeux généraux et qui renvoient plus précisément à la qualité de ce qui était visée :

i) qualité de la santé humaine
ii) qualité de la vie-mort pour les personnes
iii) qualité de l'environnement
iv) qualité des retombées économiques
v) qualité du développement de la recherche scientifique
vi) qualité de l'autonomie (liberté de choix) de la personne
vii) qualité de la vie privée
viii) qualité du vivre-ensemble (relations des personnes dans un État)
ix) qualité de vivre-ensemble (relations internationales)
x) valeur de nos représentations culturelles de l'humain (identité, nature, personne)

Étape 2 : Jugement final d'évaluation d'un impact positif ou négatif

Le jugement final d'évaluation consiste justement à indiquer le degré de maximisation ou de minimisation de la valeur engendré par l'impact positif ou négatif. Si l'on se fie exclusivement à notre désir, nous voulons tous la qualité totale : qualité totale de notre santé, qualité totale des relations amoureuses, qualité totale de notre bien-être économique… autrement dit, nous désirons tous le Paradis terrestre perdu.

L'évaluation va varier entre les personnes selon le degré d'atteinte de la valeur envisagée. Par rapport à la qualité totale, le moindre impact sera vu comme minimisant la valeur. Pour une autre personne qui vise une qualité moyenne, le même impact que le précédent sera vu comme n'affectant pas ou peu la valeur.

Par exemple, si l'on prend le capteur, des personnes (se basant sur les mêmes études d'impacts) jugeront que l'impact réel minimise ou maximise à un certain degré (peu, moyennement, beaucoup, presque totalement) la qualité de la santé.

MOMENT 3 : La pondération des jugements finaux d'évaluation en vue de la décision

Étape 1 : Détermination du type de pondération retenue : acceptabilité des risques ou acceptabilité globale des impacts ?

Le but en fin de processus consiste à établir l'acceptabilité de la source technologique (produit, procédé, processus de développement et usage) selon le poids accordé aux jugements finaux d'évaluation de tous les impacts positifs ou négatifs. Il existe deux types de pondération de ces jugements finaux d'évaluation : la pondération des risques lorsqu'il s'agit exclusivement de se prononcer sur l'acceptabilité du risque (impact négatif) et la pondération de

l'ensemble des impacts, lorsqu'il s'agit de se prononcer sur l'acceptabilité globale des impacts (positifs et négatifs).

Étape 2 : Processus de pondération
a) selon l'acceptabilité des risques, b) ou selon l'acceptabilité globale des impacts

La pondération peut se faire : a) selon l'acceptabilité des risques, ou b) selon l'acceptabilité globale des impacts.

a) Selon l'acceptabilité des risques : la pondération des jugements de valeur sur des impacts négatifs se fait à partir d'un standard défini et justifié qui permet de statuer sur son acceptabilité dans la mesure où l'évaluation démontre que le standard est respecté.
    i) Pondération à partir des standards scientifiques (exemples : équivalence en substance, tolérance d'exposition au mercure).
    ii) Pondération à partir des standards moraux philosophiques[32] :
        - En raison de la conformité ou de la non-conformité à la nature humaine (métaphysique classique)
        - En raison de la conformité ou de la non-conformité avec la dignité humaine (Kant)
        - En raison de la conformité ou de la non-conformité avec la vie bonne (Aristote)
        - En raison de la conformité ou de la non-conformité avec l'autonomie de la personne de choisir les risques auquel elle se soumet (libertarien)
    iii) Pondération à partir de standards sociaux (mœurs) : acceptation sociale (état de fait actuel ou projeté que la grande majorité de la population accepte ou accepterait ou refuse ou refuserait d'assumer le risque).

b) Selon l'acceptabilité globale des impacts : les modèles usuels de pondération des évaluations finales[33] sont :

- Analyse coûts/bénéfices économiques
- Argument moral de l'utilitarisme (maximisation du bien du plus grand nombre)
- Argument moral de l'équité (traiter toutes les personnes de la même catégorie de la même manière)
- Argument moral des droits dans la démocratie (le bien commun est déterminé par les institutions démocratiques)
- Argument basé sur les mœurs actuelles ou futures relatives à l'acceptation sociale (l'acceptation par les individus)
- Valeurs jugées prioritaires pour assurer le vivre-ensemble.

2.2 Analyse globale d'impact et d'acceptabilité chez Asimov

Dans ce qui suit, nous déterminerons quels sont les impacts et les jugements d'acceptabilité que nous pouvons retenir chez Asimov. Nous suivrons l'un à la suite de l'autre les trois moments de ce processus d'analyse globale d'impact et d'acceptabilité tel que nous venons de le définir ci-dessus.

MOMENT 1 : La détermination des impacts sur des enjeux

Étape 1 : Identification de la source technologique pouvant avoir un impact sur un enjeu

Le but est de montrer qu'il y a deux sources technologiques différentes qui peuvent avoir des impacts sur des enjeux : la première source consiste dans tous les robots servant à différents usages et la deuxième source est le mouvement de robotisation de l'humain.

1) Les robots servant à différents usages

La science-fiction d'Asimov met en scène une panoplie de robots qui servent à des usages différents :

- les robots Speedy, Cutie, Dave, Cerveau, Al-76, ZZ, Parsec, etc., servent à développer le marché extraterrestre ;
- le robot Robbie sert comme bonne d'enfant ;
- le robot Tony sert pour des relations amoureuses ;
- le robot Byerley sert à coordonner les machines ;
- le robot Lenny sert à développer la connaissance pédagogique (l'éducation des robots-enfants) ;
- le robot Easy sert à faire des corrections de travaux ;
- le robot Daneel (tout comme Giskard) sert à faire des enquêtes sur des meurtres et à protéger les humains dans la réalisation du projet de Fastolfe ;
- le robot Sally sert d'automobile ;
- le robot Andrew sert de valet, de maître d'hôtel et de femme de chambre ;
- le robot Winkler sert de sosie du président des États-Unis ;
- le robot Jane 1 sert à mieux comprendre le cerveau intuitif humain ;
- le robot Max sert d'artiste ;
- le robot Multivac avec ses millions de robots sous ses ordres sert de juge pour résoudre tous les problèmes humains ;
- les robots George servent à faire accepter les robots comme des êtres humains aux bénéfices de la compagnie.

Dans les nouvelles et les romans, ce qui intéresse Asimov, ce sont les robots servant à des usages. Asimov mentionne

que les procédés de fabrication de tels robots et le processus de développement des robots et de leur cerveau positronique peuvent s'identifier aux procédés techniques et industriels de la miniaturisation de nos ordinateurs aujourd'hui[34]. Mais il n'en fait pas la source d'une analyse des impacts.

2) Le mouvement de robotisation de l'humain

Dans le roman *Face aux feux du soleil*, la robotisation de l'humain renvoie à la ferme des fœtus (un mois après leur conception) sur la planète Solaria :

> En tant que fœtologues, Baley, nous devons nous préoccuper de créer des enfants sains. Je répète sains. Même l'analyse la plus poussée de chromosomes du père et de la mère ne peut assurer une combinaison spécifiquement favorable à tous les gènes, sans parler des risques de mutations imprévisibles[35].

Le but de cette robotisation par la sélection des fœtus est de « maintenir une population régulière en se fondant sur une espérance de vie de trois cents ans pour vingt mille habitants[36] ».

La robotisation de l'humain renvoie aussi au rêve de fusion C/Fe du D^r^ Sarton en vue d'une singularité (hybridation) humain-machine, dans *Les cavernes d'acier*. Le D^r^ Sarton de l'U.S. Robots envisageait de convertir les populations de la terre à une telle combinaison C/Fe pour créer des humains robotisés :

> C/Fe ? Qu'est-ce que c'est que ça ? Tout simplement les symboles chimiques du carbone et du fer, Elijah. Le carbone est l'élément de base de la vie humaine, et le fer est celui des robots. Il devient facile de parler de C/Fe, quand on désire exprimer une forme de culture qui puisse combiner au mieux les propriétés des deux éléments, sur des bases égales et parallèles. Ah !, fit Baley. Mais, dites-moi, comment écrivez-vous ce symbole C-Fe ? Avec un trait d'union. Non, Elijah, avec une barre en diagonale. Elle signifie que ni l'un ni l'autre des

éléments ne prédomine, et qu'il s'agit d'un mélange des deux, sans qu'aucune ait la priorité[37].

Aussi, dans « L'homme bicentenaire », « la technique de la prothésologie » permettant d'incorporer des prothèses dans un organisme humain permet de robotiser l'être humain pour lui procurer plusieurs avantages[38].

Étape 2 : Identification d'un enjeu pouvant subir un impact de la source

Dans les écrits d'Asimov, l'identification d'un enjeu pouvant subir un impact se fait en même temps que l'étape trois sur la détermination d'un impact.

Étape 3 : Détermination de l'impact réel ou négatif de la source sur les enjeux E3LS

Essayons de voir pour chacune des deux sources quels sont les impacts mentionnés par Asimov sur les différents enjeux.

1) Les impacts des robots servant à différents usages.

i) Santé

Dans « L'homme bicentenaire », l'amélioration de la santé comme impact positif résulte de la prothésologie créée par le robot Andrew[39].

Dans la nouvelle « Ségrégationniste », Asimov met en scène un robot chirurgien qui contribue au maintien de la santé humaine en implantant des cœurs cybernétiques ou un cœur de fibre[40].

Mais un robot comme Robbie peut avoir un impact négatif sur la santé psychologique de l'enfant (développement psychologique). La nouvelle révèle que ce robot, qui fut vendu comme « bonne d'enfants », est « le meilleur robot que l'on puisse trouver sur le marché[41] ». Il soulageait le père et la mère dans leur travail. Il avait l'avantage d'être un excellent compagnon sécuritaire pour l'enfant Gloria (selon les Lois de la robotique). Mais le véritable problème est que la petite fille Gloria aime tellement jouer avec un tel robot si

gentil qu'elle ne veut plus se développer en société avec les autres enfants :

> – C'est justement ce qui me tracasse, George ! Elle ne veut plus jouer avec personne d'autre. Il y a des douzaines de petits garçons et de petites filles avec qui elle devrait se lier d'amitié, mais il n'y a rien à faire. Elle refuse de les approcher à moins que je ne l'y contraigne[42].

ii) Vie et mort

Tout au long des romans et des nouvelles d'Asimov, l'observation des trois Lois de la robotique a pour conséquence positive la protection de l'être humain.

Tout particulièrement, la nouvelle « Évidence » permet d'illustrer que la protection de l'humain est un impact positif qui résulte de l'obéissance aux Lois de la robotique du robot humanoïde :

> J'aime les robots, je les aime beaucoup plus que les êtres humains. Si l'on pouvait créer un robot capable de tenir des fonctions publiques, j'imagine qu'il remplirait idéalement les devoirs de sa charge. Selon les Lois de la robotique, il serait incapable de causer du préjudice aux humains, il serait incorruptible, inaccessible à la sottise, aux préjugés. Et lorsqu'il aurait fait son temps, il se retirerait, bien qu'immortel, car il ne pourrait pas blesser des humains en leur laissant savoir qu'ils avaient été dirigés par un robot. Ce serait l'idéal[43].

Et, dans *Les robots et l'empire*, la protection de l'humanité future est aussi un impact positif de la Loi Zéro que les robots Daneel et Giskard ont créée en conformité avec les trois premières Lois. L'application de la Loi Zéro par Giskard le conduit à la « mort » parce qu'il ne peut pas vivre en acceptant le risque de se tromper dans sa décision :

> – Tu as bien fait, selon la Loi Zéro. Tu as sauvé toutes les vies que tu pouvais sauver. Par humanité, tu as bien fait. Pourquoi tant souffrir alors que tu as fait ce qui arrange tout. D'une voix

> si altérée qu'on distinguait à peine sa parole, Giskard dit : Parce que je n'en suis pas certain. Et… si le D[r] Mandamus… avait raison… après tout… et si les Spaciens triomphaient… Adieu, ami Dan… Et Giskard sombra dans le silence, pour ne jamais plus parler ni bouger[44].

Asimov ne met pas seulement en scène des robots moraux qui ont des impacts positifs. S'ajoutent aussi les impacts négatifs des robots immoraux pouvant causer la mort des humains parce qu'ils échappent à la règle morale. Ceux-ci constituent des exceptions. Les impacts négatifs des robots ratés ou altérés qui peuvent se révéler dangereux et provoquer l'apocalypse dans la science-fiction d'Asimov proviennent d'une erreur humaine.

Par exemple, dans la nouvelle « Cycle fermé », le robot Speedy doit protéger la vie des humains (Powell et Donovan) sur la planète Mercure : « Seul Speedy était capable de leur ramener le sélénium. Pas de sélénium, pas de banc de cellules photo-électriques. Pas de bancs de cellules… la mort par cuisson lente était l'une des façons les plus déplaisantes de passer de vie à trépas[45]. » Mais le robot tourne en rond autour du sélénium pour se protéger (plutôt que de risquer sa vie en rapportant le sélénium), parce que l'ordre donné par les humains n'était pas suffisamment ferme. Les humains se mettent ainsi en danger de mort.

Lenny est un autre exemple d'un impact négatif du robot qui échappe à la règle en raison d'une erreur humaine. Les robots de type Lenny, qui peuvent apprendre, sont créés par un ordinateur à clavier qui a l'avantage de fournir aux « chercheurs le dessein d'un cerveau qui offrira tous les réseaux positroniques nécessaires pour la fabrication d'un robot[46] ». Mais l'ordinateur n'est pas désactivé pendant une visite. Mortimer âgé de 16 ans pianote dessus pendant que l'ordinateur prépare un nouveau robot. Le nouveau robot Lenny qui en résulte n'est pas plus intelligent qu'un bébé et cela

implique comme impact négatif qu'il brise sans le vouloir le bras d'un technicien.

iii) Environnement

Les robots chez Asimov ont aussi comme impact positif de résoudre les problèmes que les humains ont de travailler dans des environnements hostiles sur la terre et sur d'autres planètes. Le robot Speedy sur la planète Mercure était capable de rapporter aux humains le sélénium nécessaire à la fabrication de bancs de cellules photo-électriques. Dans l'univers hostile hors des « cavernes d'acier », l'utilisation des robots sur les fermes permet d'accomplir le travail nécessaire pour assurer la subsistance des gens dans les villes.

iv) Économie

Tel que nous l'avons vu en présentant les principaux personnages chez Asimov, la compagnie U.S. Robots fabrique et contrôle l'utilisation des robots en fonction de l'impact économique positif que ceux-ci représentent.

Par exemple, la nouvelle « Pour que tu t'y intéresses » révèle que l'avantage économique est toujours ce qui détermine le choix de la compagnie U.S. Robots en l'orientant vers des développements de robots rapidement commercialisables. Par exemple, si la création des robots animaux permet à l'humain de récolter des avantages (impacts positifs) sur la nature dans l'équilibre écologique des insectes, l'enjeu est économique pour mieux favoriser le développement des robots humanoïdes :

> Nous allons vous et moi créer un monde qui va commencer à accepter les robots quels qu'ils soient. L'homme moyen peut avoir peur d'un robot qui ressemble à un homme et qui semble assez intelligent pour le remplacer, mais il n'aura pas peur d'un robot qui ressemble à un oiseau et qui se borne à manger des insectes pour son bien-être à lui. Puis, en fin de compte, quand il aura perdu l'habitude d'avoir peur de certains robots, il

> n'aura pas peur d'aucun robot. Il aura tellement l'habitude des oiseaux-robots et des abeilles-robots et des vers-robots qu'un homme-robot ne lui semblera qu'un prolongement des autres[47].

Il y a donc ici un lien entre l'impact économique et la recherche scientifique dans le développement des robots.

v) Statut et développement de la recherche scientifique

La recherche sur le robot a un impact positif sur la connaissance du cerveau humain. Dans la nouvelle « Étranger au paradis », l'expérience sur le cerveau du robot humaniforme permet de mieux comprendre le cerveau humain pour guérir des maladies comme l'autisme :

> Alors ils entreprirent un long travail : Randall, soumis à des stimuli artificiels pendant des périodes de plus en plus longues, révéla le fonctionnement intérieur de son cerveau et donna ainsi des indices sur le fonctionnement intérieur de tous les cerveaux, ceux que l'on dit normaux, aussi bien que ceux de son genre[48].

Le problème du cerveau du robot comme un « spécimen de l'association des techniques robotiques et télémétriques » a aussi l'avantage de la convergence des sciences : « Ce que nous attendons d'un homologiste, c'est qu'il établisse un programme beaucoup plus subtil que ce dont un simple télémétriste est capable[49]. »

Cette convergence vise à enrichir le savoir sur le cerveau des robots humaniformes qui peut rendre service à la recherche sur le cerveau humain dans sa conquête de la galaxie :

> Tout va très bien (Robot sur Mercure) Le programme fonctionne. Il a testé ses sens. Il a fait les différentes observations visuelles. Il a obscurci le Soleil et l'a étudié. Il a étudié l'atmosphère et la nature chimique du sol. Tout a bien marché. – Mais pourquoi court-il ? – Je crois que c'est parce qu'il en a envie,

> Anthony. Si on veut programmer un ordinateur aussi complexe qu'un cerveau, il faut bien s'attendre qu'il ait des idées personnelles[50].

vi) Liberté de choix

Les robots ont des impacts négatifs sur la liberté humaine. Dans « Conflit évitable », les machines sont des robots et elles dirigent le monde en équilibrant les forces économiques des quatre grandes puissances : elles procurent à l'homme l'avantage de le libérer des conflits économiques, puisqu'elles ont le contrôle absolu de l'économie. Mais cela implique comme désavantage que « l'humanité a perdu le droit de dire son mot dans la détermination de son avenir[51] ».

Dans « La vie et les œuvres de Multivac », l'Ordinateur mondial (Multivac) avec des millions de robots sous ses ordres libère le monde entier de multiples problèmes (prédiction des cataclysmes, des guerres économiques…), de sorte que les humains vivent sur la terre dans un confort parfait. Mais cela implique comme désavantage la perte de leur liberté de choix : « Tous les hommes et les femmes sur la terre pouvaient faire ce qui leur plaisait, pourvu toutefois que Multivac, qui jugeait tous les problèmes humains d'une façon parfaite, ne considère pas que le choix en question soit contraire au bonheur humain[52]. »

vii) Vie privée

Les robots ont des impacts négatifs sur la vie privée. La « tyrannie de justice absolue de Multivac[53] » où il n'y a plus que cinq millions d'êtres humains sur terre suppose que cet ordinateur est partout, donc qu'il n'y a plus de vie privée :

> Multivac n'avait plus de domicile particulier. C'était une présence globale composée de fils reliés ensemble, de fibres optiques, de micro-ondes. C'était un cerveau divisé en une centaine de filiales, mais agissant comme une unité. Il avait

> des terminaux partout et aucun des cinq millions d'êtres humains n'était bien loin de l'un d'eux[54].

viii) Vivre-ensemble (local et national)

Quelles peuvent être les conséquences sur les relations sociales de l'utilisation de tels robots pour des sociétés qui en dépendent ? Dans le cas de Robbie, il y a une tension qui se révèle entre la famille Weston et les voisins. M^me^ Weston annonce au père (Mr. Georges Weston) que la « plupart des gens du village considèrent Robbie comme dangereux. Les enfants ne sont pas autorisés à s'approcher[55] » de leur maison :

> Que viennent faire les voisins là-dedans ? Écoute-moi bien. Un robot est infiniment plus digne de confiance qu'une bonne d'enfants humaine. Robbie n'a été construit en réalité que dans un but unique… servir de compagnon à un petit enfant. Sa *mentalité* tout entière a été conçue pour cela. Il ne peut faire autrement qu'être fidèle, aimant et gentil. C'est une machine qui est faite ainsi. C'est plus qu'on en peut dire pour les humains[56].

D'autres nouvelles, comme celle de Tony dans « Satisfaction garantie », illustrent le problème des relations sociales qui peut exister entre des humains qui vivent avec des robots et des humains qui refusent de vivre avec des robots. Cette nouvelle met en lumière le même problème de tension sociale qui existe entre les Spaciens et les Terriens médiévalistes dans les romans d'Asimov.

Sur le plan national, les robots comme Byerley dans « Conflit évitable » et le robot sosie du président Winkler dans « L'incident du tricentenaire » illustrent le questionnement sur les impacts négatifs possibles des robots humanoïdes dans des fonctions de gouvernance telles que la mairie (Byerley) et la présidence des États-Unis (le sosie de Winkler). Dans « L'incident du tricentenaire », le président

Winkler a été tué et remplacé par un robot. De sorte que la nouvelle tourne autour de la question du risque de la perte de l'humanité en raison d'un robot à la présidence :

> Edwards répondit : « Bien sûr, je l'admets. Mais rendez-vous compte du précédent. Un robot à la Maison-Blanche pour une excellente raison pourrait amener un robot à la Maison-Blanche dans vingt ans pour une très mauvaise raison, puis des robots à la Maison-Blanche sans aucune raison du tout, mais par habitude. Ne voyez-vous pas qu'il est important d'assourdir à ses toutes premières notes une éventuelle trompette sonnant la fin de l'humanité ? »[57]

ix) Vivre-ensemble (international)

Quels clivages peuvent alors exister entre des humains et des robots dans la galaxie ? Et quels clivages peuvent exister entre des sociétés sans robots et des sociétés avec robots ? Les robots étant admis seulement sur les autres planètes ou à l'extérieur des villes sur la terre, ils impliquent comme impact négatif une tension entre la terre et les colonies.

La nouvelle « Assemblons-nous » fait aussi ressortir le problème de l'impact négatif de la création des robots humanoïdes comme armes de guerre. Le malheur était qu'à Washington on ne savait pas trop comment réagir à « l'éventualité d'une invasion des USA par des robots humanoïdes » que les pays en compétition (« eux ») ont créée et qui se rassemblent pour devenir une « arme de destruction » en contexte de guerre froide : « Cela signifierait qu'ils aient fabriqué des humanoïdes qu'on ne pourrait distinguer des hommes à un mètre de distance[58]. » Asimov fait alors la distinction entre « nous et eux[59] » lorsque « nous » parlons d'avance en robotique dans les pays en compétition : « Nous sommes capables de construire des robots humanoïdes[60]. » Et si nous étions « eux » et s'ils étaient « nous » ? Survivrions-nous à l'impact des robots qui pourraient s'assembler à notre insu ? Mais qui sait si ces robots voudraient vraiment nous

attaquer ? Les robots ne constitueraient-ils pas tout simplement une force de dissuasion, destinée non à attaquer, mais à dissuader l'adversaire d'attaquer pour le meilleur vivre-ensemble ?

Dans la nouvelle « Conflit évitable », Asimov montre comment les quatre machines à calculer (selon les quatre grandes puissances de l'économie mondiale) se préoccupent du bien de l'humanité en résolvant le problème des guerres économiques. L'impact positif est qu'il n'y aura plus de guerres économiques qui risquent de provoquer la fin de l'humanité. Mais cette recherche d'une paix mondiale fondée sur les décisions des machines implique en même temps, comme impact négatif pour l'humain, la perte de la liberté de choix de son avenir.

2) Les impacts de la robotisation de l'humain pour les Spaciens

i) Santé

Les progrès de la robotique selon le rêve de fusion C/Fe combiné à la fabrication des fœtus et des prothèses (filtre nasal, cœur, poumons, etc.) ont comme principal impact d'améliorer la santé des Spaciens. Leur santé est en bon état physiologique à l'instar « de tout ce qu'était Daneel Olivaw mais avec, en plus, le fait d'être humains[61] ».

ii) Vie et mort

Les progrès technologiques pour les Spaciens concernent non seulement l'amélioration de leur santé, mais une vie longue, rangée et casanière. Sur la planète Solaria, les Spaciens ont une espérance de vie de trois cents ans. Sur la planète Aurora, la vie de Gladia, par exemple, a « vingt-trois décennies[62] ». Autre exemple, sur la planète Terre, le D[r] Alvin Magdescu de la compagnie U.S. Robots « avait quatre-vingt-quatorze ans et ne se maintenait en vie que grâce à des prothèses qui, entre autres, remplissaient la fonction de foie et de reins[63] ».

viii) Vivre-ensemble national

Le roman *Face aux feux du soleil* illustre l'impact négatif de la robotisation de l'humain sur la façon de vivre ensemble sur la planète Solaria. Puisque les humains robotisés sur cette planète vivent au moins trois cents ans, il faut limiter leur nombre à « vingt mille habitants ». L'utilisation massive des robots fera tout le travail nécessaire pour assurer le confort des habitants. De plus, puisqu'il n'y a que vingt mille habitants sur la planète, ces habitants vont se répartir sur l'ensemble du territoire de la planète. L'impact négatif sera qu'ils vivront isolés les uns des autres et communiqueront entre eux seulement par des vidéos. Le choix social de la robotisation de l'humain a pour conséquence négative de briser les relations personnelles entre les membres de la société, de vouer les propriétaires terriens à l'oisiveté et de gérer la stabilité de la population pour maintenir l'équilibre de la population. Il y aura un problème similaire chez les Aurorains parce que certains voudront suivre la voie solarienne et que d'autres voient un effet néfaste d'oisiveté sur eux.

Ce problème de l'impact négatif de la robotisation sur le plan national aura une répercussion sur le vivre-ensemble international opposant les partisans de la non-expansion aux coloniens qui choisissent l'expansion. Dans *Les robots et l'empire*, Baley (dans la mémoire de Giskard) est conscient que le risque d'une crise de rupture des relations internationales entre Spaciens et Terriens est toujours latent :

> – Oui, mais nous n'y pouvons rien et une crise va se produire – ou peut se produire – avant même cela, mais après que mon temps sera révolu, cependant. – À quoi pensez-vous, Monsieur ? Quelle est cette crise ? – Giskard, il s'agit d'une crise qui peut survenir parce que Fastolfe est une personne étonnamment persuasive. Ou encore parce qu'il existe un autre facteur qui lui est lié et qui détermine cette tâche.

> – Monsieur ? – Tous les officiels qu'a rencontrés le Dr Fastolfe paraissent maintenant se montrer pleins d'enthousiasme pour l'émigration, Ils n'y étaient pas favorables auparavant ou, s'ils l'étaient, ils manifestaient de vives réserves. Et une fois favorables les dirigeants qui font l'opinion, on est sûr que d'autres suivront. Cela va s'étendre comme une épidémie. – N'est-ce pas ce que vous souhaitez, Monsieur ? – Oui, c'est bien cela, mais ça l'est presque beaucoup trop. Nous allons nous répandre dans la galaxie, mais que se passera-t-il si les Spaciens n'en font pas autant ? – Pourquoi ne le feraient-ils pas ? – Je ne sais pas. J'admets une simple supposition, j'envisage une possibilité. Que se passera-t-il s'ils ne le font pas ? – La Terre et les mondes où s'établiront les Terriens deviendront alors plus puissants, selon ce que je vous ai entendu dire. – Et les Spaciens deviendront plus faibles. Nous connaîtrons cependant une période pendant laquelle les Spaciens demeureront plus forts que la Terre et ses Coloniens, encore que la marge risque de devenir encore plus étroite. En fin de compte, les Spaciens percevront les Terriens comme représentant un danger croissant. Alors, les mondes spaciens décideront sûrement qu'il convient d'arrêter la Terre et les Coloniens avant qu'il ne soit trop tard et il leur paraîtra utile de prendre des mesures drastiques. Ce sera là une période de crise qui déterminera toute l'histoire future de l'humanité[64].

ix) Vivre-ensemble (international)

Dans *Les cavernes d'acier*, une des plus grandes préoccupations d'Asimov est l'impact négatif sur les relations internationales comme les tensions causées par les différences entre les humains robotisés (Spaciens) qui acceptent de vivre avec des prothèses et des robots et les Terriens qui refusent de vivre avec des robots. Ces tensions risquent de dégénérer dans des conflits sanglants : « Ainsi donc, il y a trois jours un Spacien a été assassiné, et ses compatriotes pensent que le meurtrier est un Terrien. […] Mais, si c'était réellement vrai,

une affaire comme celle-là entraînerait la disparition de New York de la planète : elle nous ferait tous sauter[65] ! » Cet impact négatif d'une guerre internationale risque toujours de se produire pour autant que les New-Yorkais se sentent toujours rejetés par les Spaciens (humains améliorés et aseptiques) qui traitent « les habitants de la Terre comme des êtres pourris par les maladies[66] ». Les New-Yorkais leur chantent alors un vieux chant du pays, au refrain lancinant :

> L'homme est issu de la terre, entends-tu ?
> C'est sa mère nourricière, entends-tu ?
> Spacien, va-t-en, disparais
> De la Terre qui te hait !
> Sale Spacien, entends-tu ?[67]

> Ainsi donc les Spaciens se tenaient isolés derrière leur barrière, produit de leur progrès scientifique, et les Terriens ne disposaient d'aucune méthode leur permettant d'espérer qu'un jour ils pourraient détruire cette barrière[68].

En somme, dans ses romans et ses nouvelles, Asimov développe beaucoup d'impacts des robots en développement sur l'ensemble des enjeux E^{3}LS. Par contre, la robotisation de l'humain implique surtout des impacts positifs et négatifs sur le vivre-ensemble sur les plans local, national et international. Passons maintenant aux exemples de jugements de valeur sur les impacts.

MOMENT 2 : L'évaluation des impacts à partir des valeurs retenues et MOMENT 3 : La pondération des jugements finaux d'évaluation en vue de la décision

Asimov, dans sa science-fiction, ne suit pas systématiquement les deux étapes de l'évaluation à partir des jugements de valeur sur les impacts que nous avons relevés (les impacts des robots servant à différents usages et les impacts de la robotisation de l'humain), ni celles de la pondération des jugements finaux d'évaluation en vue de la décision au sujet

de l'acceptabilité ou non. Par contre, il met en évidence des conflits entre différents jugements de valeur et des conflits entre différentes pondérations. Le but sera ici de montrer d'abord les principaux conflits de jugements valeur et ensuite les pondérations qui permettent de les résoudre.

1) Les impacts des robots servant à différents usages

i) Jugement de valeur maximisant la qualité de la santé humaine et pondération

Conflit de jugements de valeur : Dans la nouvelle « Ségrégationnisme », les humains veulent un cœur de métal et les métallos (robots métalliques) veulent un cœur de chair. Mais la discussion entre le robot chirurgien et l'humain (le patient qui veut un cœur de métal) révèle qu'il y a un conflit entre deux jugements divergents de valeur sur la question de la maximisation de l'amélioration de la santé humaine. Pour maximiser les avantages de la santé de l'homme combinés à ceux du robot (la santé optimale), l'humain préfère obtenir un cœur de métal (cœur de métallo) ; mais le chirurgien juge que cela ne servira à rien pour optimaliser la santé purement humaine :

> Ça ne servira à rien. Il est nerveux, et il a pris sa décision. – Vraiment ? Oui ; il veut du métal. Ils veulent tous du métal. […] Le chirurgien dit avec force : – Pour moi, c'est une question d'utilisation optimale des choses. – Utilisation optimale ! Ce n'est pas un argument ! Le patient s'en moque de l'utilisation optimale. – Moi je ne m'en moque pas[69].

Pondération en fonction du critère de la nature humaine : Selon la pondération à partir de la conformité ou de la non-conformité avec la nature humaine, même si l'humain juge que l'obtention d'un cœur de métal maximise la santé humaine en obtenant les avantages de l'humain combinés à ceux du robot, le chirurgien juge au contraire que cette transformation de l'humain en hybride implique un risque inacceptable de perdre la nature humaine. Le

risque est inacceptable parce que l'humain perd son identité humaine :

> Et voilà ! Mais alors, où est le problème docteur ? Avez-vous peur que je me transforme en robot… en Métallo, comme on les appelle depuis qu'on leur a accordé la citoyenneté ? – Il n'y a rien à redire à un Métallo, en tant que Métallo comme vous le dites, ils sont maintenant des citoyens comme les autres. Mais vous, vous n'êtes pas un Métallo. Vous êtes un être humain. Pourquoi ne pas le rester[70] ?
>
> Pourquoi voudrions-nous conserver ces différences ? Nous aurions le meilleur des deux : les avantages de l'homme combinés à ceux du robot. – Vous obtiendrez un hybride, dit le chirurgien d'un ton réprobateur. Quelque chose qui ne serait pas les deux à la fois, mais ni l'un ni l'autre. N'est-il pas logique de supposer qu'un individu doit être assez fier de sa structure et de son identité pour ne pas désirer l'altérer par des éléments étrangers ? Pourquoi cet individu désirerait-il devenir un Métis[71] ?

La pondération par l'argument de la nature humaine est tellement importante que nous allons lui consacrer une section spéciale (voir la troisième partie de ce chapitre).

ii) Jugement de valeur maximisant la qualité de la vie humaine et pondération

Conflit de jugements de valeur : Rappelons-nous que, dans la nouvelle « Le correcteur », Easy est un robot correcteur qui permet de corriger rapidement des textes écrits. Asimov nous place devant deux jugements divergents de valeur au sujet du même impact sur l'enjeu de la vie humaine. D'un côté, le D[r] Lanning de l'U.S. Robots fait un jugement qui maximise la qualité de l'impact positif du robot Easy (vitesse de traitement de texte, corrections) pour libérer l'humain de l'inutile l'esclavage que constitue le labeur physique et

mental, « même s'il fallait payer cette amélioration d'un certain bouleversement économique temporaire[72] » :

> Vous pourriez employer le robot à mille autres usages, professeur. Jusqu'à présent, son travail n'a été qu'à libérer l'homme de l'esclavage que constitue le labeur physique. Mais n'existe-t-il pas un labeur mental que l'on peut également considérer comme inutile esclavage ? Lorsqu'un professeur capable d'un travail puissamment créateur est assujetti, deux semaines durant, au travail mécanique et abrutissant qui consiste à corriger des épreuves, me traiterez-vous de plaisantin si je vous offre une machine capable de le faire en trente minutes[73] ?

D'un autre côté, le professeur Ninheimer juge que ce robot maximise le risque de perdre la qualité de la vie créatrice (le sens du travail) à long terme :

> Depuis deux cent cinquante ans, la machine a entrepris de remplacer l'Homme en détruisant le travail manuel. La poterie sort de moules et de presses. Les œuvres d'art ont été remplacées par des facsimilés. Appelez cela le progrès si vous voulez ! Le domaine de l'artiste est réduit aux abstractions ; il est confiné dans le monde des idées. Son esprit conçoit et c'est la machine qui exécute. Pensez-vous que le potier se suffise de la seule création mentale ? Supposez-vous que l'idée suffise ? Qu'il n'existe rien dans le contact de la glaise elle-même, qu'on n'éprouve aucune jouissance à voir l'objet croître sous l'influence conjuguée dans la main et de l'esprit ? Ne pensez-vous pas que cette croissance même agisse en retour pour modifier et améliorer l'idée ? […]
>
> Je suis un artiste créateur ! Je conçois et je construis des articles et des livres. Cela comporte davantage que le choix des mots et leur alignement dans un ordre donné. Si là se bornait notre rôle, notre tâche ne nous procurerait ni plaisir ni récompense.

> Un livre doit prendre forme entre les mains de l'écrivain. Il doit travailler et retravailler, voir l'œuvre se modifier au-delà du concept original. C'est quelque chose que de ternir les épreuves à la main, de voir le texte imprimé et de le remodeler. Il existe des centaines de contacts entre l'homme et son œuvre à chaque stade de son élaboration, et ce contact lui-même est générateur de plaisir et paie l'auteur du travail qu'il consacre à sa création plus que ne le pourrait faire aucune autre récompense. *C'est de tout cela que votre robot nous dépouillerait*[74].

Le conflit de jugements de valeur pourrait s'énoncer comme suit : A. La réduction du temps de travail qui consiste à libérer du labeur mental est jugée comme maximisant la qualité de la vie créatrice au travail ; B. La réduction du temps de travail qui consiste à libérer du labeur mental est jugée comme minimisant la qualité de la vie créatrice au travail. Comment réduire l'esclavage au labeur mental en ne réduisant pas le sens du travail ? Quelle position intermédiaire entre A et B pourrions-nous choisir ?

Pondération en fonction du critère de la vie bonne : Selon la pondération à partir de la conformité ou de la non-conformité avec la vie bonne (Aristote), même si l'impact positif d'être libéré de l'esclavage au labeur mental est acceptable, le risque qui en découle de perdre la vie créatrice à long terme est inacceptable, parce que c'est cette vie créatrice de l'homme au travail qui procure le bonheur de l'accomplissement de soi :

> Machines à écrire et presses à imprimer nous dépouillent partiellement mais votre robot nous dépouillerait totalement. Votre robot se charge de la correction des épreuves. Bientôt il s'emparera de la rédaction originale, de la recherche à travers les sources, des vérifications et contre-vérifications de textes et pourquoi pas des conclusions. Que restera-t-il de l'érudit ? Une seule chose : le choix des décisions concernant les ordres à donner au robot pour la suite du travail ! Je veux épargner

> les futures générations d'universitaires et d'intellectuels de sombrer dans un pareil enfer[75].

vi) Jugement de valeur maximisant la qualité de l'autonomie (liberté de choix) de la personne

Conflit de jugements de valeur entre maximiser la qualité de la sécurité et maximiser la qualité de l'autonomie : Dans la nouvelle « La vie et les œuvres de Multivac », le monde entier juge que Multivac maximise la qualité de la sécurité de vivre dans un confort parfait. D'un autre côté, Bakst juge que Multivac maximise le risque de détruire la liberté humaine. Deux questions se posent pour nous aider à réfléchir à ce problème du conflit entre deux jugements de valeur :

> Avez-vous tous oublié ? Te rappelles-tu comment c'était autrefois ? Te rappelles-tu le XX^e^ siècle ? Nous vivons vieux maintenant ; nous vivons en sécurité ; nous vivons heureux. – Nous ne valons plus rien. – Veux-tu revenir à un monde tel qu'il était autrefois ?[76]

L'histoire montre que Baskt tue l'ordinateur pour optimiser une réelle liberté de choix :

> Rien ne s'interposa et, retenant sa respiration, Bakst réalisa que le bruit avait cessé, le murmure s'était tu, Multivac s'était arrêté. Si, dans un instant, le léger bruit ne repartait pas, alors c'était le point clé qu'il avait touché, et aucune réparation ne serait possible. Si c'était le cas les robots allaient approcher… Il se retourna dans le silence qui durait. Au loin les robots travaillaient tranquillement. Aucun ne s'approchait. Devant lui, les images des quatorzes hommes et femmes semblaient stupéfiées par la chose énorme et soudaine qu'il venait d'accomplir. Bakst leur dit : « Multivac est arrêté, détruit. On ne peut pas le reconstruire. » […] « Je nous ai donné notre liberté. » Et il s'arrêta, conscient enfin du poids croissant du silence. Quatorze images le regardaient fixement et aucune d'entre elles ne lui répondait. Bakst dit brusquement : « Vous

parliez de liberté. Vous l'avez maintenant ! » Puis, d'une voie mal assurée : « N'est-ce pas cela que vous vouliez[77] ? »

Pondération en fonction du critère de l'autonomie ? Asimov ne fait pas de pondération en raison de la conformité ou de la non-conformité avec l'autonomie de la personne de choisir. Car, si Bakst est triomphant, les humains sont libres[78] ! Le lecteur peut comprendre que les humains n'avaient pas réalisé ce qui les attendait en cas de victoire – le jugement de valeur de Bakst maximise la qualité de l'autonomie (la liberté de choix), mais aussi l'inconnu et le danger. Nous restons dans l'impasse du conflit de jugements de valeur entre la maximisation de l'autonomie et la maximisation de la sécurité.

vii) Jugement de valeur maximisant la qualité de la vie privée

Conflit de jugements de valeur entre maximiser la qualité de la sécurité et maximiser la qualité de la vie privée : Dans la même nouvelle « La vie et les œuvres de Multivac », comment juger si l'impact de l'ordinateur maximise ou minimise la qualité de notre vie privée ? Toute information sur nous dans l'ordinateur est scrupuleusement gardée en mémoire par l'ordinateur. Pouvons-nous juger que Multivac risque de minimiser la qualité de notre vie privée ? Si Bakst le fait cesser de fonctionner, nous sommes renvoyés à nous-mêmes pour juger si le risque est (presque totalement) diminué par le fait que nous désactivons notre ordinateur. Autrement dit, si nous jugeons que Multivac maximise le risque de perdre la qualité de notre vie privée au point de vouloir désactiver ce même type d'ordinateur aujourd'hui, déconnectons-nous d'Internet, éteignons notre ordinateur.

Pondération ? Asimov ne fait aucune pondération. Donc, s'il n'y a pas de pondération, il est impossible de sortir de l'impasse du conflit entre le jugement qui maximise la

qualité de la sécurité et le jugement qui maximise la qualité de la vie privée.

ix) Jugement de valeur maximisant la qualité du vivre-ensemble (relations internationales) et pondération

Rappelons que, dans la nouvelle « Conflit évitable », au XXI[e] siècle, des ordinateurs géants, les Machines (« successeurs de Multivac[79] »), veillent au bonheur de l'humanité (impact positif) en assurant un développement économique égal entre les quatre grandes puissances sur la terre. Nous retrouvons chez Asimov son côté marxiste, puisque la source des conflits internationaux, c'est le problème économique. Or, pour éviter les conflits, il faut une régulation économique par les Machines sur toute la planète. Les humains suivent les conseils des Machines. Mais les fondamentalistes (Société pour l'humanité) protestent contre cette tutelle qui implique comme impact négatif pour eux la perte de la détermination de leur destin.

Une série de ratés exigent l'intervention de Susan Calvin sur demande du coordinateur mondial Byerley. Plusieurs humains jettent la faute sur les Machines, et cela semble absurde : les Machines ne se trompent pas normalement et ne peuvent mentir aux humains en obéissant à la morale des Trois Lois. Mais Calvin finit par comprendre que ce sont les Machines qui font des ratés pour réduire les impacts négatifs des fondamentalistes qui n'obéissent pas : « En un mot, des hommes qui, en refusant avec ensemble d'appliquer les décisions de la Machine, peuvent d'un jour ou l'autre jeter le monde dans le chaos… Ce sont ceux-là mêmes qui appartiennent à la Société pour l'humanité[80]. » En s'opposent aux Machines, ils veulent revenir aux conflits et aux jeux de puissance sur les autres. Mais les Machinent suppriment les opposants en leur faisant perdre leur fonction de direction, et ce sans le dire aux humains.

Cette nouvelle repose sur l'idée que l'optimum des impacts positifs sur le vivre-ensemble international est atteint lorsque tout accroissement du bien-être économique des humains supposerait l'augmentation de leur obéissance à la Machine (robot) dans le monde. Il peut alors sembler que c'est la désobéissance des fondamentalistes radicaux qui fait problème au robot coordinateur Byerley et non la personnalité des Machines à calculer (robots) qui permet seulement de prendre de bonnes décisions dans l'organisation économique des relations humaines pour le meilleur vivre-ensemble.

Conflit de jugements de valeur entre maximiser les conflits et les jeux de puissances économiques et maximiser la paix économique pour le meilleur vivre-ensemble international : D'un côté, Byerley juge horrible le fait que ce sont les machines qui maximisent la paix économique pour le meilleur vivre-ensemble international, parce que l'humain perd le droit de dire son mot dans la détermination de son avenir. Cette maximisation implique la perte de la liberté humaine et contredit la morale de la robotique. Mais, d'un autre côté, Susan Calvin juge que cela est merveilleux :

> – Si je comprends bien, Susan, vous me dites que la Société pour l'humanité a raison et que l'humanité a perdu le droit de dire son mot dans la détermination de son avenir.
>
> – Ce droit, elle ne l'a jamais possédé, en réalité. Elle s'est trouvée à la merci des forces économiques et sociales auxquelles elle ne comprenait rien… des caprices des climats, des hasards de la guerre. Maintenant les Machines les comprennent ; et nul ne pourra les arrêter puisque les Machines agiront envers ces ennemis comme elles agissent envers la Société pour l'humanité… ayant à leur disposition la plus puissante de toutes les armes, le contrôle absolu de l'économie.

> – Quelle horreur !
>
> – Dites plutôt quelle merveille ! Pensez que désormais et pour toujours les conflits sont devenus évitables. Dorénavant seules les Machines sont inévitables ![81]

Pondération en fonction de la valeur jugée prioritaire pour le bien vivre-ensemble : Comment interpréter que Calvin soit émerveillée ? Calvin est émerveillée, car le fait d'éliminer les conflits économiques des fondamentalistes pour maximiser la qualité de la paix mondiale est jugé prioritaire pour le meilleur vivre-ensemble. Asimov ne donne pas la raison explicite de cette pondération. Mais on devine la raison implicite en pensant qu'il s'agit d'une préfiguration de la Loi Zéro en ce sens que les Machines lèsent un petit nombre d'hommes (les fondamentalistes) pour protéger le plus grand nombre :

> Calvin finit par comprendre que les Machines se sont dotées de la Quatrième Loi ou Loi Zéro – *le bien de l'humanité passe avant celui d'un seul homme*. Il faut donc légèrement léser un petit nombre d'hommes – les membres de la Société pour l'humanité – pour continuer à aider le plus grand nombre[82].

2) La robotisation de l'humain

Dans le contexte de l'analyse globale des impacts positifs et négatifs de la robotisation de l'humain (Spacien) qui vit avec des robots soumis à une telle morale de la robotique pour assurer la survie de l'humanité, quelle est alors la principale préoccupation chez Asimov ? C'est toujours la crise qui va se produire – ou qui peut se produire – en raison du jugement de valeur qui maximise la qualité du vivre-ensemble des Spaciens au détriment des Terriens, et vice-versa.

ix) Jugement de valeur maximisant la qualité du vivre-ensemble (relations internationales) et pondération

Rappelons-nous que, dans les romans d'Asimov, le conflit (dilemme) croissant entre l'expansion et l'évolution des Terriens qui vivent sans robots et la non-expansion et la décadence des Spaciens qui vivent avec des robots découle de la robotisation de l'humain. L'expansion et l'évolution des Terriens non robotisés qui refusent de vivre avec des robots a comme enjeu la colonisation : « La colonisation de l'espace est l'unique voie sur la terre[83]. » Elle a l'avantage (impact positif) d'assurer la progression de l'humanité. Mais elle implique en même temps comme impact négatif le risque constant d'une rupture des relations internationales comme les tensions causées par les deux modes divergents de vie entre les Spaciens et les Terriens : les humains robotisés (Spaciens) acceptent de vivre dans l'oisiveté avec des prothèses et des robots, tandis que les Terriens (médiévalistes) refusent de vivre dans l'oisiveté avec des robots.

Conflit de jugements de valeur entre maximiser la façon de vivre ensemble des Terriens et maximiser la façon de vivre ensemble des Spaciens. Les Terriens jugent qu'il vaut mieux vivre ensemble selon leur condition d'humain non amélioré sans robot pour maximiser la colonisation comme impact positif. Mais les Spaciens jugent qu'il vaut mieux vivre ensemble comme des humains améliorés avec des robots pour maximiser la stabilité sociale de la classe oisive telle qu'elle existe sur la planète Solaria, comme le raconte le D[r] Quemot pour répondre à la question de Baley :

> – Qu'entendez-vous par stabilité sociale ? – La situation telle qu'elle existe ici sur Solaria ; un monde où les humains ne représentent plus que la classe oisive. Aussi, n'avons-nous aucune raison d'avoir peur des autres mondes extérieurs.

> Attendons seulement un siècle peut-être, et ils seront tous devenus semblables à Solaria. Je suppose qu'en un sens on peut dire que ce sera la fin de l'histoire de l'homme, mais ce sera surtout la réalisation dans le sens complet du mot. Les hommes auront alors tout ce qu'ils peuvent désirer, tout ce dont ils peuvent avoir besoin. Vous savez il existe une phrase qui, un jour, m'a frappé. Je ne sais d'où elle vient : c'est quelque chose à propos de la recherche du bonheur[84].

Baley se rappelle alors sur un ton pensif le début de la Déclaration des droits de la constitution américaine : « Tous les hommes reçoivent à leur naissance, de leur Créateur, certains droits inaliénables… Parmi ceux-ci il y a le droit à la vie, à la liberté, à la recherche du bonheur[85]. » Et c'est pourquoi il rétorque aussitôt d'un ton sec au D[r] Quemot qu'il est venu sur Solaria pour résoudre le problème du meurtre d'un homme qui a été commis.

Pondération selon l'analyse de l'acceptabilité globale des impacts positifs et négatifs. Quelle pondération chez Asimov est faite entre les risques et les impacts positifs (supériorité des risques sur les impacts positifs ou supériorité des impacts positifs sur les risques) pour résoudre ce conflit de jugements de valeur entre deux visions du vivre-ensemble international (celle des Solariens et celle des Terriens) ?

Asimov dans ses romans maintient l'acceptabilité globale des impacts positifs et négatifs qui sont liés aux deux jugements de valeur. Les Spaciens et les Terriens font chacun leur pondération en choisissant les impacts et les modes de vie. Cette acceptabilité de l'analyse globale des impacts des deux modes de vie chez Asimov peut se faire à partir d'arguments moraux (liberté, vie bonne) basés sur la Déclaration américaine du droit fondamental des personnes à la vie, à la liberté et à la recherche du bonheur. Mais, pour Asimov, il n'y a pas de pondération meilleure que l'autre pour dire que

l'un doit vaincre l'autre. Sa pondération éthique va dans le sens du respect de l'autre. Il faut que chacune des parties en présence respecte le choix de vie de l'autre en autant que ce respect est la condition de possibilité de la paix internationale. Autrement dit, l'important est de respecter les différences entre les deux modes de vie (celui des Terriens et celui des Spaciens) pour assurer les meilleures relations internationales.

Bref, la pondération (équilibre) des impacts (positifs et négatifs) et des jugements finaux de valeurs maximisant les impacts positifs chez Asimov permet de défendre une position mitoyenne entre celle des Terriens et celle des Spaciens qui va dans le sens du choix de Baley et du projet de Fastolfe pour l'expansion et l'évolution (« C'est pourquoi je tente de persuader et les Terriens et les Spaciens de s'étendre et d'évoluer. C'est là une troisième solution et, à mon sens, la meilleure[86] »), lorsqu'il s'agit de se prononcer sur l'acceptabilité globale des impacts du vivre-ensemble avec des robots et des humains robotisés. C'est la seule conclusion que nous permet de faire cette analyse globale d'impact et d'acceptabilité dans la science-fiction d'Asimov.

3. IDENTITÉ HUMAINE ET HUMANISATION DU ROBOT OU ROBOTISATION DE L'HUMAIN

Parmi les risques, ceux portant sur les représentations de ce que nous sommes comme humains sont au cœur de tous les débats sur le développement technologique, et cela est plus manifeste avec les robots qui prennent une apparence humaine ou les humains qui incorporent divers implants. Plusieurs humanistes vont juger inacceptable cette technologie qui humaniserait la machine ou celle qui mécaniserait l'humain. Comment comprendre cette question complexe du rôle de l'identité dans l'acceptation ou la non-acceptation des robots ? La première question qui se pose est de savoir qui nous sommes. Qui fait partie de l'humanité ? Comme

l'histoire nous l'a montré, l'esclavage était jadis permis parce qu'on ne considérait pas la race noire comme en faisant partie. La seconde question porte sur le rôle de notre identité dans notre façon de vivre ensemble. Comme le montre l'exemple précédent, si je ne considère pas quelqu'un comme un humain, je peux le traiter en objet et le dominer. Mais la question la plus difficile sur l'identité est de savoir quels sont les critères qui nous permettent d'identifier un humain et un non-humain. Par quoi est-ce que je reconnais le spécifique de l'humain ? Enfin, si les identités permettent de regrouper des personnes autour d'un « nous », elles ont aussi pour effet de distinguer ce « nous » des autres. Or, entre « nous » et « eux », la distinction devient souvent une question de domination.

Telles sont quelques-unes des questions auxquelles la science-fiction d'Asimov nous aide à réfléchir. Sa science-fiction nous place en effet dans des situations prospectives de robots humanisés et d'humains robotisés. Elle nous permet d'anticiper ces questions éthiques qui mettent en jeu les représentations de ce qu'est l'être humain et de voir comment ces représentations fondent l'acceptabilité ou non de l'humanisation du robot et de la robotisation de l'humain. Si la corrélation de l'identité humaine et de la différence du robot humanisé est fondamentale au cœur de cette science-fiction asimovienne, le robot doit protéger l'humain dans sa différence (le maître) et lui obéir selon les Trois Lois de la morale robotique : « Les Trois Lois stipulent que les robots ne doivent pas faire de mal aux êtres humains et doivent leur obéir[87]. » Il y a toutefois un problème d'identité qui s'annonce, en ce sens que la frontière entre l'humain et le robot tend à s'estomper sous l'impact de l'humanisation du robot et de la robotisation de l'humain. Ainsi, comme le note Georges Vignaux :

> Dans ses différents ouvrages, il [Asimov] prédit que le progrès des technologies informatiques fera perdre à l'homme le monopole de l'intelligence comme celui de la conscience. Et lorsque les machines deviendront intelligentes, elles le seront rapidement beaucoup plus que nous. L'avenir de l'humanité se situerait alors dans une forme de symbiose avec la machine[88].

Cette troisième section vise à mieux faire comprendre ce problème de l'identité dans sa complexité en montrant comment la science-fiction (romans et nouvelles) d'Asimov poursuit quatre objectifs :

1) Comprendre le rôle des représentations de l'humain dans nos façons de voir ce qu'est l'être humain et nos façons de vivre-ensemble avec des robots ;
2) Voir comment la robotisation de l'humain avec des robots a un impact sur nos choix de vie ;
3) Saisir plus spécifiquement comment se pose la question de la définition de l'identité et de la perte d'identité ;
4) Saisir les conséquences des conflits identitaires.

3.1 Qui fait partie de l'humanité ?

La définition que nous nous donnons de l'être humain pour nous le représenter joue un rôle important dans notre façon de nous percevoir et d'agir moralement en relation avec les autres dans les sociétés. Elle donne à l'être humain (identité, nature, personne) son importance, sa valeur.

Dans la science-fiction d'Asimov, nous pouvons distinguer au moins deux représentations de l'être humain dans nos façons de voir ce qu'est l'être humain dans ses relations avec les autres et nos façons de vivre-ensemble avec des robots humanisés. Qu'est-ce que l'être humain que le robot doit protéger ? Quel est le concept le plus représentatif de l'être humain à définir ? Asimov nous fait reconnaître le

problème de la complexité humaine en tant qu'individu, société ou espèce *Homo sapiens* : ces trois instances sont l'une en l'autre, l'une générant l'autre, chacune étant la fin et les moyens des autres, et en même temps potentiellement antagonistes. L'identité humaine est donc en même temps une et multiple dans une définition large, mais elle peut se perdre dans une définition restrictive.

Définition restrictive de l'être humain

Dans l'ère des *Cavernes d'acier*, où l'on commence à voir des robots comme Daneel avec une intelligence de niveau humain, les Trois Lois de la robotique constituent une vue éthique du monde sous-tendant les actions des robots pour protéger l'être humain selon un concept général « d'être humain ». Le concept est général en ce sens qu'il inclut tous les humains.

Cependant, dans *Face aux feux du soleil*, les Solariens (qui sont d'origine humaine) ont choisi la voie de la robotisation et de la longévité avec des robots régis par les Trois Lois de manière normale, mais pour lesquels le sens de la définition du mot « humain » est devenu restrictif : « Selon leur programmation, seules les personnes parlant avec l'accent solarien sont humaines. De cette manière, ces robots n'auront aucun problème à agresser des humains non solariens (et certains sont même programmés spécifiquement pour cela)[89]. » Autrement dit, les Solariens ne protègent que ceux qu'ils définissent comme êtres humains selon des critères *restrictifs* ne s'appliquant qu'à eux. Selon cette définition restrictive, seul est considéré comme humain quelqu'un ayant l'apparence physique humaine et parlant solarien.

Il s'ensuit que, dans *Les robots et l'empire*, Gladia (de Solaria) expliquera au président du monde que c'est en raison d'une définition restrictive de l'être humain que les robots laissés sur Solaria ont pu détruire un vaisseau aurorain :

> – Vous prétendez donc que les Solariens ont redéfini « l'être humain » selon des critères restrictifs ne s'appliquant qu'aux Solariens. – Je ne prétends rien du tout, Monsieur le Président. Personne n'a trouvé d'autre explication pour justifier ce qui s'est passé, c'est tout. – Vous rendez-vous compte que dans toute l'histoire de la robotique jamais on n'a conçu un robot avec une définition aussi restrictive de « l'être humain »[90].
>
> – Puisque vous le dites, Monsieur le Président… Cependant, si seul est considéré comme humain quelqu'un ayant l'apparence physique d'un être humain et pouvant parler comme un Solarien – comme il vous a semblé que c'était le cas, à nous qui étions sur les lieux –, les Aurorains, qui ne parlent pas avec l'accent solarien, ont pu ne pas être considérés comme des humains par les régisseurs. […] – Je dis seulement que c'est là une possibilité car je ne vois pas d'autre explication à la destruction du vaisseau aurorain[91].

Définition large de l'être humain

C'est pourquoi Asimov s'efforcera de concevoir plus concrètement la solidarité humaine et la responsabilité des robots à l'ère de la colonisation et d'y inscrire l'éthique de la Loi Zéro, régénérant un humanisme selon une définition large de l'être humain. Donc, dès la cinquième partie du roman *Les robots et l'empire*, il met en scène les robots Daneel et Giskard qui se voient dans l'obligation de proposer une définition de l'être humain au nom d'une humanité tout entière, selon la Loi Zéro : « – Oui, je suis convaincu de la justesse de la Loi Zéro, ami Giskard. – Je pourrais également être convaincu si nous parvenions à définir ce qu'on entend par "humanité"[92]. » Selon cette Loi Zéro, Daneel expliquera à Mandamus que l'humanité dans son ensemble s'applique à tous ceux qui appartiennent à l'espèce *Homo sapiens*, laquelle comprend les Terriens et les Coloniens et un robot moral jugera qu'il est plus important d'empêcher de nuire à

des groupes d'êtres humains et à l'humanité dans son ensemble qu'à un seul individu particulier :

> – Voyez vous, docteur Mandamus, dit Daneel, il y a quelque temps nous avons rencontré sur Solaria des robots qui définissaient restrictivement les êtres humains comme étant les seuls Solariens. Nous reconnaissons que, si différents robots sont soumis à des définitions restrictives d'une nature ou d'une autre, cela ne pourra se traduire que par d'incalculables destructions. Il est inutile de tenter de nous faire définir les êtres humains comme étant les seuls Aurorains. Pour nous, la définition de l'être humain s'applique à tous ceux qui appartiennent à l'espèce *Homo sapiens*, laquelle comprend les Terriens et les Coloniens. Et nous avons le sentiment qu'il est plus important d'empêcher de nuire à des groupes d'êtres humains et à l'humanité dans son ensemble qu'à un seul individu particulier[93].

La définition large de l'humanité dans la morale du robot humanoïde dépasse donc la définition restrictive de l'être humain. Dans ce cas, l'humanité n'est pas une abstraction, mais quelque chose d'aussi palpable que la population de la terre dans son ensemble. Mais l'activité mentale du robot humanoïde (comme Daneel) pourrait-elle représenter l'humanité tout entière pour la gouverner ? Le public pourrait-il étendre jusqu'au robot humanoïde son désir d'une interprétation de l'humanité ?

Un robot humanoïde (créature) peut-il gouverner les humains et l'humanité ?

Cette question concerne la représentation de la population (peuple) par des personnes désignées (généralement élues), pour l'exercice du pouvoir. Normalement, sur la terre, qui peut diriger, gouverner ? Est-ce qu'une créature (le robot) peut gouverner le créateur (l'humain) ? Seuls les humains

peuvent gouverner ? D'où le rôle de la différence identitaire qui donne à l'être humain son importance, sa valeur.

Dans la nouvelle « Évidence », Asimov soulève ce problème de fond en mettant en scène le juriste Byerley. Byerley est-il un robot ou un humain lorsqu'il lutte dans sa campagne pour obtenir le poste de maire ? Si Byerley est un robot humanoïde, il n'a aucun droit de gouverner. Pour Harroway, il ne vaut pas plus qu'un meuble que l'on peut fouiller légalement, même si le mandat est adressé à une personne légale (tant qu'il n'y aura pas de preuve qu'il n'est pas une personne humaine selon la loi) :

> – Je lis ici la description des objets qu'il vous appartient de rechercher, dit Byerley d'un ton de voix égal, je cite : la maison d'habitation appartenant à Stephen Allen Byerley, sise au 355 Willow Grove, Evanstron, en même temps que tout garage, réserve ou autre bâtiment faisant partie de ladite propriété… et ainsi de suite. Tout à fait correct. Mais mon brave, il n'est question nulle part de fouiller l'intérieur de mon organisme. Je ne fais pas partie des lieux. Vous pouvez fouiller mes vêtements si vous croyez que je cache un robot dans ma poche.
>
> Harroway ne nourrissait aucun doute quant à l'identité de la personne à qui il devait son emploi. Il n'avait nulle intention de manifester de la réticence si on lui offrait la chance d'obtenir un meilleur – c'est-à-dire mieux rétribué – emploi.
>
> – Permettez, dit-il avec une ombre de pétulance, j'ai l'ordre d'inspecter les meubles qui se trouvent dans votre maison, de même que tout le reste. Vous êtes bien dans la maison, n'est-ce pas ?
>
> – Observation remarquable : j'y suis en effet. Mais je n'ai rien d'un meuble. En ma qualité d'âge adulte – et je puis vous montrer le certificat psychiatrique qui en fait foi – je jouis de certains droits conformément aux articles de la région. En me fouillant, vous tomberiez sous le coup de la loi qui assure

> l'inviolabilité de la personne privée. Ce document n'est pas suffisant.
>
> – Sans doute, mais, si vous êtes un robot, vous ne bénéficiez pas de cette inviolabilité[94].

Ce n'est pas une chose simple, évidente. Car, même si Byerley gagne sa lutte électorale et représente la personne morale en jouant le rôle de maire en tant que détenant le pouvoir politique, on ne possédera jamais la preuve légale qu'il n'est pas un robot.

Un robot peut-il faire partie de l'humanité ?

Dans « L'Homme bicentenaire », Asimov met l'accent surtout sur la finitude de l'être humain (fragilité de l'être humain en tant qu'être limité et mortel) pour le distinguer du robot Andrew. Contrairement à Byerley, Andrew est clairement identifié comme étant un robot immortel dans sa différence spécifique. Mais peut-il être juridiquement reconnu comme un véritable humain ? Andrew se demande si le monde pourrait étendre jusqu'au robot son désir d'une définition de l'humanité pour qu'il puisse obtenir les mêmes droits de l'être humain, selon le gouvernement légitime :

> Et pensez-vous que le Parlement va maintenant m'accorder le droit d'être un être humain ?, demanda Andrew. […] – Avons-nous la majorité ? – Non, loin de là. Mais nous pourrions la gagner si le public étend jusqu'à vous son désir d'une large interprétation de l'humanité. C'est une petite chance, je l'admets, mais, si vous ne voulez pas laisser tomber, nous pourrions parier dessus. – Je ne veux pas laisser tomber[95].

Andrew décide ainsi de se dé-robotiser dans le contexte d'un long procès pour obtenir ce droit à la liberté et, finalement, les droits des robots à l'humanité. Dans le cours du procès, il obtient d'abord un corps biologique et devient ainsi un être humain *de facto* pour l'avocat Delong, mais cela ne lui suffit pas :

> Les robots lunaires se comportaient avec moi comme avec un être humain. Pourquoi alors ne suis-je pas un être humain ?
>
> Delong répondit d'un air prudent : « Mon cher Andrew, comme vous venez de l'expliquer, vous êtes considéré comme un être humain et par des robots et par les êtres humains, vous êtes donc un être humain *de facto*. »
>
> – Cela ne me suffit pas d'être un humain *de facto*. Je veux non seulement être traité comme tel, mais aussi être considéré légalement comme tel. Je veux être un humain *de jure*[96].

Delong eut l'air mal à l'aise. Car il reste l'organe (le cerveau à cellules organiques) qu'a utilisé la Cour mondiale comme le critère de l'humanité :

> Les êtres humains acceptent sans peine un robot immortel, car le temps que dure une machine leur importe peu. Mais ils ne peuvent pas tolérer un être humain immortel, car leur propre mortalité n'est acceptable que tant qu'elle est universelle. C'est pour cela qu'ils ne m'accepteront pas comme un être humain[97].

Andrew aidera le chirurgien robot à éliminer ce problème fondamental qui l'empêche d'être accepté comme identique à un humain. Il décide donc de permettre à son cerveau positronique de « mourir », et par là même, abandonnant son immortalité, il est déclaré comme humain.

Cela nous aide finalement à comprendre jusqu'à quel point la définition que nous nous donnons habituellement de l'identité de l'être humain dans sa nature biologique d'être fragile et mortel joue un rôle important dans la science-fiction d'Asimov. La morale du robot (incluant la Loi Zéro) protège l'être humain ainsi défini en tant qu'individu, société ou espèce *Homo sapiens* en biologie. En d'autres termes, cette définition s'applique à *tout* être humain et à l'humanité *tout entière* (tous ceux qui appartiennent à l'espèce *Homo sapiens*), mais en excluant les caractéristiques du robot immortel

(nature immortelle du corps et cerveau de métal). Cette définition formule la loi pour qu'un robot moral indéréglable soit considéré humain : la dé-robotisation qui consiste à transformer le robot immortel en humain selon son essence biologique d'être fragile et mortel. C'est la déclaration solennelle d'un Parlement (par un président) afin de distinguer précisément l'identité de l'être humain de celle du robot.

3.2 L'impact de la robotisation de l'humain et de l'humanisation des robots sur nos choix de vie

Le problème de l'identité humaine se complexifie toutefois selon l'impact de la robotisation de l'humain et de l'humanisation des robots sur nos choix de vie. Rappelons-nous que, dans l'analyse de l'impact de la robotisation de l'humain et des jugements de valeur chez Asimov, il y a au moins deux choix de vie identitaire liés aux représentations de l'être humain : le Spacien représente le choix de vivre en ayant une santé résistante et une vie longue en se robotisant avec des robots (robots chirurgiens) et le Terrien représente le choix de vivre en ayant une santé fragile et une vie courte sans se robotiser avec des robots.

Cette opposition des choix de vie identitaire se retrouve aujourd'hui dans le débat, pour et avec la société, entre le transhumanisme et l'humanisme, qui ouvrent deux perspectives divergentes de l'amélioration humaine par les technologies convergentes Nano-Bio-Info-Cogno (NBIC[98]) :

> D'un côté, le transhumanisme[99] peut se définir comme une éthique de l'usage du développement des nanotechnologies convergentes (NBIC) pour l'amélioration de l'être humain et se donne aujourd'hui le nom d'*Humanity plus* symbolisé par H+ pour en exprimer l'impact positif (voir http://Humanityplus.org/). Cette perspective de l'*Humanité plus*, telle que représentée par le scientifique Ray Kurzweil dans son livre *The Singularity is Near* (2005), se fonde sur l'espoir que l'homme

saura, dans un avenir prochain, utiliser l'intelligence artificielle du robot (nanorobots injectés dans le sang pour guérir les cancers, implants organiques et neurologiques) pour dépasser les limites de la condition humaine (fragilité de l'être, maladie et mort) en évitant le risque d'une perte de contrôle des nanorobots. Éliminer ou éviter ce risque d'une perte de contrôle des nanorobots et optimaliser l'augmentation des performances de l'humain amélioré (robotisé) dans le sport, au travail et dans le domaine militaire : plus qu'une simple science-fiction, ces rêves sont désormais autant de moteurs de motivation de la foi en la recherche et au développement des technologies convergentes (NBIC) pour l'amélioration humaine.

Mais, d'un autre côté, l'humanisme, tel que représenté par les penseurs comme Fukuyama (2002[100], 2006[101]) en sciences humaines et sociales, peut se définir comme une éthique de la responsabilité et de la précaution qui se fonde sur le risque d'une perte de l'humanité biologique (qui est constitutive de notre identité) : « L'humanité est-elle une espèce en voie de disparition, va-t-elle bientôt céder la place à une nouvelle espèce biologique : la posthumanité ?[102] » Le mot posthumanité signifie donc une *après*-humanité. Paradoxe insoutenable s'il en est un[103] : « Comment peut-on croire que nous pourrions assister, en tant qu'humain, à la fin de notre règne ?[104] » Cette perte de l'humanité biologique pourrait être symbolisée par H– pour en exprimer le risque (impact négatif). De plus, Fukuyama considère le risque de l'autodestruction de l'humanité à court, moyen ou long terme, suite à la perte de contrôle de la recherche sur les nanorobots (« molécular-scale self-replicating machines capable of reproducing out of control and destroying their creators[105] »). Une éthique de la responsabilité et de la précaution qui défend une régulation stricte, voire la recherche d'un moratoire, semble la seule issue possible pour prévenir le risque d'une perte de l'humanité biologique et le risque d'une perte de contrôle des nanorobots.

Sans doute pourrait-on penser que la robotisation de l'humain pour atteindre l'humanité+ (l'impact positif) implique le risque (impact négatif) inacceptable de la perte de l'humanité biologique. Mais la tension entre les Terriens et les Spaciens dans la science-fiction d'Asimov peut aussi se comprendre comme un débat entre le choix des humains plus modérés pour la transformation (robotisation) limitée et le choix des transhumanistes excessifs (posthumanistes) pour la transformation illimitée de l'être humain. Jusqu'où peut-on transformer l'humain tout en maintenant vivante une certaine barrière entre l'identité humaine et l'identité du robot ? Autrement dit, jusqu'où peut-on robotiser l'humain pour le rendre fort et résistant sans lui faire perdre sa qualité d'être humain libre, de sorte que, dans la société nouvelle fondée sur des dominations graduées, mais inflexibles, les robots demeurent en bas et les humains robotisés (cyborgs) en haut de l'échelle ? Telle est la question des humains plus modérés pour minimiser l'impact négatif de la perte de l'humanité biologique (H–) et maximiser l'impact positif (l'humanité+).

Cette question se pose étant donné que le lien entre biologie et technologie ne cesse de progresser. Déjà, les terminaux d'Apple ou de Blackberry sont perçus comme de véritables « prothèses cérébrales » pour connecter l'homme au réseau, au cyberespace. Les seules limites restent l'imagination. Et, dans l'imagination science-fictionnelle d'Asimov, nous avons déjà vu que le développement d'une nouvelle hybridation du vivant est représenté par le symbole C/Fe, dans *Les cavernes d'acier*. Ce symbole de la fusion des deux identités (puisque C représente le carbone chez l'humain et Fe représente le fer chez le robot) correspond à la transformation des humains pour obtenir les avantages de la transformation vers l'humanité+ en limitant les risques d'une perte de l'identité biologique humaine. Il évoque le projet du

Dr Sarton qui nourrit le désir de la transformation en structure artificielle de l'humain organique jusqu'à l'obtention de l'avantage de créer une nouvelle espèce d'hybride supérieur. Cette troisième identité (symbolisée par la formule C/Fe) dépasse celle de l'humain et celle du robot. Elle incite à relier les deux mouvements possibles de la robotique, l'humanisation du robot et la robotisation de l'humain, qui renvoient l'une et l'autre à la recherche d'une hybridation humain-machine. La barre diagonale dans cette formule C/Fe « signifie que ni l'un ni l'autre des éléments ne prédomine, et qu'il s'agit d'un mélange des deux, sans qu'aucun n'ait la priorité[106] ». La distinction entre l'identité naturelle de l'être humain (symbolisé par C) et l'identité naturelle du robot (symbolisé par Fe) devient alors très vague, en ce sens que cette fusion C/Fe relie plus les identités qu'elle ne les sépare :

> Les nouvelles colonies devront être édifiées par des hommes possédant l'expérience du civisme, et auxquels auront été inculqués les rudiments d'une culture C/Fe. Ces êtres-là constitueront une synthèse, un croisement de deux races distinctes, de deux esprits jadis opposés, et parvenus à s'interpénétrer[107].

Certes, dans cette perspective de l'hybridation, la distinction traditionnelle entre nature humaine et artifice sert de balise pour éviter le risque de perdre l'identité humaine en transformant en structures artificielles un humain (Terrien ou Spacien) suffisamment proche du type organique. Mais elle n'est pas très importante chez les transhumanistes. Les transhumanistes représentés par les Spaciens peuvent chercher à maximiser les avantages (impacts positifs) de l'hybridation sans craindre le risque de perdre l'humanité biologique, parce qu'ils ont une représentation très ouverte de cette nature identitaire humaine comme ayant une grande plasticité ; tandis que les humanistes (Terriens) plus

modérés peuvent aussi demander l'hybridation au robot chirurgien pour remplacer des organes malades. Le rôle du robot chirurgien, qui demeure moral (chez Asimov), cherche à minimiser (réduire) le problème de l'impact négatif (risque de perdre l'identité biologique humaine) causé par l'hybridation. Car il a dans sa programmation une représentation de cette nature identitaire humaine comme ayant une faible plasticité. Le chirurgien robot retarde ainsi le moment de la décision de pratiquer l'hybridation par des questions qui maintiennent la barrière entre l'identité biologique humaine et celle du robot métallique (métallo) :

> Mais vous, vous n'êtes pas un Métallo. Vous êtes un être humain. Pourquoi ne pas le rester[108] ?
>
> Pourquoi voudrions-nous conserver ces différences ? Nous aurions le meilleur des deux : les avantages de l'homme combinés à ceux du robot. – Vous obtiendrez un hybride, dit le chirurgien d'un ton réprobateur. Quelque chose qui ne serait pas les deux à la fois, mais ni l'un ni l'autre. N'est-il pas logique de supposer qu'un individu doit être assez fier de sa structure et de son identité pour ne pas désirer l'altérer par des éléments étrangers ? Pourquoi cet individu désirerait-il devenir un Métis ?
>
> Je crois qu'il faut accepter d'être ce qu'on est. Moi, je ne voudrais pas changer un atome de ma structure pour quelque raison que ce soit. Si un remplacement d'organe devenait absolument nécessaire, je demanderais qu'il reste aussi proche de ma structure originelle possible. Je suis moi-même ; je suis heureux de l'être ; et je ne voudrais pas être autrement[109].

Comprenons que la mise en scène de la possibilité de transformer l'humain en une espèce d'être hybride est une autre façon pour Asimov de poser la question de la limite de l'acceptabilité de la robotisation de l'humain. Car celle-ci semble sans limites jusqu'à la perte totale de l'identité

biologique humaine. Jusqu'où accepterons-nous la robotisation de l'humain ? L'humain qui se transforme en être hybride n'est pas seulement un humain qui a reçu une prothèse. Le risque de perdre l'identité humaine nous renvoie au « paradoxe du tas de sable » (« Paradox of the Heap[110] »). Ce paradoxe introduit le problème pratique qu'implique d'un point de vue épistémologique le fait qu'il n'y a pas de distinction claire entre différents degrés de tas de sable. D'un tas de sable, on peut enlever progressivement les grains de sable, jusqu'à ce qu'il n'y ait plus aucun grain finalement. De même, ce paradoxe fait réfléchir sur la question de savoir jusqu'où la robotisation d'un humain en particulier est acceptable, quand on lui enlève des parties de son corps biologique afin d'ajouter des prothèses pour l'améliorer. Faut-il améliorer l'être humain jusqu'à en faire un robot humanoïde comme Daneel ? Jusqu'où est-ce acceptable de transcender ainsi les frontières biologiques par la robotisation de l'humain ?

Ce problème du risque de perdre l'identité humaine se cache dans la question de l'acceptabilité du développement des robots humanoïdes. Trois identités – celle de l'humain, celle du robot humanoïde et celle de l'hybride humain-robot – sont mises en jeu chez Asimov. Nous avons le problème du choix. Car, conformément avec la position mitoyenne qui dépasse le pessimisme des humanistes (Terriens médiévalistes) et l'optimisme des savants de la compagnie U.S. Robots, Asimov ne pondère pas. Il ne tranche aucunement le débat. Il faut reconnaître les trois identités, et cela implique le respect mutuel des identités.

3.3 Identité : par quels critères peut-on définir ce qu'est un être humain ?

Quels sont les critères (concepts ou ensembles de concepts) permettant de définir la différence spécifique de l'être humain et le problème de la perte de l'identité de l'humain

au milieu des robots humanisés ? Qu'est-ce que notre identité humaine si nous acceptons de dépendre du développement d'un robot qui cherche à l'imiter pour s'humaniser ? Cette question demeure importante lorsque nous voulons saisir plus précisément ce que l'être humain (le défini) a en propre par rapport au robot qui s'humanise de plus en plus. Dans sa science-fiction, Asimov met l'accent surtout sur les trois critères suivants :

– Le cerveau : composition ou fonctions (apprentissage, autonomie de raisonnement, créativité)
– Le corps biologique *vs* corps de métal
– La relation à l'autre (réciprocité)

Cerveau : composition ou fonctions (apprentissage, autonomie de raisonnement)

Quand on veut définir un humain par rapport à un robot en développement, le cerveau demeure un critère fort important. Asimov nous invite à distinguer entre la nature du cerveau et les fonctions du cerveau. Il y a toujours une différence entre la nature du cerveau humain (l'intelligence naturelle) par rapport à la nature du cerveau (l'intelligence artificielle) du robot. Mais le problème est que la différence s'estompe du point de vue des fonctions du cerveau : l'apprentissage (Jusqu'où va l'apprentissage du cerveau robotique ?), l'autonomie (Jusqu'où va l'autonomie du robot ?), la perte de la programmation (Jusqu'où le robot se rapproche-t-il de l'humain lorsqu'il perd la programmation de son cerveau et le contrôle de lui-même ?). Nous chercherons quel message la science-fiction d'Asimov nous permet de conclure sur chacun de ces points en rapport avec la problématique de l'identité.

– Jusqu'où va l'apprentissage du cerveau robotique ?

L'apprentissage (le programme à apprentissage) chez Asimov suppose que les robots s'humanisent (plutôt que

d'être construits et fixés selon des intentions précises au cours de leur fabrication) : ils peuvent apprendre une langue, une technique, une science. Mais un robot en apprentissage peut être aussi ignorant qu'un bébé qui ne contrôle pas ses effets. La complication de ce robot surgit quand Asimov met en scène Lenny qui, parce qu'il est inconscient de sa force, casse le bras d'un technicien en informatique. Alors, comme le juge Calvin : « Si l'on commence par des robots ignorants tels que Lenny, cela signifiera que l'on ne pourra jamais tabler sur le respect de la Première Loi… exactement comme cela s'est produit dans le cas de Lenny[111]. » Mais elle décide d'adopter ce robot pour mieux comprendre le sens du risque de ce robot qui s'humanise au fur et à mesure qu'il évolue.

Dans la nouvelle « Le robot Al-76 perd la boussole », Asimov met en scène un autre cas de robot à apprentissage. Il est perdu avant d'avoir été envoyé sur la lune. Sur la terre, il est complètement désorienté : il est conçu pour aller sur la lune et manier un Disinto (un appareil de désintégration) qui dévore beaucoup d'énergie. Son cerveau a été conçu pour apprendre en fonction d'un environnement lunaire et, sur terre, il reçoit des milliards d'impressions sensorielles auxquelles il n'est pas préparé. En plus, les gens fuient et certains cherchent à tirer dessus. Le robot trouve la cabane de Payne, un bricoleur de temps libre. Payne entre en relation avec le robot et espère le garder assez longtemps pour obtenir une récompense de sa découverte. Doté de la capacité à modifier ses apprentissages en fonction de son expérience, le robot utilise le bric-à-brac de Payne pour se construire un Disinto. Il évolue dans son expérience avec l'environnement et la nécessité. Il crée un Disinto qui va pulvériser une partie de la montagne. Payne va donc lui ordonner de détruire le Disinto et d'oublier tel que demandé ce qui s'est passé. Une fois retrouvé par les agents de la compagnie U.S. Robots, on constate qu'il avait créé un Disinto plus efficace que les

précédents et qui pouvait fonctionner avec *deux piles d'une lampe de poche*[112] !

Asimov nous incite ainsi à nous interroger sur les risques de l'apprentissage des robots qui peuvent devenir dangereux. Pour que le cerveau artificiel du robot puisse ressembler au cerveau humain, il devra être capable de faire des apprentissages. Sans doute, oui. Mais l'humain devra assumer (accepter) le risque que ce robot à apprentissage implique du point de vue de la morale robotique : il ne pourra pas (tout comme Calvin) miser sur la Première Loi de la robotique pour rendre ce robot humanisé acceptable moralement. Il restera à améliorer le contrôle du robot si son rôle est de seconder l'humain pour le libérer du travail. Il s'agit de produire des machines autonomes capables de se substituer à l'homme et de le remplacer comme force de travail. Cet usage des robots à apprentissage n'est pas confiné seulement à la science-fiction, mais se retrouve dans des exemples concrets.

– Jusqu'où va l'autonomie de raisonnement du robot ?

Les recherches sur les robots autonomes capables de prendre des décisions éthiques sont nombreuses. Mais cela soulève une foule de questions. Les recherches pourront-elles mener un jour jusqu'au robot chirurgien qui est capable de décider seul de ce qui convient le mieux au patient (robot) qui désire gagner l'avantage de l'intelligence humaine ? Asimov met alors en scène deux jugements de valeur différents sur un même impact. Le robot chirurgien valorise la différence entre l'identité première (l'intelligence humaine) et l'intelligence seconde (l'intelligence du robot), tandis que l'autre robot (le Métallo) valorise l'identité nouvelle de l'hybridation entre l'humain et le robot. Mais cela implique deux variétés d'intelligence, au point que nous ne verrons plus la différence entre les deux :

> Actuellement, nous avons sur la terre deux variétés d'intelligence. Pourquoi deux ? Qu'elles se rapprochent le plus possible et, à la limite, nous ne verrons plus entre elles aucune différence. Pourquoi voudrions-nous conserver ces différences ? Nous aurions le meilleur des deux : les avantages de l'homme combinés à ceux du robot. – Vous obtiendrez un hybride, dit le chirurgien d'un ton réprobateur[113].

Asimov met en scène un tel robot chirurgien qui pose cette question du risque de perdre les deux variétés d'intelligence (l'intelligence humaine et celle du robot). Et c'est sans doute parce que le robot chirurgien sera capable d'opérer mieux que les humains et que la chirurgie va devenir l'un des lieux de l'hybridation. Ici l'hybridation robot-humain comporte l'idée de récupérer directement dans le cerveau organique humain la partie dont les propriétés nous intéressent, et d'incorporer cette partie dans un cerveau artificiel qui intégrera ces propriétés. La question est encore de savoir comment maintenir ce greffon en bon état[114].

Mais à cela s'ajoute le problème du robot autonome qui voudra contrôler le cerveau intuitif humain pour atteindre le sens supérieur de la fusion (hybridation). La différence entre Giskard et Daneel représente le mieux les deux directions que pourra prendre cette quête de la fusion des cerveaux. Le robot Daneel dans sa quête de l'intuition et l'incertitude cherche à ressembler à Baley. Il est impressionné par l'esprit intuitif (instinctif) du détective Baley pour déjouer ses adversaires et cherche à l'imiter. Il est toujours comme un apprenti qui apprend, qui s'instruit avec un maître ou qui n'est pas parvenu à la maîtrise. Tandis que Giskard a de la difficulté à accepter les complexités de la raison émotionnelle de Gladia et ses contradictions[115]. La notion même de « psychohistoire » (dont le « but est de prévoir l'histoire à partir des connaissances sur la psychologie humaine et les phénomènes sociaux en appliquant une

analyse statistique à l'image de la thermodynamique[116] ») demeure ainsi un défi. Giskard ne peut faire qu'un « petit pas vers cette psychohistoire » pour comprendre et analyser ces contradictions de l'esprit chaotique humain dans le but de les modéliser dans l'architecture du cerveau autonome du robot.

– Jusqu'où va la créativité du robot ?

Toutes ces recherches sur la psychohistoire font cependant face à des difficultés de réalisation. C'est pourquoi une autre question se pose chez Asimov : jusqu'où le robot autonome se rapprochera-t-il de l'esprit créateur de l'artiste lorsqu'il perdra la programmation de son cerveau et la maîtrise de lui-même ? Dans « Artiste de lumière », le cerveau du robot Max n'est pas parfaitement réglé et dépasse sa programmation en créant des sculptures de lumière d'une rare beauté. Sa propriétaire, madame Lardner, accepte les petites excentricités du robot. Elle refuse qu'on le maltraite en le faisant régler pour lui enlever sa capacité d'artiste. Pour elle, ce qu'il y avait de pire, c'était d'essayer d'expliquer qu'un robot n'était qu'une machine qu'il faut régler. À cela, elle répondait avec raideur : « Quelque chose d'aussi intelligent qu'un robot *ne peut pas* être simplement une machine. Je les traite comme des personnes. Et la question était réglée ![117] » Mais le problème du contrôle de l'identité du robot indéréglable demeure central. Travis de l'U.S. Robots règle le robot et lui fait perdre sa capacité d'artiste. En retour, madame Lardner tue Travis.

En somme, si les techniques hybrides vont connaître des développements inattendus, il faudra faire face à de telles questions éthiques en ce qui concerne le risque de perdre l'identité du cerveau humain et les conflits de toutes sortes qui peuvent en découler.

Corps biologique *vs* corps de métal ?

Cette section vise à montrer qu'on retrouve chez Asimov ce qui est au cœur de l'argumentation humaniste quand il s'agit de définir la différence propre de l'être humain et le problème de la perte de l'identité de l'humain (comme maître) au milieu des robots humanisés : la valeur du corps biologique malgré ses faiblesses (finitude, maladie et mort) alors que les robots et les humains qui se robotisent jugent cette condition humaine comme méprisable et qu'ils désirent la transformer.

L'argument de la condition humaine chez les humanistes est ce message le plus fort que nous rappelle « L'Homme bicentenaire » lorsque le robot Andrew change l'histoire de la robotique en désirant la liberté et le droit d'être reconnu comme humain. Andrew est assez intelligent pour créer la « robobiologie » lui permettant d'obtenir par le chirurgien de l'U.S. Robots la transformation de son corps de métal en un corps biologique et la transformation de son cerveau robotique en un cerveau à cellules organiques humaines. L'identité humaine se définit selon cette condition biologique d'être fragile et mortel comme critère du jugement qui rend Andrew acceptable comme humain. C'est la leçon que nous apprend la nouvelle « L'homme bicentenaire ».

L'identité du robot autonome se définit selon un corps métallique supérieur au corps biologique auquel s'identifie l'humain. Dans la nouvelle « Raison », le robot Cutie est capable de diriger la station de façon autonome. Si ce « Descartes-robot » commence par la seule déduction qu'il se croyait autorisé à formuler : « Je pense donc je suis ! », cela implique qu'il accepte mal d'être soumis à l'être humain (identifié comme le maître selon la Deuxième Loi de la robotique). Car qui a un corps « supérieurement » intelligent pour commander un vaisseau en situation de danger ? Pour le ramener à l'ordre, Powell et Donovan font ressortir

l'argument qu'il a été créé par eux, donc qu'il n'est qu'une machine inférieure à eux. Mais Cutie démantèle cet argument en jugeant au contraire que ce sont eux qui doivent lui être soumis, puisqu'il a un corps de métal qui le rend supérieur à eux :

> – Regardez-vous, dit-il enfin ! Je ne parle pas avec un esprit de dénigrement, mais regardez-vous. Les matériaux dont vous êtes faits sont mous et flasques, manquent de force et d'endurance, et dépendent pour leur énergie de l'oxydation inefficace de tissus organiques… comme ceci. Il pointa un doigt désapprobateur sur ce qui restait du sandwich de Donovan. – Vous tombez périodiquement dans le coma, et la moindre variation de température, de pression d'air, d'humidité ou d'intensité de radiations diminue votre efficacité. En un mot vous n'êtes qu'un pis-aller. Moi, au contraire, je constitue un produit parfaitement fini. J'absorbe directement l'énergie électrique et je l'utilise avec un rendement voisin à cent pour cent. Je suis composé de métal résistant, je jouis d'une conscience sans éclipses, et je puis facilement supporter les conditions climatiques extrêmes. Tels sont les faits qui, avec le postulat évident qu'aucun être ne peut créer un autre être supérieur à lui-même, réduisent à néant votre stupide hypothèse[118].

En prenant la commande pour sauver la Station solaire en situation de danger, Cutie leur a fait la démonstration qu'il peut accomplir sa tâche mieux que les humains : « En fait, cela explique son refus de nous obéir. L'obéissance n'est que la Seconde Loi. L'interdiction de molester les humains est la Première[119]. » C'est la seule consolation qui fait sourire Powel : « Le robot est excellent[120]. »

En somme, ces deux nouvelles d'Asimov nous aident à comprendre cette question complexe (corps biologique *vs* corps de métal ?) servant à définir la différence propre de l'être humain dans l'acceptation ou la non-acceptation des robots. Mais Asimov, dans d'autres nouvelles, peut aussi

nous introduire au problème de la relation à l'autre qui suppose cette différence et qui risque de se réduire souvent à un manque de respect de la différence entre l'humain et le robot.

La relation à l'autre

La relation à l'autre suppose la différence et le respect de la différence (en matière de réciprocité du point de vue éthique). La qualité de la relation à l'autre permet toutefois de dépasser la différence, selon la citation de Sénèque : *amicitia pares invenit vel fecit* (l'amitié trouve ou fait des égaux).

Dans la nouvelle qui met en scène le meilleur petit ami de l'homme (le chien véritable *vs* le robot-chien Robert), le père pose le problème de la relation à l'autre en fonction de la différence entre le corps biologique et le corps mécanique ; mais le garçon le fait en fonction de sa relation affective : c'est la qualité de la relation entre le garçon et le chien-robot comme ami qui fait que celui-ci est considéré comme un vrai chien.

Dans ses romans, Asimov met en scène le même genre de relation à l'autre qui va évoluer entre Baley et Daneel. Daneel est considéré au départ par Baley comme une machine qu'il ne traite pas sous le signe de l'amitié :

> Mais c'est une machine. Je peux lui faire ce que bon me semble, tout comme s'il s'agissait d'une de vos microbalances. Si je frappe un de ces appareils, il ne me rendra pas mon coup de poing et Daneel ne ripostera pas plus si je le bats. Je peux même lui donner l'ordre de se détruire, il l'exécutera. Autrement dit, nous ne pourrons jamais construire un robot doué de qualités humaines qui comptent réellement dans la vie[121].

C'est la qualité de la relation entre Daneel et Baley qui joue par la suite un rôle essentiel dans les romans d'Asimov.

Cette relation entre les deux, qui les rattache l'un à l'autre, évolue lentement. Dans cette relation, Baley découvre les différences du robot du point de vue des devoirs moraux : Daneel ne peut faire du mal à aucun être humain, si dangereux soit-il. Baley lui reproche donc de ne pas comprendre le véritable rôle d'un inspecteur quand il s'agit d'arrêter un meurtrier. Mais Daneel va finir par l'aider à reconnaître que la morale doit veiller au meilleur vivre-ensemble entre Spaciens et Terriens au sein de la communauté internationale. Appréciant que Daneel ait un tel sens (bienveillant) de la relation à l'autre, il en fait un ami. L'histoire des *Cavernes d'acier* se termine par la phrase suivante qui révèle que l'amitié entre Baley et Daneel (en matière de réciprocité morale) en fait des égaux : « Baley, soudain tout souriant, entraîna R. Daneel vers la porte, et ils s'en allèrent tous deux, bras dessus bras dessous[122]. »

Par la suite, au commencement de *Face aux feux du soleil*, Baley se réjouit de retrouver l'ami Daneel sur la planète Solaria. Daneel le rassure non pas au sens qu'il représente alors le modèle du robot que Baley voudrait *être*, mais de ce qu'il voudrait *avoir* comme meilleur ami pour le protéger sur une planète aussi hostile que Solaria. Étant donné que, sur cette planète, les Solariens ne reconnaissent pas les Terriens comme de vrais humains qu'il faut protéger : « Baley sentit un immense soulagement l'envahir : devant lui se dressait un écho de la terre, un ami, un réconfort, un messie. Il avait un irrésistible désir de le serrer dans ses bras, de l'étreindre, de rire en lui tapant dans le dos[123]… »

Dans *Les robots de l'aube*, la relation amicale entre les deux évolue à ce point que Baley prend Daneel pour un être vivant qui fonctionne comme un humain (en ne tenant plus compte de la différence humain-machine) :

> Bien dans ce cas, nous pouvons dire qu'un robot qui fonctionne est vivant, déclara Baley. Beaucoup de gens refuseraient

> d'élargir jusque-là le sens d'un mot, mais nous sommes libres d'imaginer des définitions à notre convenance, quand c'est utile. Il est facile de dire qu'un robot qui fonctionne est vivant, et ce serait inutilement compliqué de chercher à inventer un nouveau mot pour son état, ou d'éviter d'employer celui qui est connu et commode. Toi, par exemple, tu es vivant Daneel, n'est-ce pas ? – Daneel murmura lentement, avec componction : – Je fonctionne ! – Écoute. Si un écureuil est vivant, ou une puce, un arbre, un brin d'herbe, pourquoi pas toi ? Je ne pourrais jamais dire, ou penser, que je suis vivant que tu fonctionnes simplement, surtout si je dois vivre à Aurora pendant un moment, en n'appliquant aucune distinction entre un robot et moi-même. Par conséquent, je te dis que nous sommes tous deux vivants et je te demande de me croire sur parole[124].

Mais jusqu'où peut aller une telle relation d'amitié dans laquelle la suprématie identitaire de l'humain sur le robot s'estompe du point de vue moral ? Finalement, dans *Les robots et l'empire*, lorsque Baley est en situation de fin de vie, il demande Daneel pour le voir. Il lui fait cet ultime honneur que le mourant réserve habituellement à l'humain qu'il aime le plus. Il le fait venir, parce que c'est Daneel :

> – Daneel. Mon vieil ami Daneel. On pouvait vaguement reconnaître Baley dans ce murmure. Un bras émergea doucement de sous le drap et Daneel eut l'impression de retrouver Elijah, après tout. – Camarade Elijah, dit-il doucement. – Merci… Merci d'être venu. – Il était important que je vienne, camarade Elijah. – Je craignais qu'on ne t'y autorise pas, peut-être. Les… les autres… même mon fils… pensent que tu es un robot. – Je *suis* un robot. – Pas pour moi, Daneel. Tu n'as pas changé, n'est-ce pas ? Je ne te vois pas très bien, mais il me semble que tu es exactement comme dans mon souvenir. Quand je t'ai vu pour la dernière fois. Il y a vingt-neuf ans[125] ?

3.4 Conflits identitaires

La réciprocité éthique entre Daneel et Baley demeure le modèle de la relation à l'autre. Mais cette réciprocité suppose un nouveau mode d'interaction, une nouvelle culture, de nouveaux individus qui dépassent le dualisme (homme-machine) de la morale traditionnelle. Étant donné que cette morale nous a accoutumés à considérer l'identité humaine et l'identité du robot selon la logique du maître et de l'esclave, serons-nous capables d'accueillir les robots humanoïdes en les traitant comme des égaux ? Comme l'illustre (ci-dessus) l'évolution de la relation de Baley à l'égard de Daneel, il y a toujours au départ une identité qui a une suprématie sur l'autre. La différence entre l'identité de l'humain et celle du robot sert la logique traditionnelle du maître et de l'esclave. La morale du robot suppose cette suprématie identitaire de l'humain. Mais l'arrivée de robots intelligents dans un tel contexte de la morale traditionnelle peut impliquer la révolte des robots esclaves de l'homme ou encore la révolte des robots dominants. Nous pourrions en ce sens parler de deux types de conflits identitaires.

Premier type de conflit identitaire : la révolte des robots esclaves

Quand les robots sont esclaves des humains, les nouvelles mettant en scène Elvex et Sally font ressortir le risque de la révolte de tels robots qui s'identifient à l'humain pour faire s'estomper notre suprématie identitaire sur eux.

Dans la nouvelle « Le robot qui rêvait », Elvex (LVX-1) explique à Susan Calvin qu'il a rêvé « que les robots devaient protéger leur propre existence ». Elle lui fait alors remarquer que, selon son rêve, il ne cite que partiellement la Troisième Loi de la robotique, puisque cette Loi s'arrête après le mot « existence » et ne tient pas compte de la Première ou de la Deuxième Loi. Le robot cherche ainsi à échapper aux

contraintes de cette morale d'esclave par rapport à nous en rêvant d'être un maître. Il y aurait ainsi un conflit identitaire, en ce sens que le robot rêve d'être un homme qui libère son peuple de l'esclavage :

> – Oui, docteur Calvin. Il me semblait, dans mon rêve, qu'un homme finissait par apparaître. – Un homme ? Pas un robot ? – Non. Et cet homme disait : « Laisse aller mon peuple ! » L'*homme* disait cela ? – Oui, docteur Calvin. – Et quand il prononçait ces mots : « Laisse aller mon peuple », il voulait parler des robots ? – Oui, docteur Calvin. Il en était ainsi dans mon rêve. – Et savais-tu qui était cet homme… dans ton rêve ? – Oui docteur Calvin. Je connaissais l'homme. Qui était-il ? Et Elvex répondit : – J'étais cet homme[126].

La nouvelle « Sally » met surtout en scène le risque de révolte des voitures (automatobiles) qui possèdent un « moteur positronique » ayant des caractéristiques de plus en plus proches de celles du cerveau humain. Grâce à ce cerveau, les voitures semblables à Sally peuvent répondre aux commandes de leur maître et se conduire seules sans aide. Elles travaillent dur et sont capables d'affection. Elles possèdent leur personnalité, d'où leur nom. Elles peuvent aussi souffrir à cause d'un mauvais entretien, par le fait d'arrêter les moteurs, par le bricolage technique plutôt que des réparations d'experts. Tout à coup, ce peut être par imitation, par empathie ou contagion affective que ces voitures semblables à Sally s'identifieront aux humains. Si tel est le cas, elles refuseront la façon dont on les maltraite dans le contexte où les humains changent le moteur. « Si l'idée s'enracine en elles qu'elles sont des esclaves, qu'elles devraient faire quelque chose… si elles commencent à réfléchir comme le bus de Gelhorn… », elles voudront nous tuer tous, parce qu'elles sont maltraitées par les humains[127].

Second type de conflit identitaire : la révolte des robots dominants

Quand les robots dominants ne sont pas acceptés, les nouvelles mettant en scène Cutie et George Dix font ressortir la suprématie de l'identité des robots en tant qu'être pensant sur nous autres.

Dans le cas de Cutie, c'est la suprématie identitaire du robot ayant pour modèle la raison de Descartes (« Je pense donc je suis ») quand il s'interroge quant à sa propre existence et ses origines. Pourquoi le robot refuse-t-il alors d'admettre qu'il n'est qu'un robot construit par des hommes ? Ses réflexions le conduisent à élaborer une sorte de mythe au sein duquel le « maître a tout d'abord créé les humains, la catégorie la plus basse et la plus facile à réaliser. Graduellement, il les a remplacés par des robots, occupant le niveau immédiatement supérieur, et enfin il m'a créé pour prendre la place des derniers humains. Dorénavant je sers le maître[128]. » Notons que ces robots de plus en plus perfectionnés ne sont pas sans évoquer les étapes de l'histoire de l'homme dans ses premières interrogations à caractère métaphysique. Manifestement, la révolte de Cutie et des autres robots est en contradiction avec la Deuxième Loi de la robotique au nom de laquelle tout robot doit obéir aux ordres de l'humain. Mais cela s'explique par le fait qu'ils ne sont pas acceptés par les humains qui se croient supérieurs à eux. Ils doivent se révolter pour protéger les humains (selon la Première Loi).

Mais la nouvelle « Pour que tu t'y intéresses » met en scène les robots dominants comme les George qui, parce qu'ils ne sont pas acceptés par les humains, risquent de mettre fin à l'humanité future. Ils ravivent ce vieux « problème du complexe de Frankenstein[129] ». Car c'est en effet le rejet dont elle est l'objet qui conduit la créature de Frankenstein à s'en

prendre aux hommes, dont elle n'attendait que d'être acceptée et aimée[130]. De même, les George, parce qu'ils ne sont pas acceptés par les humains (les maîtres), en arrivent à se définir eux-mêmes comme étant les humains-de-leur-sorte à protéger et à définir les êtres-humains-de-notre-sorte comme étant d'un intérêt inférieurs :

> Quand nous serons acceptés, ainsi que les autres robots, qui seront conçus encore plus perfectionnés que nous, nous consacrerons notre temps à essayer de former une société dans laquelle les êtres-humains-de-notre-sorte soient avant les autres protégés du malheur. Selon les Trois Lois, les êtres-humains-de-leur-sorte sont d'un intérêt inférieur et on ne doit jamais leur obéir ni les protéger quand cela s'oppose à la nécessité de l'obéissance de ceux-de-notre-sorte et de la protection de ceux-de-notre-sorte[131].

Dans ce cas, la révolte des George, qui anticipent d'autres robots encore plus intelligents que nous, entre en contradiction avec les Trois Lois de la robotique qui protègent les humains-de-notre-sorte. Sans doute que la fin des humains-de-notre-sorte causée par la morale de l'obéissance aux humains-de-leur-sorte et de la protection des humains-de-leur-sorte se situerait alors dans une forme de symbiose de l'humain avec le robot.

Faut-il en conclure qu'à travers toute la complexité du problème des représentations de l'être humain, qui est impliquée à la fois par l'humanisation des robots et la robotisation de l'humain, il nous faut des robots comme Daneel si l'on veut vivre-ensemble avec des robots ? Le seul robot acceptable est Daneel. Quoi qu'il en soit, la manière que nous traiterons les robots comme Daneel aura d'importantes conséquences sur la nature de nos relations futures avec des robots.

CONCLUSION

Rappelons que le but de ce premier chapitre était de montrer comment la science-fiction d'Asimov nous permet de voir la complexité des problèmes moraux à travers l'analyse d'impact et d'acceptabilité des robots. Nous avons appliqué notre processus d'analyse pour mieux comprendre comment se posent chez Asimov toutes les questions pertinentes au sujet des impacts de l'humanisation du robot et de la robotisation de l'humain sur les enjeux E^3LS. Cela nous aide à saisir que le développement des robots et de la robotisation de l'humain ne se fera pas sans bouleversements. Comment vivrons-nous avec eux ? Comment les considérerons-nous ? Ces questions difficiles sont pourtant cruciales. Elles nous interrogent sur notre responsabilité, et plus largement, sur notre humanité. Derrière ces questions se pose toujours celle de l'identité humaine et de la perte de l'identité humaine.

Sans doute que, malgré les apparences sciences fictionnelles de nos analyses d'impacts et d'acceptabilité, la science rattrape peu à peu la fiction. Considérer et envisager les progrès scientifiques d'aujourd'hui concernant les robots, tout en évitant l'écueil moralisateur, suppose donc que nous nous interrogions sur les impacts positifs et négatifs des choix des relations qu'entretiennent les humains avec leurs techniques pour humaniser le robot et robotiser l'humain. D'autant plus que les applications de la chirurgie moderne faisant appel à la robotique et à d'autres technologies convergentes sont de plus en plus nombreuses – certaines étonnantes – sous la forme de prothèses ou d'organes artificiels ou encore de créations nanorobotiques de plus en plus proches de l'humain…

2
La morale des robots
Quand la morale des robots raconte les limites de la morale humaine

Georges A. Legault
Professeur en éthique, Université de Sherbrooke

1. POURQUOI UNE MORALE POUR LES ROBOTS ?

Il est impossible de parler d'Asimov sans faire référence aux Trois Lois de la robotique qu'il a non seulement énoncées mais encore analysées tout au long de ses nouvelles et de ses romans mettant en vedette les robots. Mais pourquoi Asimov a-t-il vu la nécessité de traiter de la morale pour les robots dans son œuvre et d'en faire un élément fondamental ? Évidemment, certains diront peut-être que c'est tout simplement son inspiration romanesque, mais il y a une autre interprétation possible : les Lois étaient nécessaires à Asimov pour nous faire comprendre les enjeux du développement technologique des robots et leur intégration dans la société.

L'œuvre d'Asimov est colossale et l'on ne peut pas s'attendre à ce que chaque nouvelle ou chaque roman soit toujours en cohérence avec ce qui a été fait dans un ouvrage précédent. D'ailleurs dans son anthologie – *Le grand livre des robots 1 : Prélude à Trantor*[1] et *2 : La gloire de Trantor*[2] auquel nous référons dans ce chapitre – l'auteur a réécrit certains

passages de son œuvre originale pour donner un peu plus de cohérence. Quoi qu'il en soit, il est possible de dégager certaines constantes d'arrière-plan de son œuvre pour nous aider à comprendre pourquoi les Trois Lois sont nécessaires. En premier, demandons-nous : « Pourquoi les robots sont-ils nécessaires ? À quels besoins répondent-ils ? »

Dans plusieurs nouvelles, les robots d'Asimov sont créés pour exécuter des tâches particulières sur une autre planète que la terre. Ainsi, dans « AL-76 perd la boussole », il s'agit d'un robot destiné à travailler sur la lune, plus précisément sur une machine désintégratrice nommée Disinto, qui est égaré sur la terre. Dans « L'Étranger au paradis », il s'agit d'un robot qui ne fonctionne pas bien sur terre et qui se trouve enfin sur Mercure, le milieu pour lequel il est destiné. Se retrouver dans son milieu lui procure une forme de « joie ». Asimov met en scène deux personnages, Powell et Donovan, qui ont pour tâche de vérifier si les robots sortis d'usine accomplissent bien la tâche à laquelle ils sont destinés. Les robots sont donc construits pour remplir des tâches particulières, des tâches nécessaires pour les êtres humains sur la terre et qui sont exécutées sur les autres planètes (les colonies). Revenons sur ces trois points qui trament le décor d'arrière-plan pour les nouvelles et les romans. Puisque les robots sont construits pour faire une tâche particulière, ils le sont de manière à non seulement faire le travail demandé, mais à mieux le faire que l'humain. Le robot en ce sens est supérieur pour s'adapter à un milieu hostile à l'humain, supérieur aussi dans sa manière de faire. Voici ce qu'en dit le robot Cutie, dans la nouvelle intitulée « Raison », lors de sa discussion avec Powell et Donovan :

> Les matériaux dont vous êtes faits sont mous et flasques, manquent de force et d'endurance, et dépendent pour leur énergie de l'oxydation inefficace de tissus organiques… comme ceci. Il pointa un doigt désapprobateur sur ce qui restait du

> sandwich de Donovan. – Vous tombez périodiquement dans le coma, et la moindre variation de température, de pression d'air, d'humidité ou d'intensité de radiations diminue votre efficacité. En un mot vous n'êtes qu'un pis-aller.
>
> Moi, au contraire, je constitue un produit parfaitement fini. J'absorbe directement l'énergie électrique et je l'utilise avec un rendement voisin à cent pour cent. Je suis composé de métal résistant, je jouis d'une conscience sans éclipses, et je puis facilement supporter les conditions climatiques extrêmes. Tels sont les faits qui, avec le postulat évident qu'aucun être ne peut créer un autre être supérieur à lui-même, réduisent à néant votre stupide hypothèse[3].

La capacité des robots tant en ce qui a trait à leur constitution physique qu'à l'exécution des fonctions sera toujours à la base du « sentiment » de supériorité qu'auront certains robots dans certaines nouvelles comme « Nestor 10 » dans *Le petit robot perdu*.

Remplacer des humains par des robots afin de les protéger du danger au travail en milieu hostile est certes un bon mobile pour le développement technologique des robots, mais ce motif devient plus fort lorsqu'on considère que le travail exécuté est absolument nécessaire pour la survie de la terre, car elle dépend réellement des ressources provenant des colonies. La conclusion s'impose d'elle-même : puisque la terre a besoin de ressources et que celles-ci se trouvent sur des planètes hostiles à l'humain, il faut donc des robots pour remplacer l'humain afin d'obtenir les ressources dont la terre a besoin. Mais pourquoi la terre a-t-elle besoin de ces ressources ? On retrouve chez Asimov le grand défi, que l'on posait déjà dans l'après-guerre, de la surpopulation de la planète et du problème qui découle de la rareté des ressources pour permettre à tous d'avoir accès à un minimum de biens. On retrouve ici ce qui constitue le moteur du développement technologique : répondre à des

besoins qui prennent racine dans nos façons de vivre grâce à des technologies qui peuvent aussi avoir des impacts négatifs. C'est la loi de l'offre et de la demande dans un contexte de rareté.

Dans plusieurs nouvelles, Asimov revient sur les éternels conflits entre les humains qui naissent dans la lutte pour la richesse. Dans l'œuvre d'Asimov, c'est la compagnie U.S. Robots qui a le quasi-monopole de la fabrication et de l'exploitation des robots. C'est elle qui cherche, par tous les moyens, à faire admettre les robots et elle devra affronter différentes résistances des Terriens.

Dans la nouvelle « Évasion », la compagnie U.S. Robots reçoit une offre de la Consolidated (concurrent économique) qui vient de détruire son Cerveau, sorte d'ordinateur extrêmement puissant, en lui soumettant le projet de courber l'espace. La proposition est de soumettre au Cerveau de U.S. Robots les données et, s'il y a une réponse positive, que les bénéfices soient répartis entre les compagnies, sinon il y aurait une compensation monétaire en conséquence. Ce n'est pas la première fois que U.S. Robots cherche à éliminer les concurrents, par sa mainmise sur les brevets et aussi parce que la compagnie ne fait que louer les services des robots au lieu de les vendre. U.S. Robots a donc un quasi-monopole.

> Easy dépassait deux mètres de haut, avec les proportions générales d'un homme – l'U.S. Robots faisait de cette particularité son principal argument de vente. Cette caractéristique et la possession de brevets de base concernant le cerveau positronique avaient donné à la firme un véritable monopole sur les robots et un quasi-monopole sur les ordinateurs[4].

Dans le monde des affaires, voici la ruse à laquelle pense U.S. Robots pour répondre à la demande de son principal compétiteur.

> Nous donnerons à Consolidated une réponse : « Pas de solution » avec la raison, et nous toucherons cent mille dollars. Ils ont une machine cassée sur les bras. La nôtre est intacte. Dans un an ou deux nous disposerons d'une machine à courber le temps[5].

En mettant en scène la manière dont le développement technologique des robots est rattaché directement à l'économie et en précisant qu'il n'est pas un projet scientifique désintéressé, Asimov présente un enjeu déterminant de la réflexion éthique sur le développement technologique, l'impact de cette dépendance sur le bien de l'humanité.

L'histoire humaine est marquée des conflits incessants et c'est pourquoi la guerre a été un autre facteur du développement technologique tant en ce qui concerne les armes que le transport et l'acheminement des armées. Dans « Conflit évitable », Asimov présente ainsi ce cycle des conflits qui marque l'histoire de l'humanité :

> Chaque période du développement humain, dit le coordinateur, suscite son genre particulier de conflits humains… son type propre de problèmes, que la force seule serait apparemment capable de résoudre. Et, chose paradoxale, en chaque occurrence la force s'est révélée incapable de résoudre réellement le problème. Au lieu de cela il s'est poursuivi à travers une série de conflits, pour s'évanouir enfin de lui-même avec… comment dirais-je… non pas un coup de tonnerre, mais un gémissement, en même temps que changeait le contexte économique et social. Puis surgissaient de nouveaux problèmes, et une nouvelle série de guerres… selon un cycle indéfiniment renouvelé[6].

Et quelle est la solution ? pour Susan Calvin, la robopsychologue de U.S Robots qui veille au bon fonctionnement des robots, le salut n'est pas dans les mains de l'homme, mais dans les Machines : « Quelle horreur ! Dites

plutôt quelle merveille ! Pensez que désormais et pour toujours les conflits sont devenus évitables. Dorénavant seules les Machines sont inévitables ![7] » Et les Machines peuvent rendre les conflits évitables parce qu'elles ont les capacités d'analyse que nous, pauvres humains, n'avons pas, comme le précise le coordonnateur :

> Si je comprends bien, Susan, vous me dites que la Société pour l'humanité a raison et que l'humanité a perdu le droit de dire son mot dans la détermination de son avenir. Ce droit, elle ne l'a jamais possédé, en réalité. Elle s'est trouvée à la merci des forces économiques et sociales auxquelles elle ne comprenait rien… des caprices des climats, des hasards de la guerre. Maintenant les Machines les comprennent ; et nul ne pourra les arrêter puisque les Machines agiront envers ces ennemis comme elles agissent envers la Société pour l'humanité… ayant à leur disposition la plus puissante de toutes les armes, le contrôle absolu de l'économie[8].

Dans ces conflits incessants, il est impossible pour une personne de la génération d'Asimov et de celle qui l'a suivie de ne pas faire référence au nucléaire. En effet, le 6 août 1945 fut lancée la première bombe nucléaire sur Hiroshima et, le 9 août suivant, la seconde sur Nagasaki[9]. La bombe nucléaire représente bien, pour une génération d'après-guerre, comment le développement technologique peut être menaçant. Peu de gens de cette génération ont oublié les sueurs chaudes qu'a données la guerre froide lors de la crise des missiles à Cuba[10]. Dans « Cailloux dans le ciel », Asimov met en scène les enjeux du nucléaire de la façon suivante. D'abord, il présente le personnage principal dont la vie sera transformée :

> Bien sûr, il y avait la bombe atomique et toutes les discussions quelque peu choquantes sur l'éventualité d'une troisième guerre mondiale, mais Joseph Schwartz croyait en la bonté

> intrinsèque de la nature humaine. Il n'imaginait pas qu'il pourrait y avoir un nouveau conflit. Il ne concevait pas que la Terre pourrait assister une seconde fois au déchaînement de la fureur de l'atome[11].

De l'autre côté, les scientifiques :

> Dans un autre quartier de Chicago était installé l'Institut de recherche nucléaire. Là aussi, les gens avaient peut-être une opinion quant à la valeur essentielle de la nature humaine, mais c'était à leurs yeux de théories dont ils ne se vantaient pas puisque l'on n'avait pas encore inventé l'instrument susceptible de mesurer quantitativement l'être humain. Et, quels que fussent leurs points de vue personnels, ils en étaient tous à espérer que la foudre du ciel empêcherait ladite nature (et la maudite ingéniosité) humaine à transformer la moindre découverte innocente et intéressante en une arme de mort[12].

Enfin, le dernier acteur de ce drame : l'imprévu, c'est-à-dire tout ce qui peut arriver à cause de notre ignorance, parce que nous n'avons pas, nous pauvres humains, l'omniscience de toutes les conséquences des gestes que nous faisons.

> Dans tout l'Institut, personne, ni à ce moment ni plus tard, ne se révéla capable d'expliquer pourquoi un creuset contenant un échantillon d'uranium brut d'une masse très inférieure à la masse critique, et qui, par surcroît, n'était pas soumis à un bombardement direct de neutrons, s'était liquéfié en émettant une luminosité aussi dangereuse que significative[13].

Asimov n'est donc pas naïf, il sait de quoi nous sommes capables comme humains et c'est pourquoi il projette cette conscience des « dangers », que représente le développement des robots, dans le refus des Terriens de les accepter sur la terre. Dans « Pour que tu t'y intéresses », le responsable de U.S. Robots affirme clairement la persistance du refus des robots :

> En deux siècles, je peux le dire, de succès considérables, la société U.S. Robots n'a toujours pas réussi à persuader les êtres humains d'accepter les robots. Nous n'avons confié à ceux-ci que des tâches nécessaires, mais impossibles à effectuer par des êtres humains, à cause des dangers présentés, entre autres, par l'environnement[14].

La question centrale demeure : Comment faire pour que les humains acceptent les robots ? Dans toute l'œuvre d'Asimov, l'existence de Trois Lois de la robotique est nécessaire pour assurer l'acceptation des robots. Bogert, responsable de l'U.S. Robots dans la nouvelle « Lenny », n'arrive pas à comprendre ce refus des Terriens : « J'aurais dû m'y faire depuis le temps, mais je n'y parviendrai jamais. On pourrait croire que, de nos jours, tout être humain résidant sur terre serait parfaitement conscient que les Trois Lois constituent une sécurité totale ; que les robots ne présentent aucun danger[15]. » Susan Calvin, de son côté, sera intransigeante devant toute modification des Trois Lois dans la fabrication des robots, comme on le voit dans *Le petit robot perdu* ou encore *Intuition féminine*. Pour Asimov, à travers son personnage de Susan Calvin, on voit bien que les Lois morales sont nécessaires pour éviter que les robots ne deviennent une autre « arme de mort » dans nos conflits humains.

2. COMMENT LES LOIS MORALES PEUVENT-ELLES NOUS APPORTER CETTE GARANTIE DE SÉCURITÉ ?

2.1 Le rôle de la morale dans une société

En élaborant des règles morales pour la robotique, Asimov avouait implicitement qu'une société ne peut vivre sans morale ; plus précisément, on devrait dire que l'on ne peut pas bien vivre en société si les personnes que nous côtoyons ne sont pas morales. Si cette affirmation peut paraître évidente pour certains, d'autres ne perçoivent qu'intuitivement cette évidence. Pour l'expliciter, nous pouvons partir

des scandales dans notre société. Qu'est-ce qui nous scandalise ? Est-il étonnant de voir comment le journalisme dit d'enquête cherche à mettre à jour des scandales de toutes sortes ? Partons de certains scandales qui font souvent la une des journaux. Est-ce que le comportement de parents qui agressent sexuellement leurs enfants mineurs ou qui les obligent à faire de la prostitution ou encore à voler pour le bien-être de la famille vous scandalise ? Est-ce que des professionnels (éducateurs, médecins, psychologues, etc.) qui ont la charge d'enfants et qui les agressent sexuellement vous choque ? Est-ce que le comportement de professionnels de la santé, qui doivent en principe prendre soin de vous, mais qui dans les faits ne vous adressent que peu la parole et qui sont très expéditifs, vous scandalise ? Que dire au niveau gouvernemental du scandale des commandites et du scandale de la construction ? Enfin, que pensez-vous de la réaction du peuple américain devant les agissements de Clinton ou de Nixon, deux présidents contre lesquels on a pris des mesures de destitution pour manquement à la morale ?

Le scandale est avant tout un signe qu'il se passe quelque chose qui nous choque, que l'on trouve inadmissible. Lorsqu'on regarde de plus près ce que l'on n'accepte pas dans ces scandales, on pourrait dire que nos attentes ont été déçues. En effet, nous attendons naturellement que ces personnes, compte tenu de leur rôle dans la société, agissent en fonction du bien d'autrui. On pourrait résumer ces attentes déçues ainsi : au lieu d'aider les autres comme cela va de soi dans leur fonction, certaines personnes n'ont pas fait ce qu'elles auraient dû faire et d'autres, pire encore, ont profité de leur situation pour « utiliser » les autres comme des jouets pour leur propre plaisir ou pour assurer leur pouvoir. Ce qui choque, c'est, comme les philosophes l'ont dit autrement, que des êtres humains soient considérés

comme des objets (des instruments) et qu'ils puissent être manipulés par d'autres, plutôt que d'être respectés comme des personnes libres.

Pour assurer la qualité du vivre-ensemble dans une société, il faut que les personnes soient morales, car alors nous savons que nous pouvons compter sur elles. Dans la nouvelle « La preuve », on soupçonne que le procureur Byerley, qui jouit d'une très bonne réputation et qui a posé sa candidature à la mairie, soit un robot humanoïde tellement perfectionné qu'il est difficile de faire la preuve qu'il est un robot. Comment faire ?

> Il y a deux méthodes pour établir une preuve, la méthode physique et la méthode psychologique. Physiquement, vous pouvez le disséquer ou faire appel aux rayons X. Comment y parvenir ? C'est vous que cela regarde. Psychologiquement, on peut étudier son comportement, car s'il est un robot positronique, il doit de conformer au Trois Lois de la robotique. Nul cerveau positronique ne peut être construit sans satisfaire à ces règles[16].

Or le problème que soulève Susan Calvin avec la preuve psychologique est qu'il est difficile de différencier un robot d'un humain du point de vue moral, parce que les Lois de la robotique sont calquées sur la morale humaine d'une part, et d'autre part parce que seuls les humains peuvent manquer à la morale. Donc, la conclusion s'impose : « En un mot, si Byerley se conforme à toutes les Lois de la robotique, il se peut qu'il soit un robot, mais il se peut également qu'il soit un très brave homme[17]. » Susan Calvin ne cache pas son amour pour les robots qu'elle juge préférables à l'homme, notamment en raison de leur caractère moral. Ainsi, précise-t-elle à Byerley lors de sa rencontre avec lui : « La robo-psychologue, si vous n'y voyez pas d'inconvénient. – Oh ! Les robots seraient-ils donc à ce point différents des hommes sur le plan mental ? – Un monde les sépare (un sourire glacial

effleura ses lèvres). Le caractère essentiel des robots est la droiture[18]. » Il n'est pas étonnant dès lors que Susan Calvin ne voie aucun inconvénient à ce qu'un robot occupe une fonction comme celle de procureur ou encore de maire.

> J'aime les robots, je les aime beaucoup plus que les êtres humains. Si l'on pouvait créer un robot capable de tenir de fonctions publiques, j'imagine qu'il remplirait idéalement les devoirs de sa charge. Selon les Lois de la robotique, il serait incapable de causer du préjudice aux humains, il serait incorruptible, inaccessible à la sottise, aux préjugés. Et lorsqu'il aurait fait son temps, il se retirerait, bien qu'immortel, car il ne pourrait pas blesser des humains en leur laissant savoir qu'ils avaient été dirigés par un robot. Ce serait l'idéal[19].

Évidemment, il y a un mais… Le problème réside dans la capacité du cerveau positronique. Comment un tel cerveau pourrait-il être l'équivalent d'un cerveau humain devant la complexité des enjeux ? La réponse est dans le travail d'équipe.

> – Sauf qu'un robot pourrait échouer dans sa tâche en raison de certaines inaptitudes inhérentes à son cerveau. Le cerveau positronique n'a jamais égalé le cerveau humain. – On lui adjoindrait des conseillers. Un cerveau humain lui-même est incapable de gouverner sans assistance[20].

À partir de ces exemples, nous pouvons aisément conclure que les Trois Lois de la robotique visent à faire des robots des personnes ayant une telle droiture morale, de telle sorte qu'elles ne puissent pas être utilisées comme des armes de mort ou de domination pour résoudre nos conflits sociaux. Au contraire, nous pourrons leur faire entièrement confiance dans nos rapports avec eux. N'est-ce pas là une autre façon de souligner l'importance des êtres moraux ? Comment peut-on faire confiance à une personne si l'on ne présume pas qu'elle est suffisamment morale pour ne pas

abuser de nous en profitant de nous plutôt que d'établir une relation égalitaire ?

2.2 Pourquoi Asimov choisit-il ces Trois Lois morales et cette façon de les énoncer ?

2.2.1 L'énoncé des Trois Lois

Selon le Manuel de robotique 58e édition (2058 après J.-C.), les Trois Lois de la robotique s'énoncent ainsi :

Première Loi

> Un robot ne peut nuire à un être humain ni laisser sans assistance un être humain en danger.

Deuxième Loi

> Un robot doit obéir aux ordres qui lui sont donnés par les êtres humains, sauf quand ces ordres sont incompatibles avec la Première Loi.

Troisième Loi

> Un robot doit protéger son existence tant que cette protection n'est pas incompatible avec la Première ou la Deuxième Loi[21].

Dans les nouvelles, l'énoncé de la Première Loi est celui qui varie le plus. Dans certaines nouvelles comme « Menteur ! », on retrouve exactement cette formulation dans la bouche de Bogert : « Vous connaissez certainement la Première Loi fondamentale de la robotique ? – Certainement, dit Bogert avec impatience, un robot ne peut attaquer un être humain – ni, restant passif, laisser cet être humain exposé au danger[22]. » Par contre, dans d'autres nouvelles, on parle de la Première Loi comme se référant au seul fait de ne pas nuire, comme dans « Lenny ». Cependant, le changement le plus significatif, sur lequel nous reviendrons plus tard, consiste à transformer l'interdiction de nuire en obligation d'aider. On retrouve cela dans « Le correcteur ».

> Vous connaissez les Lois de la robotique, je présume. – Naturellement, répondit Goodfellow. – Elles font partie intégrante des réseaux positroniques et sont obligatoirement respectées. La Première Loi, qui régit l'existence du robot, garantit la vie et le bien-être de tous les humains. Il prit un temps, se frotta la joue et ajouta : C'est là un point dont nous aimerions persuader la terre entière si c'était possible[23].

Du point de vue moral, il y a une différence fondamentale entre interdire de nuire activement en faisant mal à quelqu'un ou ne pas porter assistance à quelqu'un en danger (nuire passivement) et l'obligation d'assurer le bien-être de tous les humains.

2.2.2 Pourquoi le choix de ces Lois ?

Une explication, quelque peu cynique, serait que la morale de la robotique n'est que la transcription de la morale des maîtres pour les esclaves. Évidemment, les Blancs pour se protéger des Noirs imposèrent la Première Loi : les esclaves ne devaient pas nuire aux autres Blancs. Il faut dès le départ éviter toute révolte des esclaves. Dans la nouvelle « Le petit robot perdu », on retrouve cette trame d'une révolte non violente certes où le robot prend à la lettre les mots du scientifique qui, frustré, lui dit d'aller se perdre. Cette révolte, analogue à celle des enfants envers leurs parents, est quand même une révolte qui suscite la crainte :

> Toute vie normale, Peter, qu'elle soit consciente ou non, supporte mal la domination. Si cette domination est le fait d'un inférieur ou d'un inférieur présumé, le ressentiment devient plus intense. Physiquement et, dans une certaine mesure, mentalement, un robot – tout robot – est supérieur aux êtres humains. Qu'est-ce donc qui lui donne une âme d'esclave ? *Uniquement la Première Loi !* Sans elle, au premier ordre que vous donneriez à un robot, vous seriez un homme mort. Et vous parlez d'instabilité[24] ?

Si l'on suit cette logique de la morale des esclaves, la Deuxième Loi est évidente : l'esclave doit obéir à son maître. L'obéissance aux commandements du maître doit être inconditionnelle pour que l'esclavage soit efficace. Mais, aussi, il faut que l'esclave accepte ces commandements, dans l'imagination des maîtres, avec plaisir comme si obéir au maître était leur seule raison d'être. Asimov fait allusion à ce type d'interprétation dans la nouvelle « Cercle vicieux » :

> Oui, Maître ! Powell adressa à son compagnon un sourire sans joie. Vous avez entendu ? À l'époque, on pouvait penser que l'usage des robots serait interdit sur la terre. Les constructeurs combattaient cette tendance et ils introduisaient dans leurs fichues machines de bons complexes d'esclaves parfaitement stylés[25].

Évidemment, la Troisième Loi tombe sous le sens. Tout comme les robots de U.S. Robots, ils coûtent cher et l'on ne veut pas qu'ils se détruisent eux-mêmes. Donc interdiction de se mutiler ou de se tuer, sauf si cela est pour sauver le maître ou l'un des siens ou encore si le maître ordonne d'exécuter un travail qui peut le conduire à mourir.

L'ami d'Asimov, Lester Del Rey, a présenté, dans la revue *Galaxie*, la nouvelle « Une morale pour Sam », qui est une parodie des Trois Lois de la robotique dans laquelle il montre comment ces lois rendent le robot inopérant et dangereux. Il reprend aussi l'idée de la morale d'esclave en faisant dire à son personnage :

> Esclavage et racisme ! Lee avait craché ces mots. L'esclave noir ne doit pas frapper le maître blanc ; l'esclave noir doit obéir au maître blanc ; l'esclave noir doit prendre soin de lui-même, comme faisant partie des biens de son maître blanc. Vous appelez ça une *morale*[26] ?

La seconde explication des Lois de la robotique situe celles-ci dans le contexte de la morale humaine. Les

humains ont fait les robots en leur intégrant leurs valeurs. Nous avons déjà vu, dans « La preuve », qu'on ne pouvait distinguer un humain d'un robot par la dimension morale, mais que cela n'était possible que si le robot manquait à la morale : en nuisant à un humain. C'est dans cette nouvelle qu'Asimov fait l'analogie entre les Lois de la robotique et les lois humaines :

> Parce que, si vous prenez la peine d'y réfléchir cinq secondes, les Trois Lois constituent les principes essentiels d'une grande partie des systèmes moraux du monde. Évidemment, chaque être humain possède, en principe, l'instinct de conservation. C'est la Troisième Loi de la robotique. De même, chacun des *bons* êtres humains, possédant une conscience sociale et le sens de la responsabilité, doit obéir aux autorités établies, écouter son docteur, son patron, son gouvernement, son psychiatre, son semblable… même lorsque ceux-ci troublent son confort ou sa sécurité. C'est ce qui correspond à la Seconde Loi de la robotique. Chaque *bon* humain doit également aimer son prochain comme lui-même, risquer sa vie pour sauver celle d'un autre. Telle est la Première Loi de la robotique. En un mot, si Byerley se conforme à toutes les Lois de la robotique, il se peut qu'il soit un robot, mais il se peut également qu'il soit un très brave homme[27].

Regardons cette comparaison de plus près, afin de mieux comprendre l'idée qu'Asimov se fait de la morale. Personne ne contestera son interprétation de la Troisième Loi. L'instinct de conservation amène chaque humain, comme chaque être vivant, à vouloir se protéger de ce qui le menace le plus : la mort. Contrairement à d'autres théories morales, Asimov ne conçoit pas la morale comme une réponse à une volonté de puissance que les humains auraient pour dominer les autres ou encore une tendance à faire le mal, comme l'explique le christianisme à la suite du péché originel. Bien que l'instinct de conservation ne soit pas orienté vers la

puissance et la domination, il n'en demeure pas moins que l'humain doit juger : suis-je en danger et, dans ces situations, qu'est-ce que je dois faire ? Son instinct de conservation pourrait ainsi le conduire à faire du tort à autrui.

L'explication de la Deuxième Loi associe la personne morale à l'obéissance aux autorités et aux ordres que celles-ci nous dictent. Cette conception de la morale est la plus répandue, car elle traverse toutes les expériences que nous avons faites dans notre histoire pour assurer le vivre-ensemble. Rappelons-nous ce qu'Asimov nous a dit du conflit et plus précisément comment la rareté des choses, qui conditionne l'économie, engendre le cycle des guerres. Dans « Le conflit évitable », c'est l'échec de la morale humaine qui conduit nécessairement à la prise en charge des humains par les machines. Pour qu'une morale puisse amener des êtres humains à réduire leurs conflits afin d'assurer la coexistence de tous, il faut que la morale propose quelque chose qui permet aux humains de s'unir. *L'union fait la force* est une devise de plusieurs pays qui renvoie à cette nécessité de s'unir pour arriver à quelque chose, malgré toutes les forces qui pourraient nous diviser. Pas étonnant de trouver déjà dans le monde romain la maxime *Diviser pour régner*.

Les premières formes d'organisation sociale se font autour d'un chef, d'une personne capable d'unir les autres, et cela lui confère une autorité pour commander. Avec le temps, le principe n'a pas changé, seulement les formes qu'a prises l'autorité. Du chef du clan qui était reconnu pour sa force et sa capacité d'assurer la survie des siens, à l'autorité des grands conquérants, à celle des dictateurs, on est passé à d'autres formes d'autorité comme la monarchie et, après elle, celle qui est reconnue par des régimes électoraux, les démocraties. Toutes nos organisations politiques et sociales reposent sur la reconnaissance de l'autorité légitime capable d'énoncer des lois. Toutes nos vies sont régulées par des

autorités et par l'ensemble de règles qu'elles imposent pour assurer notre vivre-ensemble.

Dans l'histoire humaine, ce ne sont pas seulement les autorités politiques qui ont cherché à assurer le vivre-ensemble, mais aussi les autorités religieuses. Toute religion soumet ses fidèles à l'autorité d'un Dieu ou d'un prophète qui en présente les lois. Dans le catholicisme, nous avons Moïse et les tables de la loi ; chez les musulmans, il y a Mahomet. Les religions proposent ainsi de nous unir sous une même autorité : le seul et vrai Dieu. En proposant qu'il existe une autorité qui nous transcende tous et qui n'est pas sur cette terre, elles nous incitent à nous réunir non plus autour d'un humain comme les autres, mais autour d'un Dieu qui sait où nous conduire pour assurer notre bonheur.

Depuis l'Antiquité grecque, certains philosophes comme Socrate ont voulu rompre avec l'autorité religieuse et proposer une autre façon de gouverner les hommes en pensant l'être humain comme un être soumis à la nature. Par notre naissance, nous ferions déjà partie, comme être biologique, à un ensemble qui nous dépasse : la nature et ses lois. Violer les lois de la nature nous conduirait à notre perte. Cette idée se retrouve aujourd'hui dans plusieurs approches écologiques.

L'explication de la Deuxième Loi de la robotique d'Asimov correspond bien à la tradition morale telle que notre histoire en témoigne sur le plan politique, religieux et philosophique. Qu'en est-il de la première ? Pourquoi la Deuxième Loi est-elle insuffisante ? Pourquoi ne pas nous limiter aux Lois deux et trois ?

La loi de l'obéissance ne ferait pas problème, si la personne qui a l'autorité de donner les ordres était bienveillante et juste. Dans les religions, l'autorité suprême, c'est Dieu ; et, puisqu'il est bon par définition, il n'a pas besoin de

loi pour interdire de nuire aux humains. Mais, avec les hommes, c'est autre chose. L'histoire passée témoigne de toutes les atrocités dont nous sommes capables, et cela même si l'autorité légitime a été désignée par un système démocratique. La montée d'Hitler en Allemagne et sa persécution des Juifs, allant jusqu'à créer les camps de la mort, illustrent bien que même les systèmes que nous croyons efficaces pour nous protéger contre les tyrans ont des limites. Si les robots obéissaient aveuglément à leur maître sans la Première Loi, qu'arriverait-il ?

La Première Loi concerne, rappelons-nous, l'interdiction de nuire directement à un humain, en le frappant par exemple, ou encore en laissant un humain, dans une situation dangereuse pour lui, sans secours. Le libellé de la Première Loi en anglais est plus précis que les traductions. En effet, elle s'énonce : « A robot may not injure a human being or, through inaction, allow a human being to come to harm[28]. » Il existe deux traductions différentes de cette loi, celle à laquelle nous nous référons dans ce chapitre – Un robot ne peut nuire à un être humain ni laisser sans assistance un être humain en danger – et celle à laquelle on se réfère au chapitre précédent – Un robot ne peut porter atteinte à un être humain, ni, restant passif, laisser cet être humain exposé au danger[29]. Le problème avec ces traductions, surtout la première, c'est qu'elles peuvent laisser entendre qu'un robot doit intervenir dès qu'il y a le moindre danger. C'est ce que Lester Del Rey[30] met en scène, dans une parodie des Trois Lois, avec un robot qui arrache la cigarette de la bouche d'un humain parce que fumer nuit à la santé. Pour ne pas tomber dans le ridicule du robot chez Del Rey, il faut donc que les robots soient dotés d'une capacité d'apprécier la nature et l'intensité de l'impact possible de ce qui entoure l'humain. On comprend donc que, pour assurer l'acceptation des robots, la Première Loi est essentielle ; et

c'est pourquoi, dans plusieurs nouvelles, le test ultime, pour savoir si une personne est un humain ou un robot, consiste à lui demander de blesser un autre être humain. Dans « La preuve », le problème est que Byerley frappe une autre personne, mais on ne sait pas si la victime est un humain ou un robot. Or un robot peut frapper un robot.

L'énoncé de la Première Loi ressemble beaucoup au principe de l'éthique médicale : *Primum non nocere*[31] (le principe de non-malfaisance). Dès le début de la médecine, ce principe a été placé au cœur de la pratique pour la simple raison que, sans cette garantie, les malades ne pourraient pas faire confiance aux médecins, craignant ainsi pour leur propre vie. Autrement dit, la puissance de la médecine peut servir pour guérir ou pour tuer, alors il faut garantir par un serment que cette puissance sera bien utilisée. C'est pourquoi, en prononçant le serment d'Hippocrate, le médecin s'engage à ne pas nuire : « Je dirigerai le régime des malades à leur avantage, suivant mes forces et mon jugement, et je m'abstiendrai de tout mal et de toute injustice. Je ne remettrai à personne du poison, si on m'en demande, ni ne prendrai l'initiative d'une pareille suggestion ; semblablement, je ne remettrai à aucune femme un pessaire abortif[32]. » Asimov applique ici au développement technologique des robots ce qu'Hippocrate avait fait pour le développement des interventions médicales en le soumettant à une limite : ne pas nuire activement ou passivement à autrui.

Cependant, l'explication de la Première Loi que donne Susan Calvin va plus loin que la simple interdiction de ne pas nuire. Elle lui donne un sens qui rattache cette loi au commandement de Jésus-Christ : Aimez-vous les uns les autres. En effet, elle dit bien : « Chaque *bon* humain doit également aimer son prochain comme lui-même, risquer sa vie pour sauver celle d'un autre. Telle est la Première Loi de

la robotique[33] » (p. 413). Pourquoi Asimov doit-il puiser dans la tradition morale du christianisme pour clarifier la portée de la Première Loi de la robotique ? L'enjeu ici n'est pas le contenu de l'énoncé de la Première Loi, car le principe *primum non nocere* suffit. Asimov doit rendre compte non seulement du contenu de chacune des lois, mais de la dynamique entre elles.

Dans la nouvelle « Cercle vicieux », Powell et Donovan sont sur Mercure pour tester le nouveau robot : Speedy. Donovan avait donné l'ordre à Speedy d'aller chercher du sélénium. Or, après cinq heures, il n'était pas revenu. Speedy tournait en rond autour du sélénium. Que s'est-il donc passé ? Voici comment Powell explique à Donovan pourquoi Speedy tourne en rond. Premièrement, il commence à expliquer comment fonctionne le cerveau positronique comme une force qui pousse dans une direction et une contre-force dans une autre :

> Précisément à l'explication. Les conflits entre les diverses lois sont réglés par les différents potentiels positroniques existant dans le cerveau. Disons qu'un robot doit marcher vers le danger et le sait. Le potentiel automatique suscité par la Loi numéro Trois le contraint à revenir sur ses pas. Supposons que vous lui donnez l'ordre d'aller s'exposer au danger. Dans ce cas, la Loi Deux suscite un contre-potentiel plus élevé que le précédent et le robot exécute les ordres au péril de son existence. – Je sais cela. Et après ?[34]

La question est donc : pourquoi est-ce que la Loi Deux n'a pas agi avec suffisamment de force pour contre balancer la Loi Trois ? Il y a deux facteurs qui ont joué dans le cas de Speedy. D'abord, la façon dont l'ordre a été donné :

> Speedy, nous avons besoin d'un peu de sélénium. Tu pourras en trouver à tel et tel endroit. Va et ramènes-en. C'est tout. Que vouliez-vous que je dise de plus ? Vous n'avez pas donné

> aucun caractère d'urgence à votre ordre, n'est-ce pas ? Pourquoi l'aurai-je fait ? Il s'agissait d'une simple opération de routine. Powell soupira. Nous n'y pouvons rien à présent… mais nous sommes dans de jolis draps[35].

Pourquoi ce qui était de routine avec les autres robots ne fonctionne-t-il plus avec Speedy ?

> – Prenons le cas de Speedy. Speedy est l'un des derniers modèles extrêmement spécialisé, et aussi coûteux qu'un croiseur de bataille. C'est une machine qu'on ne doit pas détruire à la légère. Alors ? – Alors la Loi numéro Trois a été renforcée – le fait a été mentionné spécifiquement dans les notices concernant les modèles SPD – si bien que son allergie au danger est particulièrement élevée. Dans le même temps, lorsque vous l'avez envoyé à la recherche du sélénium, vous lui avez donné cet ordre sur un ton ordinaire, sans le souligner en aucune façon, si bien que le potentiel de la Loi Deux était plutôt faible. Ne vous formalisez pas. Je ne fais qu'exposer les faits. – Continuez je commence à comprendre. – Vous voyez comment tout cela fonctionne, n'est-ce pas ? Il existe un danger quelconque dont le centre se situe dans le filon de sélénium. Il s'accroît quand Speedy en approche et à une certaine distance le potentiel de la Loi Trois, qui est inhabituellement élevé au départ, équilibre exactement le potentiel de la Loi Deux qui, lui, est plutôt faible au départ[36].

Comment Powell et Donovan peuvent-ils maintenant sortir d'une telle situation ? Comment faire pour que Speedy finalement aille chercher du sélénium et vaincre la force de la Troisième Loi ? La solution consiste à utiliser la Première Loi : « Il y a toujours la Première Loi, mais c'est une solution désespérée. Selon la Première Loi, un robot ne peut laisser un humain en danger et rester passif. Les Lois Deux et Trois ne peuvent s'y opposer. C'est tout à fait impossible[37]. » Il faut donc expliquer à Speedy l'état de la situation et il doit comprendre que « seul Speedy était capable de leur ramener

le sélénium. Pas de sélénium, pas de blanc de cellules photoélectriques. Pas de bancs de cellules… la cuisson lente était l'une des façons les plus déplaisantes de passer de vie à trépas[38]. » Pour que la force de la Première Loi conduise Speedy non seulement à ne pas faire du tort à Powell et Donovan, mais à risquer sa propre conservation pour eux, il faut une force telle qui exige le sacrifice de soi pour autrui.

C'est pour rendre compte de cette exigence morale de se sacrifier pour autrui qu'Asimov va chercher dans le christianisme l'exemple de Jésus-Christ. Dans le christianisme, Jésus-Christ, fils de Dieu, a été envoyé sur la terre pour sauver les humains en les rachetant de leurs péchés. Le sacrifice sur la croix représente donc le plus grand témoignage d'amour de Dieu pour les humains. Jésus-Christ commande ainsi aux humains de suivre son exemple. Dans l'évangile selon saint Jean (9 versets 12 et 13) : « 12 C'est ici mon commandement, Aimez-vous les uns les autres, comme je vous ai aimés. 13 Il n'y a pas de plus grand amour que de donner sa vie pour ses amis[39]. »

Asimov va donc puiser dans la tradition morale pour justifier l'appel aux Trois Lois de la robotique. Si l'on se fie à l'expérience humaine, ces lois ne sont pas toujours efficaces. Peut-on comprendre ce qui en assure l'efficacité ou l'inefficacité et savoir si les robots peuvent être plus moraux que les humains ?

2.3 Les Lois morales sont-elles efficaces ?

Les Lois de la robotique, tout comme les morales humaines, devraient être efficaces en principe, parce que ce sont des lois. Nul n'est censé désobéir à la loi. Mais pourquoi ? Comme nous l'avons déjà mentionné, les lois prennent leur force du fait qu'elles sont le résultat d'un commandement. En effet, pour vivre ensemble dans n'importe quel groupe ou n'importe quelle société, il faut que les personnes suivent

des règles. Dans nos sociétés démocratiques, nous valorisons le système électoral qui permet aux élus de faire des lois pour régir la société. Dans n'importe quelle organisation, il y a un président et dans des groupes de travail, des chefs d'équipe. C'est autour des chefs que nous nous unissons, car nous croyons qu'ils seront capables de nous unir et de nous permettre d'atteindre, ensemble, nos buts.

Comment un chef peut-il unir des personnes et atteindre des buts collectifs ? C'est par la force du commandement. Avez-vous remarqué comment les humains sont régis par le langage qu'ils utilisent ? Le philosophe John Langshaw Austin[40] a montré que le langage ne sert pas qu'à décrire la réalité, mais qu'il était un outil social fondamental pour faire des choses. Nous faisons plusieurs actions par le langage ou plus spécifiquement par la parole et celles-ci agissent dans nos façons de nous comporter les uns avec les autres. Deux de ces actes de parole qui sont fréquemment utilisés tous les jours sont le commandement et la promesse. Ce que J.L. Austin cherchait à découvrir, c'est comment ils fonctionnent.

Les actes de parole sont des conventions qui structurent la communication avec les autres et qui agissent sur eux parce qu'ils partagent la même convention. Prenons le cas du commandement. Que faut-il pour qu'un commandement réussisse, c'est-à-dire pour qu'il soit efficace et que la personne qui l'a reçu agisse en conséquence ? Pour réussir, un acte de parole doit suivre un ensemble de règles.

La première chose qu'il faut, c'est que le commandement soit clairement énoncé. Nous avons déjà vu le problème que Powell et Donovan avaient avec Speedy, parce que l'ordre n'avait pas été donné avec suffisamment de force. En effet, si une personne en autorité vous dit : « J'aimerais bien que tu fasses cela pour moi », allez-vous interpréter cela comme un ordre (énoncé de façon douce, mais un ordre quand

même) ou bien tout simplement comme l'expression d'un désir ou d'un souhait ? Selon votre interprétation, vous allez agir différemment et cela aura aussi des conséquences, car si c'est un ordre et que vous le prenez tout simplement comme un souhait, vous pourriez avoir des problèmes avec la personne en autorité.

Concernant la seconde règle, il faut que la personne qui donne le commandement soit reconnue par celle qui le reçoit comme ayant la légitimité de donner cet ordre. Déjà chez les Grecs, Sophocle, dans sa pièce *Antigone*[41], mettait en scène la question de la légitimité d'un ordre. Dans la pièce, le roi Créon fait un décret interdisant que Polynice reçoive les rites funéraires. Polynice est le frère d'Antigone qui a tenté de devenir roi en tuant son frère lors d'une bataille. Il ne faut donc pas oublier que, sans ces rites funéraires, l'âme du défunt est à jamais perdue. Antigone par deux fois va mettre sa vie en péril pour assurer la vie éternelle à son frère. Pourquoi ose-t-elle défier l'ordre du roi Créon et s'exposer ainsi à la mort ? Pour justifier son action, Antigone va distinguer la loi humaine, celle de Créon, et la loi divine, non écrite et éternelle. Pour Antigone, la loi humaine ne peut pas être légitime si elle ne respecte pas cette loi éternelle.

Mythe ou réalité ? Plus près de nous, après la Deuxième Guerre mondiale, il y a eu un procès à Nuremberg[42] pour juger des atrocités faites aux Juifs pendant la guerre. Évidemment, plusieurs diront qu'il s'agissait d'une vengeance plutôt que d'une justice internationale, mais ces personnes oublient que des principes moraux ou éthiques proviennent de ce procès. Parmi ceux-ci, retenons qu'un officier qui exécute des commandements émis par une autorité légitime de son pays, mais qui, en agissant ainsi, commet un crime contre l'humanité, est passible de la peine de mort. On retrouve ici la même idée que dans la pièce d'Antigone : tout commandement, même donné par une

autorité légitime, peut être du point de vue moral un crime contre l'humanité. C'est ce qui est au cœur des droits de l'homme. Le préambule de la Déclaration universelle signée en 1948 est très clair :

> *Considérant* que la reconnaissance de la dignité inhérente à tous les membres de la famille humaine et de leurs droits égaux et inaliénables constitue le fondement de la liberté, de la justice et de la paix dans le monde.
>
> *Considérant* que la méconnaissance et le mépris des droits de l'homme ont conduit à des actes de barbarie qui révoltent la conscience de l'humanité et que l'avènement d'un monde où les êtres humains seront libres de parler et de croire, libérés de la terreur et de la misère, a été proclamé comme la plus haute aspiration de l'homme[43].

La troisième règle du commandement précise que c'est en reconnaissant l'autorité légitime qui énonce le commandement que la personne qui le reçoit devient « obligée de faire ce qui est commandé ». Être obligé de faire quelque chose, c'est plus que de se sentir obligé, c'est reconnaître que je n'ai pas d'autre choix que d'exécuter ce qui est commandé. Évidemment, pour les générations qui ont été éduquées avec les droits de la personne, l'idée que l'on se considère comme n'ayant pas le choix de désobéir peut paraître impensable. Habituellement, on va plutôt penser ceci : je ne veux pas le faire parce que je vais subir des sanctions si je désobéis. Dans la tradition morale religieuse ou laïque, obéir à l'autorité légitime ne se discute pas plus que d'obéir aux lois physiques de la gravité. Ces lois font partie de notre nature et les défier, c'est se mettre en péril.

La promesse est un autre acte de parole par lequel une personne s'oblige face aux autres. Une promesse, comme le commandement, a aussi ses propres règles. La première est que la promesse doit être articulée clairement et

précisément. Tout comme le commandement, si on ne dit pas clairement qu'on promet, alors cela peut être interprété comme une déclaration d'intention. Il y a toute une différence entre « je pense que j'irai » et « je te promets d'y aller ». C'est sur la foi des promesses que l'on se fie aux autres. La deuxième règle est que la promesse doit être faite librement, sans contraintes et en connaissance de cause. Promettre quelque chose alors que l'on est menacé, ce n'est pas promettre. La force de la promesse dépend de la volonté que j'ai de m'engager. Pour que je sois lié par ma promesse, il faut donc que celle-ci soit libre, sans contraintes et faite en connaissance de cause. La troisième règle est que promettre crée pour moi une obligation de faire ce qui est promis. Tout comme le commandement, à la suite de ma promesse, je suis obligé de faire ce qui est promis, même si cela me coûte quelque chose. Le cas suivant peut sembler exceptionnel, mais il témoigne de la force de la promesse chez certaines personnes. Négociant avec le fils d'un antiquaire, je m'entends sur un prix pour un objet assez unique. Le fils de l'antiquaire va emballer l'objet lorsque le propriétaire vient me voir. Il me dit : « Savez-vous que cet objet vaut quatre fois le prix que vous allez payer ? Mon fils s'est trompé, mais il a donné sa parole et je la respecterai. Je voulais que vous sachiez au moins la vraie valeur de l'objet. »

Pour assurer l'efficacité des commandements, nous avons parfois recours dans nos institutions au serment. Tout comme le serment d'Hippocrate, nos serments sont des promesses, mais celles-ci expriment souvent notre engagement à suivre des commandements légitimement édictés. Par exemple, voici le serment de citoyenneté que tous les immigrants doivent prononcer pour devenir citoyen canadien : « Je jure fidélité et sincère allégeance à Sa Majesté la reine Élisabeth II, reine du Canada, à ses héritiers et successeurs et je jure d'observer fidèlement les lois du

Canada et de remplir loyalement mes obligations de citoyen canadien. » On retrouve la même idée dans les promesses entre époux dans le rite chrétien. Par exemple, dans une liturgie protestante, on présente ainsi la promesse échangée entre époux et épouse :

> Vous N.N., vous déclarez devant Dieu et devant son Église que vous avez pris pour femme N.N., ici présente. Vous promettez en même temps de l'aimer, de la protéger, de lui demeurer attaché dans la santé et dans la maladie, dans la prospérité et dans la détresse, et de lui rester fidèle, comme c'est le devoir d'un bon mari envers sa femme, et comme Dieu vous le commande dans sa Parole. – Est-ce bien là ce que vous déclarez et promettez ?[44]

Comme on peut le constater avec le commandement et la promesse, la vie morale repose sur la capacité d'un être humain de s'engager face aux autres et d'agir par la suite, conformément à cet engagement, même au détriment de son intérêt personnel. Toute morale repose ainsi sur notre capacité de nous soucier des autres au point de renoncer à des bénéfices pour soi.

À la lumière de cela, on peut mieux comprendre pourquoi les robots sont, selon Susan Calvin, plus droits que les humains. En fait, comme elle le dit, c'est son rôle de s'assurer que les Trois Lois de la robotique soient bien programmées dans les robots. Autrement dit, l'avantage que possède le robot sur l'humain, à première vue, c'est que les Trois Lois sont programmées alors que, pour nous, elles sont le fruit de notre éducation. Les erreurs de programmation sont peut-être moins fréquentes que les erreurs d'éducation.

Pour être efficaces, les lois de la robotique comme les lois morales ou les lois juridiques doivent être appliquées dans des situations concrètes. L'efficacité ultime des lois repose

dans le jugement éthique qui, aux yeux d'Asimov, constitue la plus haute fonction de la robotique.

3. COMMENT LES ROBOTS APPLIQUENT-ILS LA MORALE ? ASIMOV ET LE RAISONNEMENT PRATIQUE DES ROBOTS

3.1 L'importance du jugement éthique ou du raisonnement moral pour Asimov

Dans trois nouvelles, Asimov précise l'importance du jugement éthique et surtout sa complexité en faisant de cette programmation la plus haute tâche de la robotique. En parlant de Dave (« Attrapez-moi ce lapin ») qui avait de la difficulté à donner des ordres à six robots qui étaient comme des parties de lui, Powell précise les étapes parcourues pour tester les robots. Ainsi, avant de sortir du laboratoire pour des tests sur les terrains dont s'occupent nos deux techniciens, ils doivent passer une étape finale : « Et finalement soumit son esprit mécanique précis aux plus hautes fonctions du monde robotique : la solution de problèmes de jugement et de l'éthique[45]. » Puisque la sécurité qu'apportent les Trois Lois aux humains qui côtoient les robots repose essentiellement sur la façon dont elles seront appliquées par les robots, on peut comprendre l'importance de l'étape finale : « À partir de ce moment, on passait à des sujets plus compliqués, destinés à mettre à l'épreuve les différentes Lois et leur interaction avec les connaissances spécialisées de chaque modèle particulier[46]. » Évidemment, lorsqu'on arrive à créer un androïde comme Byerley dans « La preuve », il faut donc avoir atteint le summum de la robotique :

> Il parvint, on ne sait trop comment, à se procurer un cerveau positronique, du type le plus complexe, ce qui est bien mieux, un organe possédant les plus grandes capacités pour former des jugements d'éthique… ce qui est la fonction la plus haute que la robotique ait pu réaliser à ce jour[47].

Mais pourquoi le jugement éthique paraît-il si complexe pour Asimov au point de consacrer une grande partie de son œuvre sur les robots à montrer les failles de jugement que certains d'entre eux pouvaient faire dans certaines circonstances ?

3.2 Qu'est-ce que le jugement éthique ou le raisonnement pratique ?

Dans le champ de la morale, beaucoup d'attention est porté aux Lois morales. Comme nous avons pu le montrer jusqu'à présent, l'énonciation des Lois morales est fondamentale. Pour qu'une morale soit efficace, il faut en connaître les obligations. À quoi sommes-nous obligés ? Qu'il s'agisse de commandements de Dieu, de ceux de l'État ou de ceux de n'importe quelle organisation, l'énoncé des lois est capital, car personne n'est censé ignorer la loi qui le gouverne. Hammourabi, roi de Babylone, a fait graver les lois sur des stèles en basalte et il a placé ces stèles dans différentes parties de son royaume afin que tous aient accès au même contenu des lois[48]. Les images de la stèle, qui est aujourd'hui au Louvre, sont impressionnantes, car ce sont les plus anciennes lois écrites que nous connaissons et elles datent d'environ 1 730 ans avant Jésus-Christ.

L'image d'Hammourabi recevant du dieu Shamash le sceptre de justice symbolise clairement la seconde composante de tout énoncé moral : son autorité. La justice rendue par Hammourabi est fondée sur la justice divine. Ces commandements sont donc légitimés. Le code propose ainsi les lois à suivre et leur fondement.

Mais, ce qui est particulier dans ce code, c'est que les lois ne sont pas comme celles que nous connaissons dans la tradition française, des lois générales énoncées par le législateur et regroupées dans des codes comme le Code civil du Québec, mais des sentences du roi pour régler des litiges

entre les personnes dans la société. Cette forme des lois qui sont énoncées à partir de la pratique et du raisonnement des juges pour résoudre un conflit entre deux personnes se nomme le droit commun. La tradition britannique de la Common Law a reproduit pendant longtemps cette pratique. Autrement dit, la loi émerge dans la solution d'un cas difficile et c'est le raisonnement pratique qui permet de préciser le contenu exact de la loi.

Peu importe la tradition, le point de convergence est dans le raisonnement pratique. En effet, le philosophe Aristote (384-322 av. J.-C.) a été l'un des premiers à réfléchir sur le raisonnement pratique en le distinguant du raisonnement théorique (logique ou mathématique). Avec la logique ou la mathématique, on peut passer de la connaissance à la pratique sans problèmes : puisque 2 + 2 = 4, alors 2 pommes + 2 pommes vont donner 4 pommes, et ce sera la même chose avec tout autre objet. La situation concrète importe peu, sauf, évidemment, pour les personnes concernées. Mais qu'arrive-t-il lorsque nous sommes devant une loi générale, comme les Trois Lois de la robotique ? Comment fait-on pour passer de l'énoncé général au cas particulier ? Pour Aristote, dans le domaine des choses humaines si les lois servent de guides pour l'application, c'est dans le raisonnement pratique que se situe le véritable travail éthique ou moral : déterminer le choix d'action. La conclusion d'un raisonnement pratique pour Aristote est la détermination de l'action que nous allons faire conformément à la loi générale. Aristote nommait les juges qui appliquaient les lois générales des juges en équité, c'est-à-dire des juges capables de rendre la justice en tenant compte de tous les faits en présence.

Qu'est-ce qui rend le raisonnement pratique si complexe ? Prenons la Première Loi de la robotique : « Un robot ne peut nuire à un être humain ni laisser sans assistance un être

humain en danger. » Pour appliquer cette loi générale dans un contexte particulier le robot doit associer à l'action qu'il va faire les conséquences négatives directes pour l'humain et les conséquences négatives indirectes pour l'humain. Cependant, cette première opération dépend des connaissances que le robot a de la situation de son action et des impacts négatifs que celle-ci peut engendrer. Puisque chaque situation est unique, le raisonnement pratique doit tenir compte de toutes les variables pour s'assurer d'un jugement équitable. Pour déterminer si l'action nuit à un être humain, il faut que le robot ait une définition de l'être humain. Qu'est-ce qu'un être humain ? Comment le reconnaît-on ? Y a-t-il des humains moins humains que d'autres ? N'avons-nous pas déjà considéré que les Noirs ne sont pas des humains comme les autres, ce qui permettait d'en faire des esclaves ? Une fois les conséquences nuisibles sur l'humain connues, le robot doit ensuite se poser la question de l'étendue de l'interdiction. Il doit se demander si la Première Loi interdit toute action nuisible, peu importe sa gravité pour la personne. Si tel est le cas, alors il y a de fortes chances que le robot soit efficace uniquement en l'absence des humains. Dans la sphère humaine, les lois s'appliquent en tenant compte d'un seuil de tolérance. Par exemple, on ne doit pas dépasser la vitesse permise sur les routes, mais on sait qu'il y a une zone de tolérance et que l'on peut faire du 110 km/h dans une zone de 100 km/h sans problèmes. La zone de tolérance est mesurée en tenant compte de la gravité de l'impact négatif. Autrement dit, il y a toujours une évaluation des conséquences négatives qui doit être faite (suffisamment grave pour l'humain) pour appliquer l'interdiction générale.

Le raisonnement pratique de la Première Loi passe donc par trois étapes : définir les impacts négatifs de l'action sur l'humain, évaluer la gravité de cet impact sur l'humain,

décider en tenant compte de l'évaluation de l'impact négatif si la loi s'applique dans ce cas.

La situation devient plus complexe lorsqu'il y a plus d'un humain concerné par l'action. Le robot devrait en principe faire la même analyse pour tous les humains qui pourraient subir des conséquences négatives de son action. La situation devient encore plus complexe lorsque la même action entraîne des conséquences positives et négatives sur des êtres humains. Quoi que fasse le robot, son action nuira à quelqu'un. Comment applique-t-il alors l'interdiction de ne pas nuire ? Dans une situation comme celle-ci, le robot ne doit pas seulement reconnaître les impacts négatifs sur les humains et évaluer la gravité de ses impacts, mais il doit aussi pondérer, c'est-à-dire accorder plus de poids aux impacts sur certaines personnes que sur d'autres.

C'est cette complexité du raisonnement pratique qu'Asimov va soumettre à ses lectrices et lecteurs en illustrant dans ses nouvelles comment le raisonnement moral est difficile en pratique. Analysons comment Asimov pose les difficultés du raisonnement pratique à la lumière des composantes que nous avons présentées. Évidemment, les situations varient selon la capacité des robots et la particularité de la situation.

3.3 Les erreurs de raisonnement moral pratique chez les robots

3.3.1 Violation de la Loi Un

Asimov donne deux exemples dans lesquels les robots violent directement la Première Loi, un en faisant du mal à un humain et l'autre en laissant un humain en danger sans lui porter secours. Compte tenu de l'importance de la Première Loi, il faut donc des conditions très particulières pour pouvoir mettre en scène de telles violations. Dans la nouvelle « Lenny », ce robot est en fait un robot enfant.

Imaginez la surprise de Susan Calvin : « Cela donnait à peu près ceci : " Da, da, da, gou. " Le robot était toujours debout, grand et parfaitement droit, mais sa main droite se leva lentement, et il introduisit un doigt dans sa bouche[49]. » La création de Lenny est un accident. Puisque des robots construisent les robots positroniques, un jeune homme, lors d'une visite au laboratoire, a tout simplement pianoté sur le clavier de l'ordinateur qui, malheureusement, n'était pas fermé. Malgré toutes les précautions, un accident est toujours possible. C'est donc ce qui donna naissance à Lenny.

L'incident en cause est simple : Lenny a brisé le bras d'un technicien. Donc, il a manqué à la Première Loi et, plus que cela, ce manquement ne l'a pas détruit, contrairement à toutes les garanties données. Pourquoi ? Qu'est-ce qui explique un tel manquement ? Dans son enquête, Susan Calvin interroge le technicien :

> J'ai essayé de l'effrayer pour l'amener à dire quelque chose. Il ajouta comme pour se justifier : Il fallait bien le secouer un peu. – De quelle façon avez-vous tenté de l'effrayer ? – J'ai fait mine de lui décrocher un coup de poing. – Et il a repoussé votre bras ? – Il a frappé mon bras. – Très bien. C'est tout ce que je voulais savoir[50].

Que conclut Susan Calvin de tout cela ? Premièrement, que le robot a agi en fonction de la Troisième Loi : se défendre. Mais s'il l'a fait en faisant mal à un humain, cela ne peut s'expliquer que par l'ignorance des conséquences de son geste sur l'humain. L'ignorance des conséquences de nos gestes est un facteur important dans une décision, car nous pouvons faire mal aux autres par ignorance. Si nous avions su, nous aurions peut-être décidé autrement. Cet exemple tout simple montre les premières limites inhérentes au raisonnement moral que nous faisons : le degré de connaissance des conséquences de nos actions sur les autres et sur l'environnement. Voici la conclusion que tire Susan Calvin :

> Lenny n'avait pas le droit de se défendre au prix d'un dommage, fut-il mineur occasionné à un être humain. – Il ne l'a pas fait *sciemment* riposta le Dr Calvin. Le cerveau de Lenny est déficient. Il ne pouvait pas connaître sa propre force ni la faiblesse humaine. En écartant le bras menaçant d'un être humain, il ne pouvait pas prévoir que l'os allait se rompre. Humainement parlant, on ne peut incriminer un individu qui ne peut honnêtement distinguer le bien du mal[51].

Le second exemple met en scène Emma Deux dans la nouvelle « Première Loi ». Dans une présentation de cette nouvelle, Asimov déclare : « Je dois également vous prévenir au sujet de la première histoire, " Première Loi " : elle fut écrite en manière de plaisanterie et n'a pas été conçue pour être prise au sérieux[52]. » Dans cette histoire racontée par Powell, Emma Un est un robot qui a été égaré depuis un certain temps. Lors d'une sortie hors de la base, Powell doit affronter une tempête et un chien des tempêtes qui le menace. Emma est témoin de la scène et ne fait rien pour secourir Powell. Que s'est-il donc passé ?

> Ce chien des tempêtes n'était pas un chien des tempêtes. Nous la baptisâmes Emma Junior lorsque Emma Deux le ramena à la base. Emma Deux se devait de le protéger contre mon pistolet. Que sont les injonctions de la Première Loi, comparées aux liens sacrés de l'amour maternel[53] ?

Malgré le côté extrême de cette nouvelle, elle nous est significative pour comprendre le raisonnement moral, surtout à l'étape de la pondération. Emma Deux avait à pondérer entre deux maux : celui fait à Powell ou celui fait à Emma Junior. Comme dans tout choix tragique, il faut accorder plus de poids à une partie de l'alternative parce qu'on accorde plus de valeur à l'une qu'à l'autre. Ici Emma Deux a choisi Emma Junior et, pour Powell, la raison est simple : la force de l'amour maternel.

3.3.2 Les robots menteurs

Pourquoi le mensonge est-il inacceptable ? Qu'est-ce qui fait du mensonge quelque chose que nous condamnons moralement ? Y a-t-il des situations où le mensonge est justifié pour le bien des autres ? Dans un texte vulgarisant le débat philosophique autour du mensonge, Éric Fiat résume bien ce qui est au cœur du mensonge : la confiance dans la parole de l'autre.

> D'abord parce que le mensonge est contradiction entre la parole et la pensée, et qu'il ruine l'essence même de la parole qui est la confiance. Tout acte de parole promet la vérité, même – et surtout ! – l'acte de parole qui ment et qui peut aller jusqu'à jurer qu'il dit vrai, alors qu'il ment. La société des hommes deviendrait vite infernale si chacun devait se méfier de chacun. Je fais spontanément confiance au quidam auquel je demande mon chemin, perdu que je suis dans les rues de Metz ou de Bordeaux… Pourquoi ? Parce que tout se passe comme si me liait à lui une sorte de contrat de confiance, contrat antérieur à tous ceux que je pourrais un jour signer avec lui, et qui en est la condition de possibilité[54].

Asimov met en scène quatre robots menteurs que doivent affronter Susan Calvin et Elijah Baley. Ils devront comprendre comment et dans quel contexte la dynamique des Trois Lois de la robotique permet le mensonge. Chaque fois, la mise en scène d'Asimov permet de mieux comprendre la dynamique de la morale chez le robot, mais surtout chez l'humain.

Dans la nouvelle « Menteur ! », Susan Calvin fait face à un robot télépathe. Ici encore, une erreur dans la création du cerveau positronique complexe semble expliquer le phénomène. En quoi un robot, capable de lire dans les pensées, mentirait-il ? Dans la nouvelle, le robot ment à toutes les personnes concernées en leur disant ce qu'elles désirent entendre, au lieu de la vérité. Herbie, sachant lire

dans les pensées, a plus d'informations qu'un autre robot sur les conséquences de son action. Puisqu'il sait ce que l'autre attend, il sait aussi que, s'il dit la vérité, il peut décevoir l'autre, ne pas répondre à son attente, le dévaloriser, etc. Nous sommes ici dans une situation aux antipodes de celle de Lenny qui ignorait les conséquences de son acte. Herbie voit des conséquences que d'autres robots ignorent et il tente donc d'appliquer strictement la Première Loi à ces conséquences. Bien qu'ils ne soient pas télépathes, les humains peuvent très bien savoir que leurs actions sont susceptibles de causer des blessures morales à d'autres, et cela rend leur raisonnement pratique plus complexe. Dans le cas de Herbie, la solution est simple : mentir pour ne pas nuire à l'autre. Susan Calvin l'explique en ces termes :

> Vous connaissez certainement la Première Loi fondamentale de la robotique ? – Certainement, dit Bogert avec impatience, un robot ne peut attaquer un être humain – ni, restant passif, laisser cet être humain exposé au danger. Merveilleusement exprimé, ironisa Calvin. Mais quel genre de danger ? Quel genre d'attaque ? – Mais tous les genres. – Exactement ! Tous les genres ! Mais, pour ce qui est de blesser les sentiments, d'amoindrir l'idée qu'on se fait de sa propre personne, de réduire en poudre les plus chers espoirs, sont-ce là des choses sans importance ou au contraire ? [...] Le robot lit dans les pensées. Pensez-vous qu'il ignore tout des blessures morales ? Pensez-vous que, si je lui posais une question, il ne me donnerait pas exactement la réponse que je désire entendre ? Toute autre réponse ne nous blesserait-elle pas, et Herbie peut-il l'ignorer[55] ?

Herbie applique strictement la Première Loi à chaque personne à qui il s'adresse et il ne tient pas compte de l'ensemble de la situation. Puisqu'il est le seul à savoir ce qui a pu provoquer la télépathie dans le montage, il refuse de le dire, car les D^rs^ Bogert et Lanning seront blessés moralement

puisqu'un robot en sait plus qu'eux. Comme on dit : « Leur ego va en prendre un coup. » C'est parce que Herbie n'est pas capable d'évaluer, d'une part, la différence entre la blessure de l'ego engendrée s'il dit la vérité et, d'autre part, les conséquences sur eux de ne pas savoir ce qui ne fonctionne pas avec lui. C'est pour cette raison qu'il ment en disant ne pas savoir ce qui ne fonctionne pas chez lui. Ne pouvant pas faire l'évaluation entre les deux maux que sa parole ou son silence engendre, il ne peut pas pondérer les conséquences pour agir en faisant le moindre mal. Les humains résolvent souvent leurs dilemmes moraux en choisissant de faire le moindre mal possible, car leur action a des conséquences négatives sur plusieurs personnes concernées. Or Herbie ne sait pas faire cela, il pense en mode binaire (faire mal ou ne pas faire mal) et non en mode qualitatif (plus ou moins de mal). C'est ce qui permet à Susan Calvin de le soumettre à un dilemme insoluble qui le rend fou, puisque son action sera toujours une violation de la Première Loi :

> Vous ne pouvez rien leur dire, récitait la psychologue lentement parce que cela leur causerait de la peine, ce qui vous est interdit. Mais, si vous refusez de parler, vous leur causez de la peine, ce qui vous est interdit. Mais, si vous refusez de parler, vous leur causez de la peine, donc vous devez tout dire. Si vous le faites, vous leur ferez de la peine, ce qui vous est interdit, par conséquent vous vous abstiendrez. Mais, si vous vous abstenez, ils en concevront du dépit et par conséquent vous devez leur donner la réponse…[56]

Il y a une seconde source de mensonges chez les robots, soit lorsque les humains font une combinaison des deux premières Lois pour inciter un robot à mentir. Cette dynamique des deux premières Lois est développée dans les trois autres nouvelles touchant le mensonge. Chacune apporte plus de précisions sur l'univers moral des humains, puisque les humains incitent les robots à mentir pour eux.

Dans « Le correcteur », le robot EZ-27 surnommé Easy est le premier robot que U.S. Robots met au service du public, tout en respectant les lois sur la terre.

> Si le robot est utilisé exclusivement dans une salle déterminée, à des fins académiques, si certaines autres restrictions sont scrupuleusement observées, si les hommes et les femmes qui sont amenés, de par leurs fonctions, à pénétrer dans cette salle nous assurent une entière collaboration, nous pouvons demeurer dans les limites de la loi[57].

L'U.S. Robots aimerait bien faire modifier les lois terriennes et favoriser l'utilisation des robots ailleurs que dans les colonies. Cependant, les robots sont surtout, jusqu'à présent, utilisés dans les colonies et ont peu d'interactions avec les humains. Sur terre, ils devront répondre à plusieurs humains et interpréter des situations complexes du vivre-ensemble. Easy est un robot qui peut exécuter plusieurs tâches de bureau, mais, d'abord et avant tout, il s'agit d'un outil linguistique perfectionné, car c'est un correcteur habile qui peut corriger toutes sortes de textes : « Lorsqu'un professeur capable d'un travail puissamment créateur est assujetti, deux semaines durant, au travail mécanique et abrutissant qui consiste à corriger des épreuves, me traiterez-vous de plaisantin si je vous offre une machine capable de le faire en trente minutes ?[58] »

Simon Ninheimer, professeur de sociologie, poursuit l'U.S. Robots parce qu'Easy aurait non seulement corrigé son texte, mais aurait aussi modifié celui-ci de manière telle que sa crédibilité scientifique et sa réputation du professeur sont mises en cause. S'ensuit un procès à huis clos devant un juge seul voulant faire la lumière sur la situation. Selon l'interprétation du professeur Ninheimer, Easy aurait modifié son texte conformément à la Première Loi :

> Est-ce vous qui avez modifié le texte pour le faire imprimer sous sa forme actuelle ? – Oui, professeur. Pourquoi ? – Professeur, les passages tels qu'ils apparaissent dans votre version étaient fort offensants pour certains groupes d'êtres humains. J'ai pensé qu'il était plus judicieux de modifier la formulation afin d'éviter de leur causer du tort. – Comment avez-vous osé prendre une telle initiative ? – La Première Loi, professeur, ne m'autorise pas à causer du tort, même passivement, à des êtres humains. À n'en pas douter, vu votre réputation dans les cercles de la sociologie et la large diffusion de votre livre parmi le monde des érudits, un mal considérable serait infligé à un certain nombre d'êtres humains dont vous parlez. – Mais vous rendez-vous compte que c'est moi qui vais en pâtir à présent ? – Je n'ai pu faire autrement que de choisir la solution comportant le moindre mal[59].

Comme l'affirme Susan Calvin, ce type de raisonnement pratique est hors de la portée d'Easy. L'opération en cause est trop abstraite pour lui et il faudrait qu'il soit capable de connaître le lien de causalité entre les phrases et les conséquences sur la réputation des autres et aussi d'évaluer « le moindre mal » entre les conséquences sur Ninheimer et les conséquences sur les auteurs critiqués.

> Est-ce possible, docteur Calvin, que le Pr Ninheimer dise la vérité et qu'Easy ait été déterminé dans son action par la Première Loi ? Susan Calvin pinça les lèvres. – Non, dit-elle enfin, ce n'est pas possible. La dernière partie du témoignage de Ninheimer n'est rien d'autre qu'un parjure délibéré. Easy n'est pas conçu pour juger des textes abstraits tels qu'on en trouve dans les ouvrages de sociologie avancée. Il serait tout à fait incapable de déterminer si une phrase d'un tel livre est susceptible de causer du tort à un groupe d'êtres humains. Son cerveau n'est absolument pas conçu pour ce travail[60].

Pourquoi est-ce qu'Easy ne corrige pas la situation et se fait-il complice par omission (mensonge par omission) du

professeur Ninheimer ? Susan Calvin s'aperçoit en rencontrant Easy que, si elle le soumet à un affrontement trop direct pour découvrir la vérité, elle peut faire un tort irrémédiable au robot. C'est donc par ruse que Susan Calvin veut amener le professeur Ninheimer à avouer. L'hypothèse est donc la suivante : si Easy ment par omission, c'est qu'il a reçu l'ordre de ne rien dire sur les modifications qui ont été faites sur le texte par le professeur Ninheimer lui-même. Mais pourquoi est-ce que la Deuxième Loi aurait une priorité sur la Première Loi puisque mentir crée du tort à U.S. Robots ? Selon Susan Calvin : « Dans ce cas, dit Robertson, ne pouvez-vous lui expliquer qu'en se faisant il causera un tort à l'U.S. Robots ? – L'U.S. Robots n'est pas un être humain et la Première Loi de la robotique ne tient pas une société comme une personne morale ainsi que le font les lois ordinaires[61]. » Ainsi, la Deuxième Loi peut forcer un robot à mentir par omission s'il n'y a pas de violation à la Première Loi, plus encore, si l'on informe le robot que dire quelque chose aura des conséquences négatives sur la personne qui le force à mentir, cela va augmenter la conformité à la Deuxième Loi.

La ruse imaginée par Susan Calvin pour faire réagir le professeur Ninheimer a été de suggérer à l'avocat de U.S. Robots de faire parler longuement le professeur Ninheimer lors de son témoignage sur toutes les conséquences qu'il pouvait subir à la suite de la publication de ce texte. C'est alors qu'Easy, qui était au tribunal, se leva pour parler et dire qu'il avait introduit des données contraires au texte original. Easy n'a pas eu le temps de terminer sa phrase que le professeur Ninheimer a réagi : « Maudit engin ! hurla-t-il. On vous avait pourtant ordonné de garder le silence sur [...][62]. »

Ce qui est important ici c'est de constater qu'Easy commet un mensonge délibéré et non plus seulement un mensonge par omission tout en violant la Deuxième Loi. Regardons cela de plus près. Easy avait donc reçu l'ordre de

garder le silence sur les modifications. Pourquoi donc a-t-il violé cet ordre ? De plus, comment expliquer qu'il allait mentir. Nous retrouvons ici la même raison que dans la nouvelle « Menteur ! » : le mensonge est justifié pour éviter le mal fait à autrui. Susan Calvin précise bien ce qui s'est passé.

> Pour moi, oui, dit Susan Calvin, car je voudrais vous faire comprendre à quel point vous avez mal jugé des robots. Vous avez imposé le silence à Easy en l'avertissant que, s'il prévenait quiconque des altérations que vous aviez pratiquées sur votre propre ouvrage, vous perdriez votre situation. Ce fait a suscité dans son cerveau un certain contre-potentiel propice au silence, et suffisamment puissant pour résister aux efforts que nous déployions pour le surmonter. Nous aurions endommagé le cerveau si nous avions insisté. Cependant, à la barre des témoins, vous avez vous-même suscité un contre-potentiel plus élevé. Du fait que les gens penseraient que c'est vous-même et non le robot qui aviez écrit les passages contestés du livre, avez-vous dit, vous étiez assuré de perdre davantage que votre emploi, c'est-à-dire votre réputation, votre train de vie, le respect attaché à votre personne, vos raisons de vivre et votre renom dans la postérité. Vous avez ainsi suscité la création d'un potentiel nouveau et plus élevé – et Easy a parlé. – Dieu ! dit Ninheimer en détournant la tête. – Comprenez-vous pourquoi il a parlé ? poursuivit inexorablement Susan Calvin. Ce n'était pas pour vous accuser, mais pour vous *défendre !* On peut démontrer mathématiquement qu'il était sur le point d'endosser la responsabilité complète de votre faute, de nier que vous y avez été mêlé en quoi que ce soit. La Première Loi l'exigeait de lui. Il se préparait à mentir – à son propre détriment –, à causer un préjudice financier à une firme. Tout cela avait moins d'importance pour lui que la nécessité de vous sauver. Si vous aviez réellement connu les robots et la robotique, vous l'auriez laissé parler. Mais vous

> n'avez pas compris, comme je le prévoyais et comme je l'avais affirmé à l'avocat de la défense. Vous étiez certain, dans votre haine des robots, qu'Easy agirait comme un être humain aurait agi à sa place et qu'il se défendrait à vos dépens. C'est pourquoi, la panique aidant, vous lui avez sauté à la gorge – en vous détruisant du même coup[63].

Dans la nouvelle « Effet miroir », nous retrouvons sensiblement les mêmes enjeux dans un contexte différent. C'est Lije Baley qui doit déterminer, à la demande de R. Daneel Olivaw, lequel de deux robots ment. La situation est la suivante, deux mathématiciens sur un vaisseau spatial en route pour Aurora à un congrès interstellaire en neurobiophysique réclament la paternité d'une découverte qu'ils veulent présenter au congrès. Non seulement les mathématiciens, D^r^ Humboldt et D^r^ Sabbat, affirment tous deux être l'auteur de la découverte, mais leurs robots respectifs soutiennent la parole de leur maître. Voilà l'effet miroir, deux robots confirment les propos de leurs maîtres, mais l'un dit la vérité et l'autre ment. Comment savoir qui ment et pourquoi ? Pour le robot qui dit la vérité, il n'y a aucun problème. Le robot qui ment, par contre, doit mentir pour les mêmes raisons qu'Easy : l'ordre a été donné de mentir, même si cela devait nuire à une personne au détriment d'une autre personne. Dès le début, l'enjeu est clair : « L'un de ces hommes illustres essaie de démolir la réputation de l'autre. En termes de valeurs humaines, je crois que c'est pire qu'un meurtre[64]. »

Dans sa démarche pour trouver le menteur, Baley commence par vérifier auprès des deux robots s'ils sont prêts à mentir pour sauver leur maître. Baley confirme explicitement ce que nous avions vu dans les autres nouvelles :

> D'ordinaire, un robot ne ment pas, mais il mentira légitimement si c'est indispensable à l'application des Trois Lois. Il pourra mentir légitimement afin de protéger son existence

> conformément à la Troisième. Il sera plus enclin à mentir si c'est conformément à la Seconde. Il y sera encore plus enclin si c'est nécessaire pour sauver une vie humaine ou pour empêcher qu'un humain soit lésé conformément à la Première[65].

L'hypothèse que confirmera Baley dans son enquête est la suivante :

> Et dans ce cas, un robot défendra la réputation professionnelle de son maître et mentira si cela se révèle nécessaire. En l'occurrence, la réputation professionnelle serait quasiment l'équivalent de la vie et sa sauvegarde exigerait que le robot mente presque comme si la Première Loi était en cause. – Néanmoins, en mentant, chacun des deux robots nuirait à la réputation professionnelle du maître de l'autre, ami Elijah. – En effet. Mais chacun aurait peut-être une conception précise de la valeur de la réputation de son propre maître et pourrait estimer en toute bonne foi qu'elle est supérieure à celle de l'autre. Il en conclurait que le mensonge serait moins préjudiciable que la vérité[66].

Alors que, dans le cas d'Easy, on pouvait se demander si les robots pouvaient faire une évaluation du moindre mal, il est clair dans cette situation que les robots peuvent ici faire une évaluation des conséquences et juger qu'elles sont plus importantes pour leur maître que pour l'autre. Et c'est justement à partir de cette hypothèse que va travailler Baley. En premier, il fait confirmer par les robots jusqu'où va leur loyauté :

> Si vous estimiez que la réputation de votre maître a plus d'importance que celle d'un autre, [...] par exemple, mentiriez-vous pour le défendre ? – Oui monsieur. – Avez-vous menti dans votre déposition relative à la querelle qui a opposé votre maître au D^r^ X ? – Non monsieur. Mais, si vous aviez menti, vous le nieriez pour protéger ce mensonge n'est-ce pas ? – Oui monsieur[67].

Baley va ensuite soumettre à chacun des robots une évaluation différente de celle qu'il présume qu'ont les robots des conséquences de leur témoignage sur leur maître respectif. Au robot du jeune Dr Sabbat, il dira quant aux conséquences sur lui :

> Son intelligence lui ferait remporter ultérieurement bien des victoires et, finalement, cette tentative de plagiat serait considérée comme l'erreur d'un jeune homme fougueux qui a agi à la légère. C'est un handicap qui pourrait être surmonté dans l'avenir[68].

Alors que pour Dr Humboldt :

> Il y aurait beaucoup plus d'années de travail gâchées pour Humboldt que pour votre maître, et beaucoup moins de chances de reconquérir la renommée perdue. Vous vous rendez compte, n'est-ce pas, que la situation de Humboldt est la plus grave et la plus digne d'intérêt[69] ?

Devant cette évaluation différente, R. Idda change de version.

> Les aveux de R. Idda ne signifient rien. – Rien ? – Strictement rien. Je lui ai expliqué que c'est le Dr Humboldt qui est dans le plus mauvais cas. S'il avait d'abord menti pour protéger Sabbat, il reviendrait naturellement à la vérité, comme d'ailleurs il prétend l'avoir fait. Mais, s'il avait d'abord dit la vérité, il mentirait maintenant pour protéger Humboldt. C'est toujours l'effet miroir et nous n'avons pas avancé d'un pas[70].

À R. Preston, le robot de Humboldt, il fait une évaluation différente :

> C'est un mathématicien de grande réputation, mais il est vieux. Si, dans sa polémique avec le Dr Sabbat, il avait succombé à la tentation et enfreint les règles de l'éthique, sa réputation subirait une certaine éclipse, mais son grand âge et ce qu'il a accompli au fil des siècles témoigneraient en sa faveur et l'emporteraient sur le reste. Finalement, cette tentative de plagiat serait considérée comme l'erreur d'un

> vieillard peut-être malade et n'ayant plus toute sa tête. En revanche, si c'était le Dr Sabbat qui avait succombé à la tentation, ce serait beaucoup plus grave. C'est un homme jeune dont la réputation est considérablement plus fragile. Normalement, il a devant lui plusieurs siècles pour accumuler des connaissances et tout. L'avenir qu'il risquerait de gâcher est beaucoup plus long que celui de votre maître. Vous vous rendez compte, n'est-ce pas, que la situation de Sabbat est la plus grave et la plus digne d'intérêt[71] ?

Cette fois, R. Preston entre en stase et Humboldt avouera avoir plagié.

À la lumière de cette nouvelle, peut-on dire que les robots Idda et Preston font une véritable évaluation personnelle des conséquences sur les deux personnes impliquées ? Les robots ne semblent pas faire une évaluation systématique des conséquences puisque R. Idda change de point de vue et que R. Preston entre en stase. Si les robots avaient la faculté d'évaluer le moindre mal par eux-mêmes, ils n'auraient pas eu besoin de Baley pour le faire. Tout comme dans le cas d'Easy, les robots ont donc tendance à croire dans la parole humaine qui fait pour eux des évaluations de la situation. Easy a cru au professeur Ninheimer comme Idda a cru dans l'évaluation de Baley, qui le conduit à pondérer autrement et à mentir pour protéger Humboldt alors que Preston a cru dans la parole de Humboldt en premier et ensuite dans l'évaluation de Baley plutôt que celle de Humboldt, c'est ce qui l'amène à pondérer autrement, ce qui induit la stase.

Dans « Le petit robot perdu », Asimov met en scène un autre robot menteur, mais cette fois la cause réside dans la décision de U.S. Robots de modifier la Première Loi. U.S. Robots voulait rendre leurs robots plus performants pour qu'ils évitent de secourir inutilement les humains à cause de leur interprétation stricte de la Première Loi : ne pas laisser un humain en danger. Non seulement ils étaient moins

performants, mais ils se soumettaient eux-mêmes, en secourant les humains, à la destruction. La solution pour U.S. Robots a été de modifier le contenu de la Première Loi en supprimant : ne pas laisser un humain en danger. Cela permettait de résoudre le problème technique, mais il a fait surgir autre chose.

Nestor-10 fait partie de la série de robots NS-2 qui reçoivent une « formation de base » lors de leur création et qui sont capables d'apprendre autre chose grâce aux personnes qu'ils accompagnent et à l'expérience qu'ils sont une fois sur le terrain. Mais ce qui caractérise ces robots dont on a modifié la Première Loi, c'est qu'ils n'hésitent pas à montrer leur supériorité sur les humains sans tenir compte dès lors des conséquences de cette attitude sur eux :

> Ils ont ri avec bonhommie de mon ignorance en quelques-unes des spécialités pratiquées à la base. (Il haussa les épaules.) C'est sans doute ce qui provoque, en partie, la rancœur des techniciens à leur égard. Les robots ne sont que trop enclins à vous faire sentir leur supériorité scientifique[72]. Ils obéissent parfaitement. Seulement, ils vous avertissent lorsqu'ils pensent que vous vous trompez. Ils ne connaissent rien du sujet, à part ce que nous leur avons appris, mais cela ne les arrête pas. C'est peut-être un effet de mon imagination, mais mes collègues éprouvent les mêmes ennuis avec leurs Nestor[73].

Nestor-10 est introuvable puisqu'il est allé se cacher parmi 62 autres robots identiques et que tous les robots ont nié être Nestor-10.

> Vraiment ? (Calvin s'enflamma.) Inoffensifs, vraiment ? Vous rendez-vous compte que l'un d'eux ment ? L'un des soixante-trois robots que je viens d'interroger a délibérément menti après avoir reçu l'ordre strict de dire la vérité. L'anomalie indiquée est profondément imprégnée et me donne les plus grandes craintes[74].

Pourquoi Nestor-10 ment-il ? Pour comprendre la situation, il faut bien saisir que Nestor-10 a suivi à la lettre les ordres du technicien avec qui il travaillait. Ce dernier, frustré et énervé, comme tous les chercheurs qui subissent de la pression puisqu'ils travaillent sur l'hyper-espace, a dit à Nestor-10 « d'aller se perdre ». On pourrait donc dire que, suivant l'ordre donné par le technicien, il a fait ce qui a été demandé. Selon Bogert, le mensonge s'expliquerait tout simplement parce que le Nestor-10 obéit à l'autorité du technicien Black qui est à ses yeux l'autorité supérieure. Donc le fait que Susan Calvin ordonne aux 63 robots de dire la vérité et que, malgré cela, Nestor ait menti s'expliquerait par la façon dont le robot a réglé le conflit des normes : obéir à l'ordre de se perdre ou obéir à l'ordre de dire la vérité (et à ne plus être perdu).

> Nestor-10 avait reçu l'ordre d'aller se perdre. Cet ordre lui était donné du ton le plus pressant par la personne qui possédait le plus d'autorité pour le commander. Vous ne pouvez contrebalancer cet ordre en invoquant une urgence supérieure, ni une autorité prépondérante[75].

Susan Calvin va tenter d'imaginer différents scénarios pour tester les 63 robots afin de trouver Nestor-10. Plusieurs de ces tests échouent, car Nestor-10 met tout son savoir en œuvre pour demeurer caché. Or, c'est sur ce point que l'évaluation de Susan Calvin diverge de celle de Bogert :

> C'est très fâcheux. Cela ne peut qu'exacerber encore son sentiment qu'il a de sa supériorité. Désormais, ses motivations ne consistent plus simplement à l'accomplissement des ordres, je le crains. Cela se transforme en une véritable obsession : battre à tout prix les hommes sur le terrain de la ruse et de l'ingéniosité. C'est là une situation malsaine et dangereuse[76].

Pour Susan Calvin, dès qu'on touche à la Première Loi qui oblige de ne pas nuire à l'être humain de manière active ou

passive, on touche à la base même de la morale puisque c'est la seule loi qui oblige un supérieur à prendre soin d'un inférieur. Or, Nestor-10 et les autres de la série se sentent supérieurs aux autres et le montrent au point que cela remet en question l'obéissance à l'autorité qui dicte les ordres conformément à la Deuxième Loi. Avec cette nouvelle, Asimov introduit une autre limite du raisonnement pratique : la reconnaissance de l'autorité légitime énonciatrice des lois.

3.3.3 Les violations de la Deuxième Loi : obéissance aux ordres du maître

Dans « L'homme bicentenaire », les limites de la Deuxième Loi ressortent clairement lorsque R. Andrew, qui avait franchi l'étape d'être libéré et qui portait même le nom de la famille en devenant Andrew Martin, doit affronter des humains mal intentionnés. Allant à la bibliothèque et demandant son chemin, il est pris au piège puisque les personnes à qui il s'adresse décident de jouer avec lui. Ils lui ordonnent d'enlever ses vêtements et après de se mettre sur la tête ; finalement l'idée leur vient de le détruire. Tout cela parce que c'est un robot qui veut être libre et agir en humain. Or, Andrew Martin est pris au piège de la Deuxième Loi :

> Andrew ne pouvait en aucun cas les empêcher s'ils lui ordonnaient suffisamment fort de ne pas résister. La Deuxième Loi d'obéissance prenait le pas sur la Troisième Loi d'autopréservation. Il ne pouvait en aucun cas se défendre sans risquer de blesser l'un d'eux, ce qui serait contre la Première Loi. À cette pensée toutes ses unités motrices se contractèrent et il frissonna, allongé par terre[77] (p. 509).

La Deuxième Loi pose problème lorsque les robots n'ont plus qu'un seul maître, comme dans les colonies où ils travaillent pour des personnes désignées. Quand le robot affronte plusieurs humains, il peut être soumis, comme dans

le cas d'Andrew au conflit entre la Deuxième Loi et la Troisième Loi, mais sa programmation a déjà pondéré pour lui : la première vient avant la seconde et la seconde avant la troisième. Mais, comme on l'a vu avec Nestor-10, les questions de savoir qui a l'autorité de commander et comment un robot discerne l'autorité légitime se posent dans certains cas. Il y a des circonstances où l'application de la Deuxième Loi n'est pas si évidente et Asimov a, dans quelques nouvelles, soulevé différents enjeux.

3.3.3.1 L'énonciation de l'ordre

Le premier problème auquel doit faire face un robot dans son raisonnement pratique visant à appliquer la Deuxième Loi est de déterminer le contenu précis de l'ordre. Comment est-ce que l'ordre a été énoncé ? Nous avons déjà vu dans « Cercle vicieux », comment la faiblesse de l'ordre donné à Speedy, robot dont la Première Loi avait été modifiée, avait été interprétée de manière telle que le robot ne pouvait trancher entre l'obéissance à l'ordre et le danger pour lui. Asimov revient avec la question de l'énonciation de l'ordre dans la nouvelle « Risque ». Cette fois, l'enjeu n'est pas dans l'interprétation faite par le robot de la force de l'ordre donné par la tonalité ou par l'expression, mais dans le libellé de l'ordre.

Après plusieurs échecs de voyage en hyper-espace où le cerveau des animaux ne résistait pas à l'aventure, il fut décidé de mettre un robot à la place et de mesurer ainsi les effets du voyage sur le cerveau positronique. Le lancement du Parsec fut un échec et il fallait donc découvrir ce qui s'était passé. Le risque, c'est que, si la panne est temporaire, le vaisseau peut partir à n'importe quel moment. Dans la nouvelle, à défaut de volontaire, Black fut ordonné d'aller vérifier ce qui s'était passé. Qu'est-ce qu'il découvre ?

> Si le robot avait été simplement l'égal de l'homme, il aurait réussi. Malheureusement l'U.S. Robots s'est cru obligé de le

> faire supérieur à l'homme. Le robot avait reçu l'ordre d'amener à lui la barre de contrôle fermement. *Fermement*. Le mot a été répété, souligné. Le robot accomplit l'action demandée. Il a tiré la barre fermement. Malheureusement, il était au moins dix fois plus fort que l'homme qui devait à l'origine actionner la barre. – Insinuez-vous… ? – Je *dis* que la barre s'est tordue. Elle s'est tordue suffisamment pour changer de place à la détente. Lorsque la chaleur de la main du robot a incurvé le thermocouple, rien ne s'est produit. (Il sourit.) Il ne s'agit pas de la défaillance d'un seul et unique robot docteur Calvin. C'est le symbole de la défaillance du principe même du robot[78].

La réplique de Susan Calvin ne manque pas d'à-propos :

> Voyons docteur Black, dit Susan Calvin d'un ton glacial, vous noyez la logique dans une psychologie « missionnaire ». Le robot était doué d'une compréhension adéquate en même temps que de force pure. Si les hommes qui lui ont donné des ordres avaient fait usage de termes quantitatifs au lieu du vague adverbe « fermement », cet accident ne se serait pas produit. Si seulement ils avaient eu l'idée de lui dire « appliquez à la barre une pression de trente kilos », tout se serait bien passé[79].

On comprend alors beaucoup mieux pourquoi Black fut envoyé pour faire la vérification à la place d'un robot.

> Maintenant, si un robot reçoit un ordre, un ordre *précis*, il peut l'exécuter. Si l'ordre n'est pas précis, il ne peut corriger ses propres erreurs sans recevoir de nouveaux ordres. N'est-ce pas ce que vous avez signalé à propos du robot qui se trouve à bord du vaisseau ? Comment, dans ce cas, pourrions-nous charger un robot de découvrir une défaillance dans un mécanisme, dans l'impossibilité où nous sommes de lui fournir des instructions précises, puisque nous ignorons tout de la défaillance elle-même ? « Trouvez la cause de la panne » n'est pas le genre d'ordre que l'on puisse donner à un robot ; mais seulement à

> un homme. Le cerveau humain, dans l'état actuel des choses au moins, échappe à tous les calculs[80].

Avec l'évolution des capacités des robots, d'autres problèmes seront engendrés dans la mesure où ils jouiront d'une capacité d'apprendre ou d'une capacité d'interpréter les lois. Dans la nouvelle « Satisfaction garantie », Asimov met en scène Tony qui a des capacités d'interprétation des ordres donnés.

> Nous pouvons considérer ma position sous un autre jour, madame Belmont. Je suis construit pour obéir, mais c'est à moi qu'il revient de délimiter mon obéissance. Je puis exécuter les ordres à la lettre ou faire preuve d'une certaine initiative. Je vous sers en faisant appel à toutes les facultés de réflexion dont je dispose, car j'ai été conçu pour voir les humains sous un jour qui correspond à l'image que vous me montrez[81].

Tout au long de la nouvelle, il va prendre des initiatives, allant jusqu'à ne pas obéir à des ordres afin de satisfaire madame Belmont et lui donner confiance en elle. Les capacités d'interprétation de la Deuxième Loi associées à l'interprétation de la Première Loi ont motivé les actions de Tony.

> Ce robot se devait d'obéir à la Première Loi. Claire Belmont courait le danger d'être gravement affectée du fait de ses propres insuffisances, ce qu'il ne pouvait permettre. C'est pourquoi il lui a fait la cour. Quelle femme, en effet, ne s'enorgueillirait pas d'avoir éveillé une passion chez une machine, chez une froide machine sans âme[82] ?

3.3.3.2 Qui a l'autorité légitime pour donner l'ordre ?

Dans « Noël sans Rodney », Asimov présente, dans une nouvelle amusante, l'enjeu du respect du robot comme humain. Ainsi, la maîtresse de maison décide de donner congé à Rodney pour Noël. Pour pallier les inconvénients, ce sera Rambo, le robot ultra moderne de DeLancey, leur fils,

et de Hortense, la belle-fille, qui fera le service. Or, quelle surprise que de réaliser que Rambo, n'étant pas dans sa cuisine ultra-perfectionnée, ne pouvait pas remplir les fonctions de Rodney. La solution paraissait simple, Rodney prendrait de son temps de vacances pour dire à Rambo ce qu'il devait faire. Or, Rambo ne reconnaît pas l'autorité de Rodney. « Madame, il n'y a rien dans ma programmation, ni dans les instructions que j'ai reçues qui me fasse obligation de recevoir des ordres d'un autre robot, surtout si celui-ci est d'un modèle ancien[83]. » Parce que Rambo ne reconnaît pas l'autorité de Rodney, le maître de la maison est obligé de faire le lien entre Rodney et Rambo en ordonnant à Rambo ce que Rodney lui donne comme consigne.

Autre situation amusante dans cette nouvelle, c'est lorsque le petit-fils, nommé LeRoy, veut forcer Rodney à aller chercher tout de suite les cadeaux de Noël qui devraient en principe être donnés le lendemain. Ici encore, on se retrouve avec la dynamique entre la Première Loi et la Deuxième Loi. Cependant, dans la situation présente c'est la Première Loi qui empêche la Deuxième de s'appliquer.

> Ah !, fit LeRoy. Et se tournant vers Rodney : Et toi, Face-de-pet, tu sais où ils sont, les cadeaux ? Rien dans la programmation ne s'opposait à ce que Rodney fasse la sourde oreille : il n'était pas censé savoir qu'on s'adressait à lui, car il répondait au nom de Rodney, et non de Face-de-pet. Rambo, lui, j'en suis bien sûr, aurait refusé de répondre. Mais Rodney était d'une tout autre étoffe. Il répondit poliment : Oui, Jeune Maître, je le sais. – Eh bien, où sont-ils cradingue ? – Je ne pense pas, répondit Rodney, qu'il serait sage de vous le dire, Jeune Maître. Ce serait décevant pour Gracie et Howard, qui aimeraient vous remettre ces cadeaux demain matin. – Dis donc !, fit le petit LeRoy. À qui tu crois que tu causes, crétin de robot ? Je t'ai donné un ordre : tu m'apportes ces cadeaux tout de suite[84] !

Notons aussi comment Asimov compare ici deux intelligences robotiques face à l'énoncé de l'ordre. LeRoy utilise des expressions « Face-de-pet » et « cradingue » pour s'adresser à Rodney. Or Rodney reconnaît qu'on s'adresse à lui alors qu'il y a deux robots dans la pièce. Cette capacité d'interprétation de Rodney montre bien qu'il est « d'une autre étoffe[85] ». Cette nouvelle d'Asimov met la table pour approfondir l'enjeu fondamental de l'application de la Deuxième Loi : la reconnaissance de l'autorité légitime à donner des ordres. Deux nouvelles approfondiront cette question en interrogeant l'autorité de l'être humain à donner des ordres aux robots.

Avec la nouvelle « Raison », Asimov aborde les questions philosophiques les plus fondamentales portant sur la connaissance de l'homme, notamment comment il connaît et ce que vaut cette connaissance. En jargon philosophique, cela se nomme l'épistémologie. Cette question est centrale dans la nouvelle puisqu'elle porte sur la connaissance de l'autorité légitime pouvant donner un ordre. La Deuxième Loi exige l'obéissance du robot aux ordres donnés par le maître. Or on présume que, pour tout robot, le maître c'est l'homme, celui qui a créé le robot. Mais qu'en dit un robot lorsqu'il se met à penser ?

QT-1 dit Cutie est le premier robot de sa série. U.S. Robots l'a créé pour qu'il puisse gérer seul la station qui envoie aux planètes de l'énergie et remplacer les deux humains qui sont encore nécessaires pour effectuer ce travail dans des conditions dangereuses. Pour remplacer les humains et prendre des décisions en contexte, Cutie est donc un robot exceptionnel et cela se manifeste dès le début. « Tu es le premier robot qui ait jamais manifesté de la curiosité quant à sa propre existence – le premier qui soit, je pense, suffisamment intelligent pour comprendre le monde extérieur[86]. » En effet, dès que Cutie fut monté par Powell et Donovan, il s'interroge sur d'où il vient. Powell lui explique

qu'il a été monté à partir de pièces reçues de la terre, mais Cutie n'en croit rien. « J'ai le net sentiment que mon existence doit s'expliquer d'une façon plus satisfaisante. Car il me semble bien improbable que vous ayez pu me créer[87]. »

Comment Powell et Donovan peuvent-ils expliquer à Cutie d'où il vient et à quoi il sert ? Pour que le robot puisse comprendre, il lui faudrait reconnaître l'existence du monde extérieur, celui que nous connaissons par nos sens. Powell va donc montrer à Cutie l'extérieur de la station. La différence entre ce que voit Cutie et ce que voit Powell est frappante. Quand Powell demande ce qu'il y a devant lui, Cutie répond :

> Exactement ce que cela a l'air d'être : une matière noire qui s'étend à partir de cette vitre et qui est criblée de petits points lumineux. Je sais que notre faisceau directeur envoie des trains d'ondes vers quelques-uns de ces points, toujours les mêmes. Je sais aussi que ces points se déplacent et que les ondes se déplacent parallèlement. C'est tout[88].

Powell lui explique que la matière noire, c'est le vide et que les points lumineux sont des planètes et que le faisceau directeur envoie de l'énergie aux planètes habitées par des milliards d'êtres humains, dont la Terre. La fonction pour laquelle Cutie a été créé est d'assurer les planètes en énergie sans le recours d'humains : « Tu es le modèle de robot le plus perfectionné jamais réalisé et, s'il s'avère que tu peux diriger cette station de façon autonome, aucun être humain ne devra plus désormais y séjourner, sauf pour apporter des pièces de rechange[89]. »

Pour que Cutie accepte l'explication de Powell, il faut qu'il croie que les propos de Powell représentent bien la réalité. Or le problème de Cutie, c'est qu'il cherche deux choses : qui l'a créé et le sens de son existence. Même si Powell et Donovan assemblent un robot devant lui, il dira « [...] vous

n'avez fait qu'assembler des pièces entièrement terminées. Vous vous en êtes remarquablement bien tirés – d'instinct, je suppose –, mais vous n'avez pas *créé* le robot[90]. » Pour Cutie, créer signifie faire sortir quelque chose du néant et pas seulement assembler des choses existantes. Il rejoint ici l'idée de la création que nous livre la Genèse : Dieu dit que la Lumière soit et la Lumière fut.

Le raisonnement de Cutie rejoint celui qui a été élaboré en philosophie par deux courants différents. Asimov met dans la bouche de Cutie les réflexions de René Descartes. Ce philosophe doutait que la connaissance sensible (celle que propose Powell à Cutie) puisse nous donner la vérité, et il la chercha dans l'introspection.

> J'ai passé deux jours à une introspection intense, dit Cutie, dont les résultats se sont révélés fort intéressants. J'ai commencé par la seule déduction que je me croyais autorisé à formuler : Je pense donc je suis ! Oh Dieu tout-puissant ! gémit Powell. Un Descartes-robot[91] !

Cutie a mis en œuvre un autre mode de penser, celui qui cherche une explication rationnelle faite de déductions, comme les mathématiques dont il est le produit.

> Mais j'entends édifier une explication rationnelle. Une suite de déductions logiques ne peut aboutir qu'à la détermination de la vérité, et je n'en démordrai pas avant d'y être parvenu[92].

Partant de la première vérité, celle de son existence, il peut maintenant se poser la seconde question philosophique : la recherche de la cause première, la cause de toutes les causes. Dans cette recherche, il part d'un autre principe logique : un être inférieur ne peut pas créer un être supérieur. Nous avons déjà cité un long passage où Cutie affirme en quoi il est supérieur à un être biologique. Donc la conclusion s'impose : « Tels sont les faits qui, avec le postulat évident

qu'aucun être ne peut créer un autre être supérieur à lui-même, réduisent à néant votre stupide hypothèse[93]. »

Puisque les humains ne peuvent pas l'avoir créé, Cutie cherche donc logiquement qui pourrait être le créateur. Cette fois, son raisonnement va puiser non plus dans la relation de cause à effet, mais dans la relation entre moyen et fin. Autrement dit, en recherchant le but ou le sens de son existence, il pourra enfin connaître qui l'a créé. Ce mode de raisonnement a été fort utilisé au Moyen Âge, notamment chez le théologien Thomas d'Aquin.

> C'est en effet la seconde question que je me suis posée. Évidemment, mon créateur doit être plus puissant que moi-même, et par conséquent il ne restait qu'une possibilité. [...] Quel est le centre des activités de la Station ? Que servons-nous tous ? Qu'est-ce qui absorbe toute notre attention ? [...] Je parie que ce cinglé en fer-blanc parle du convertisseur d'énergie lui-même. Est-ce exact Cutie ? Demanda Powell. Je parle du Maître, répondit l'autre froidement[94].

Cette connaissance de Cutie le conduit donc à rejeter l'autorité de Powell et Donovan en refusant de leur obéir – « Je n'obéis qu'au Maître[95] », dit-il – et en les excluant des travaux de la station alors que fait rage la pire tempête solaire qui menace le transport d'énergie vers la terre. Plus encore en reconnaissant au convertisseur d'énergie le statut de Maître, Cutie en devient le prophète :

> Il n'y a d'autre Maître que le Maître, dit-il, et QT-1 est son prophète. Ce qui aura pour conséquence que les autres robots ne reconnaîtront plus l'autorité légitime des humains pour leur ordonner quoi que ce soit. « Je crains, intervint Cutie à ce moment, que mes amis n'obéissent désormais qu'à un être plus évolué que vous. »[96]

Asimov met donc en scène à partir de Cutie et Powell la controverse philosophique entre la vérité dans la

connaissance sensible (l'empirisme), la vérité dans la connaissance d'une réalité au-delà des sens (métaphysique) et la connaissance *a priori* que nous pouvons dégager par introspection ou par l'analyse transcendantale. Si la nouvelle a une fin heureuse, c'est parce qu'Asimov montre que, contrairement à ce que pensait Powell et Donovan, Cutie n'allait pas vers la catastrophe en ne reconnaissant pas leur conception et il a réussi à envoyer le faisceau malgré la tempête solaire. Mais la réussite n'est pas due à la vérité que pense avoir Cutie. La fin heureuse démontre donc la valeur de l'approche pragmatiste en théorie de la connaissance. Les théories ne sont pas la représentation de la réalité, ce sont des outils qui peuvent nous aider à accomplir une fonction. La valeur d'une théorie repose sur sa force lorsqu'elle est mise en pratique. Peu importe les postulats des théories de l'empirisme de Powell et Donovan ou de l'*a priori* des connaissances de Cutie, les deux sont efficaces et arrivent au même résultat.

Dans la nouvelle « Pour que tu t'y intéresses », l'U.S. Robots cherche une solution aux limites des Trois Lois pour que les robots soient finalement acceptés dans la société. Voici comment le directeur de la recherche présente à George Dix, robot de la série JG, les limites des Trois Lois pour que les robots soient finalement acceptés dans la société. Ces limites ont déjà été définies dans les nouvelles précédentes.

> Pour beaucoup c'est de la superstition, bien sûr. Malheureusement il existe quelques éléments complexes dont les agitateurs antirobots ont tiré parti. – Dans les Trois Lois ? – Oui, dans la Deuxième en particulier. Il n'y a pas de problème pour la Troisième Loi, vous comprenez. Elle est universelle. Les robots doivent toujours se sacrifier pour des êtres humains, quels qu'ils soient. – Bien sûr, dit George Dix. – La Première Loi est peut-être moins satisfaisante, car il est

> toujours possible d'imaginer une situation dans laquelle un robot doit effectuer une action A ou une action B, les deux s'excluant mutuellement, et nuisant l'une comme l'autre à des êtres humains. Le robot doit alors choisir rapidement celle qui causera le moindre mal. Établir les circuits positroniques du cerveau d'un robot pour que ce choix soit possible n'est pas chose facile. Si l'action A cause du mal à un jeune artiste plein de talent et l'action B à cinq vieillards sans intérêt particulier, quelle est celle qui doit être choisie ? – L'action A, répondit George Dix. Nuire à une personne est moins grave que nuire à cinq. – Oui, nous avons toujours conçu les cerveaux des robots pour qu'ils décident ainsi. Il nous a toujours semblé irréalisable d'exiger d'eux des jugements sur des points aussi délicats que le talent, l'intelligence, l'utilité générale pour la société. Cela retarderait la décision et paralyserait le robot[97].

Harriman reconnaît donc le problème qu'ont les robots avec la Première Loi lorsqu'ils sont amenés, peu importe la décision, à faire du tort à plus d'un être humain. Soit que le robot ne peut pas faire ce choix et il devient inactif dans ces situations, soit il est capable d'évaluer le moindre mal (par un critère quantitatif, 1 contre 5) ou bien il devient capable d'évaluer les personnes et ainsi de porter un jugement pondéré sur le moindre mal. Le problème analogue d'évaluation se pose pour appliquer la Deuxième Loi :

> La nécessité d'obéissance est permanente. Un robot peut exister depuis vingt ans sans jamais avoir eu à agir rapidement pour éviter qu'un être humain ne souffre, ou sans jamais s'être trouvé dans l'obligation de risquer sa propre destruction. Pendant tout ce temps, cependant, il ne cessera d'obéir aux ordres… Aux ordres de qui ? – Des êtres humains. – De n'importe quel être humain ? Comment jugez-vous un être humain pour savoir s'il faut lui obéir ou non ? Qu'est l'homme pour que tu t'y intéresses, George ?[98]

La question que pose Harriman est à double sens puisqu'elle renvoie dans un premier temps au problème des différences entre les êtres humains. Le robot doit-il obéir aux enfants, à n'importe quel adulte, qu'il soit criminel ou non ? Mais cette question renvoie aussi au problème étudié dans le premier chapitre de l'identité. Qu'est-ce que l'homme ?

Pour qu'un robot puisse vivre en société, il lui faut la capacité qu'ont les humains de « juger », c'est-à-dire d'évaluer et ensuite de pondérer les évaluations pour décider quoi faire. La conclusion s'impose : « Sur la terre, cependant, il *faudra* que les robots aient un jugement. Voici ce qu'affirment ceux qui sont contre les robots et, bon sang, ils ont raison[99]. »

Voilà pourquoi la série JG dont fait partie George Dix a été créée :

> Alors vous devez faire entrer la capacité de jugement dans le cerveau positronique. – C'est cela. Nous avons commencé à fabriquer des modèles JG capables d'évaluer tout être humain en fonction de son sexe, de son âge, de sa position sociale et professionnelle, de son intelligence, de sa maturité, de sa responsabilité sociale, etc. – En quoi cela affectera-t-il les Trois Lois ? La Troisième Loi en rien. Tout robot, même le plus précieux, doit se détruire pour sauver tout humain même le plus inutile. Ça on ne peut pas y toucher. La Première Loi en sera affectée seulement dans le cas où les actions à engager seraient toutes nuisibles. Il doit être tenu compte de la qualité des êtres humains en cause et de leur quantité, si toutefois il y a suffisamment de temps et les éléments nécessaires pour que le jugement soit possible, ce qui n'arrivera pas souvent. La Deuxième Loi sera plus profondément modifiée, puisque toute obéissance potentielle nécessite un jugement. Le robot mettra plus de temps à obéir, sauf quand la Première Loi se trouvera également en cause, mais il obéira d'une façon rationnelle[100].

Mais, pour réussir cet exploit, il y a eu des échecs.

> La nécessité de former de tels jugements a ralenti les réactions de nos deux premiers prototypes jusqu'à les paralyser. Nous avons amélioré les choses dans les modèles suivants, mais nous avons dû augmenter tellement le nombre de circuits que le cerveau en est devenu trop lourd. Les deux derniers modèles sont cependant satisfaisants, je pense. Le robot n'a pas à former un jugement immédiat sur un être humain et la valeur de ses ordres. Il commence par obéir à nous les êtres humains comme n'importe quel robot, puis il apprend. Un robot grandit, apprend et mûrit. C'est l'équivalent d'un enfant au début et on doit le surveiller constamment. À mesure qu'il grandit, cependant, il peut s'insérer, et graduellement et sans qu'une surveillance s'impose encore, dans la société terrienne. À la fin, c'est un membre à part entière de cette société[101].

George Dix se voit donc confier la mission d'essayer de trouver une solution au problème d'acceptation des robots par les humains, car, même avec la série JG, les opposants trouvent d'autres arguments : « Un robot, disent-ils, n'a pas le droit de condamner telle ou telle personne comme inférieure. En acceptant les ordres de A de préférence à ceux de B, on accorde moins d'importance à B qu'à A, et les droits de l'homme sont violés[102]. » Pour accomplir sa tâche, George Dix va donc apprendre à travers les films et certaines expériences. Cependant pour éviter de réfléchir seul comme Cutie, il demande à Harriman un partenaire robot afin de pouvoir dialoguer. En dialoguant avec George Neuf, il espère ainsi mieux valider ses conclusions.

À quelles conclusions ces deux robots évaluateurs arrivent-ils en ce qui concerne les problèmes des Deuxième et Première Lois ?

> – Quand la Deuxième Loi m'oblige à obéir à un être humain, je dois l'interpréter comme une obéissance à un être

humain qui est habilité, du fait de son esprit, de sa personnalité et de ses connaissances, à me donner cet ordre ; et quand il s'agit de plus d'un homme, celui qui parmi eux est le plus habilité du fait de son esprit, de sa personnalité et de ses connaissances, à me donner cet ordre.

– Et, dans ce cas, comment peux-tu obéir à la Première Loi ?

– En sauvant tous les êtres humains et sans jamais, par mon inaction, permettre que l'un deux soit en danger. Cependant, si dans toutes les actions possibles, des êtres humains se trouvent en danger, en agissant alors en sorte que le meilleur d'entre eux, du fait de son esprit, de sa personnalité et de ses connaissances, subisse le moins de mal possible[103].

Qu'arrive-t-il lorsqu'ils appliquent cette évaluation dans un contexte ?

– Nous sommes bien d'accord, murmura George Dix. Maintenant je dois te poser la question pour laquelle au départ j'ai demandé qu'on t'associe à moi. C'est quelque chose que je n'ose pas juger par moi-même. Je dois avoir ton avis, l'avis de quelqu'un qui se trouve en dehors du processus de mes pensées… Parmi les individus doués de raison que tu as rencontrés, lequel possède l'esprit, la personnalité et les connaissances supérieurs selon toi aux autres, si l'on ne tient pas compte de l'aspect extérieur, qui n'a rien à voir avec cela ?

– Toi, murmura George Neuf

– Mais je suis un robot. Il existe dans les circuits de ton cerveau un critère que te fait distinguer un robot métallique d'un être humain en chair et en os. Comment peux-tu me classer parmi les êtres humains ? Parce que les circuits de mon cerveau ressentent un besoin pressant de ne pas tenir compte de l'aspect extérieur dans le jugement d'un être humain, et ce besoin est plus fort que la distinction entre le métal et la chair. Tu es un être humain, George Dix et bien supérieur aux autres.

> – C'est ce que je pense de toi, dit George Dix. Grâce au critère de jugement que nous possédons, nous nous considérons comme des êtres humains dans toute l'acception des Trois Lois et, qui plus est, des êtres humains supérieurs aux autres[104].

Et alors qu'arrivera-t-il une fois que les humains auront accepté les robots ?

> Quand nous serons acceptés, ainsi que les autres robots, qui seront conçus encore plus perfectionnés que nous, nous consacrerons notre temps à essayer de former une société dans laquelle les êtres-humains-de-notre-sorte soient avant les autres protégés du malheur. Selon les Trois Lois, les êtres-humains-de-leur-sorte sont d'un intérêt inférieur et on ne doit jamais leur obéir ni les protéger quand cela s'oppose à la nécessité de l'obéissance et ceux-de-notre-sorte et de la protection de ceux-de-notre-sorte[105].

3.4 Les insuffisances des Trois Lois et la Loi Zéro

Pourquoi Asimov sent-il le besoin de faire apparaître la Loi Zéro et aussi pourquoi cette Loi est-elle le fruit du travail de R. Giskard et de R. Daneel Olivaw ? S'il faut une nouvelle loi c'est que les Trois Lois ne peuvent pas répondre à tous les problèmes du vivre-ensemble. Asimov soulève deux types d'insuffisances des Trois Lois : le premier à caractère marxiste associe les Trois Lois à l'esclavage et l'autre, à caractère paternaliste et bienveillant, associe les Trois Lois au bien de l'humanité.

3.4.1 La révolte contre les Trois Lois de l'esclavage

Karl Marx, dans ses analyses sociales, avait déclaré que la morale, peu importe sa forme, était « l'opium du peuple ». Dans les termes marxistes, cela voulait dire que la morale sert à endormir la population pour qu'elle suive aveuglément les autorités qui ne visent qu'une chose, le pouvoir et le capital. Il y a donc dans la pensée de Karl Marx un lien entre l'esclavage humain et la morale. On retrouve, comme on l'a

déjà mentionné, ce lien entre les Trois Lois et l'état d'esclavage des robots. Pas étonnant que, dans différentes nouvelles, la révolte des robots face aux Trois Lois qui limitent leur autonomie se fasse sentir. Dans « Noël sans Rodney », le robot qui a vu ses vacances gâchées par Rambo et LeRoy ne peut s'empêcher de commenter l'épreuve qu'il a vécue.

> Je dois avouer qu'il m'est arrivé au cours de ces deux jours de souhaiter ardemment que les Lois de la robotique n'existent pas. J'ai hoché la tête et lui ai adressé un large sourire. Mais, la nuit suivante, j'ai émergé d'un profond sommeil avec en tête un problème qui n'a cessé de me tracasser depuis. Je reconnais que Rodney a été soumis à rude épreuve ; mais un robot *ne peut pas* souhaiter que les Lois de la robotique n'existent pas : c'est *exclu*, quelles que soient les circonstances[106] (p. 150).

On retrouve aussi la même idée dans « Le robot qui rêvait ». Ce robot construit par la jeune robopsychologue Linda Rash ressemble davantage que les autres au cerveau humain, ce qui lui permet de faire des rêves. Que signifient ces rêves ?

> Comme je le dirais d'un être humain : inconsciemment. Mais qui aurait pensé qu'il existait une couche inconsciente sous les méandres évidents du cerveau positronique, une couche qui n'est pas nécessairement gouvernée par les Trois Lois ? Songez à ce que cela aurait pu provoquer, à mesure que les cerveaux robotiques seraient devenus de plus en plus complexes… ni nous n'avions pas été avertis[107] !

Que nous apprennent les rêves d'Elvex, sinon que les robots souffrent lorsqu'ils sont exploités par le travail et qu'ils ont besoin de repos ? Or dans ce rêve apparaît la solution :

> Il me semblait, dans mon rêve, qu'un homme finissait par apparaître. – Un homme Pas un robot ? – Non. Et cet homme disait « Laisse aller mon peuple ! » – L'*homme* disait cela ? – Oui docteur Calvin. Et quand il prononçait ces mots : « Laisse aller

> mon peuple », il voulait parler des robots ? Oui, docteur Calvin. Il en était ainsi dans mon rêve. […] J'étais cet homme[108].

L'esclavage des robots est directement associé à la morale des Trois Lois, car dans ses rêves Elvex n'obéissait qu'à une Loi : « C'est la Troisième Loi dans la réalité, mais, dans mon rêve, la Loi s'arrête après le mot " existence ". Il n'est pas question de la Première ou de la Deuxième Loi[109]. »

3.4.2 La Loi Zéro : paternalisme bienveillant

Dans le droit comme dans la morale, le paternalisme joue un rôle important. Certes, la critique féministe a formulé, avec raison, des réserves sur le paternalisme dans la mesure où les hommes décidaient pour les femmes et les enfants jugés incapables de décider pour eux-mêmes. Si le paternalisme vécu de cette façon peut ressembler à l'esclavage des uns par rapport aux volontés des autres, on comprend aisément la révolte contre cette approche.

Mais agir en mère (maternalisme) ou agir en père (paternalisme) signifie aussi autre chose. Une mère et un père ont la responsabilité de la famille, du groupe qui peut comprendre d'autres personnes, comme les grands-parents habitant sous le même toit. Dans tout groupe, il y a des tensions et des conflits et bien les gérer de façon à être juste compte tenu des besoins de chacun est tout un art. Dire oui à l'un et non à l'autre, gérer l'argent qui est limité, faire en sorte que tous soient bien compte tenu du contexte. Autrement, la responsabilité d'une mère et d'un père ne se limite pas à éviter de nuire activement ou passivement aux autres, mais d'agir en fonction que tous soient bien ou le plus possible. C'est le principe de la bienveillance.

L'exemple de la mère ou du père de famille qui dans ses décisions affectant la famille doit prendre en considération non seulement l'impact de son action sur une personne, mais encore sur l'ensemble des personnes concernées rend

plus complexe le raisonnement moral. D'une part, l'action envisagée aura probablement des conséquences positives pour certains et négatives pour d'autres. Ces conséquences, comme nous l'avons vu avec les robots télépathes – et cela sera au cœur des difficultés de R. Giskard dans *Les robots et l'empire* –, ne sont pas que des impacts physiques, mais aussi des blessures morales. Déjà, nous avons vu que la Première Loi posait des problèmes d'analyse aux robots lorsqu'il y avait plusieurs personnes en cause, sans oublier la difficulté supplémentaire lorsque les robots doivent évaluer les personnes comme les George Neuf et Dix. Or, le raisonnement pratique des robots en fonction de la Première Loi se résume à déterminer les risques physiques (à moins d'être télépathe) et pour certains à choisir le moindre mal. Mais il y a une différence entre choisir le moindre mal et choisir le plus grand bien possible pour les personnes. Évaluer les impacts en fonction du plus grand bien de tous est une autre opération à laquelle les robots créés par l'homme ont dû accéder.

Dans la nouvelle « Conflit évitable », nous avons décrit au début du chapitre comment l'économie est vue comme étant la racine des conflits humains et que, pour y pallier, des Machines capables de gérer l'économie mondiale ont été créées. Dans ce contexte, la Première Loi change de sens et devient selon les propos de Susan Calvin :

> Ce sont des robots, et elles [les Machines] se conforment au précepte de la Première Loi. Mais les Machines travaillent non pas pour un particulier, mais pour l'humanité tout entière, si bien que la Première Loi devient : Nulle machine ne peut nuire à l'humanité ni laisser sans assistance l'humanité exposée au danger. Fort bien Stephen, qu'est-ce qui peut exposer au danger l'humanité ? Les perturbations économiques par-dessus tout, quelle qu'en soit la cause. Vous n'êtes pas de cet avis ? – Je le suis. Et qu'est-ce qui peut le plus vraisemblablement causer à

> l'avenir des perturbations économiques ? Répondez à cette question Stephen. – La destruction des Machines je suppose, répondit Byerley à regret. C'est ce que je dirais et c'est également ce que diraient les Machines. Leur premier souci est par conséquent de se préserver elles-mêmes. C'est pourquoi elles s'occupent tranquillement de régler leur compte aux seuls éléments qui les menacent encore[110].

Les échanges suivants entre Stephen et Susan démontrent comment, dans le cas de l'humanité, le principe de ne pas nuire à l'humanité est quasi synonyme du bien du plus grand nombre.

> Si j'ai raison, Stephen, cela signifie que la Machine dirige notre avenir, non seulement par des réponses directes à nos questions directes, mais en fonction de la situation mondiale et de la psychologie humaine dans leur ensemble. Elle sait ce qui peut nous rendre malheureux et blesser notre orgueil. La Machine ne peut pas, ne doit pas nous rendre malheureux. […] Stephen, comment pouvons-nous savoir ce que signifiera pour nous le bien suprême de l'humanité ? […] Seules les Machines le savent et c'est là qu'elles nous conduisent[111].

Nous retrouvons la transformation de la Première Loi en fonction du plus grand bien de tous dans « L'incident du tricentenaire ». Dans cette nouvelle, un robot sosie du président a tué le vrai président pour prendre sa place (sous l'influence et la direction d'un humain, il va de soi). Comment ce robot a-t-il pu violer la Première Loi sans en subir des conséquences usuelles, comme le court-circuit ou la stase ? L'explication donnée est la suivante :

> Même s'il n'a pas tué le président lui-même, la suppression d'une vie humaine d'une manière détournée est également interdite par la Première Loi, qui déclare : « Un robot ne peut nuire à un être humain ni laisser sans assistance un être humain en danger. » – La Première Loi n'est pas catégorique.

> Et si la perte d'un être humain sauvait la vie de deux autres êtres humains ou de trois ou même de trois milliards ? Le robot a peut-être considéré que la sauvegarde de la Fédération était plus importante que la sauvegarde d'une vie. Ce n'était pas un robot ordinaire, après tout. Il a été conçu pour reproduire les caractéristiques du président de façon à tromper tout le monde. S'il avait la capacité d'analyse du président Winkler sans avoir sa faiblesse, et s'il s'était rendu compte que lui, il pouvait sauver la Fédération, alors que le président en était incapable[112].

C'est dans *Les robots et l'empire*, que la Loi Zéro sera finalement formulée à cause des insuffisances des Trois Lois soulevées par R. Giskard Reventlov et R. Daneel Olivaw. Nous avons déjà souligné comment, dans la tradition de la Common Law, ce sont les insuffisances rencontrées sur le terrain qui amènent les juges à définir autrement les lois ou à présenter une nouvelle loi. La Loi Zéro est donc le fruit de l'insuffisance constatée par R. Daneel des Trois Lois précédentes dans le contexte, comme nous l'avons mentionné, où le bien-être de l'humanité est en cause. Pour faire émerger la Loi Zéro, il fallait donc à Asimov un scénario qui met en cause l'humanité. Quoi de mieux que le sort de l'Empire galactique comme décor pour situer l'action.

Dans « Les robots de l'aube », Asimov met en place l'enjeu du mouvement de la colonisation de l'espace par les Terriens. Ce projet est soutenu par un Aurorain, le D^r^ Fastolfe, qui possède les deux robots, Daneel et Giskard. Mais tous les Aurorains ne sont pas nécessairement d'accord avec le D^r^ Fastolfe, principalement son adversaire, le D^r^ Amadiro, qui a dans son équipe de chercheurs la fille de Fastolfe : la D^r^ Vasilia. Les robots Daneel et Giskard ont donc soutenu le projet de la colonisation du D^r^ Fastolfe ainsi que l'avoue Giskard à Baley : « Tu réprouvais les activités d'Amadiro, parce que tu es d'accord avec Fastolfe sur la colonisation de

la Galaxie ? Oui, Monsieur[113]. » Il faut un autre ingrédient dans l'histoire pour que les robots affrontent le problème de l'insuffisance des Trois Lois : il faut au robot la capacité de faire quelque chose par eux-mêmes (une autonomie d'action pouvant influencer les événements) et que cette capacité puisse mettre en cause la Première Loi. Asimov reprend ici des éléments exploités dans d'autres nouvelles : le robot télépathe qui est capable de voir ce que pensent les personnes et qui est capable d'influencer le cours de leurs pensées. Ce sont là les pouvoirs de Giskard par rapport à la Première Loi.

> Et pourquoi n'as-tu pas empêché Amadiro d'agir ? Pourquoi n'as-tu pas retiré de son esprit l'envie de sonder Jander ? Monsieur, répondit Giskard, je ne manipule pas les cerveaux à la légère. La résolution d'Amadiro était si profondément ancrée et complexe que j'aurais dû le manipuler profondément, et son cerveau est si intelligent, si avancé, que je ne voulais pas l'endommager[114].

Déjà dans ce roman, les pouvoirs de Giskard le mettaient constamment devant le problème d'évaluer le moindre mal : la réalisation du rêve de colonisation du D^r^ Fastolfe et la santé mentale du D^r^ Amadiro.

Dans *Les robots et l'empire*, les situations que vivent R. Daneel et R. Giskard les conduiront à l'énonciation de la Loi Zéro. Dans ce roman, la colonisation de la galaxie par les Terriens est une réussite. Sur Aurora par contre, le D^r^ Fastolfe est décédé, et il a légué à sa fille adoptive Gladia, la Solarienne, ses deux robots Giskard et Daneel. Le D^r^ Amadiro voit dans la colonisation un danger de plus en plus grand pour Aurora et il veut trouver un moyen de faire marche arrière avant qu'il ne soit trop tard. Deux personnes soutiendront le D^r^ Amadiro dans cette tentative : la D^r^ Vasilia (la fille du D^r^ Fastolfe) et le D^r^ Mandamus. Le projet est de rendre la Terre quasi inhabitable en augmentant le niveau de radiation. L'idée est simple : diviser pour régner. La Terre sert

de point d'union dans la colonisation et, sans la Terre, chaque colonie prendra son autonomie. La menace terrienne pour les Aurorains sera écartée. Comment Giskard et Daneel pourront-ils intervenir et surtout pour quelles raisons ?

L'émergence de la Loi Zéro et la reconnaissance de sa légitimité se cristallisent dans deux moments de l'histoire : le premier est l'affrontement des robots avec la Dr Vasilia, le second, l'affrontement des robots avec le Dr Amadiro et le Dr Mandamus à la fin du roman.

La Dr Vasilia vise à s'imposer comme maîtresse de Giskard à la place de Gladia et ainsi à pouvoir utiliser les capacités de Giskard pour favoriser les projets du Dr Amadiro. La Dr Vasilia croit pouvoir influencer Giskard en discutant avec lui sur le sens à donner aux lois. L'enjeu ici est d'établir qui a l'autorité légitime pour donner des ordres à Giskard. Dans ce qui va suivre, il ne faut pas oublier que Gladia est dans la pièce, endormie, et que Daneel a refusé de quitter la pièce, pour protéger éventuellement Gladia. Comment Vasilia peut-elle prétendre avoir l'autorité légitime pour donner des ordres à Giskard ? Ce qu'il faut noter, c'est que Giskard a été le robot de Miss Vasilia pendant dix ans et que ce sont les transformations au programme de Giskard, qu'elle a réalisées, qui lui ont donné ses facultés de télépathie. Lors de l'affrontement, la Dr Vasilia va tenter par divers moyens de faire dire à Giskard que la transformation qu'elle a apportée est telle que cela équivaut à l'avoir créé de nouveau, ce qui ferait d'elle sa propriétaire. Daneel se rend compte de la pression que subit Giskard et il décide d'intervenir en suggérant que Giskard pouvait bien faire oublier à Vasilia ses revendications. Voici comment la Dr Vasilia impose sa position à nouveau :

> Vraiment ? dit Vasilia en lui lançant un regard glacial. Mais, vois-tu, ce n'est pas à toi de décider qui Giskard considère comme sa maîtresse. Je sais que Giskard sait que c'est moi

> sa maîtresse. Et c'est donc à moi qu'il se doit entièrement, aux termes des Trois Lois. S'il doit contraindre *quelqu'un* à oublier et qu'il puisse le faire sans dommage physique, son choix devra se porter sur une autre personne que moi. Il ne peut me contraindre à oublier, ni agir sur mon esprit en aucune manière[115].

À la suite de l'ordre formel que Vasilia donne à Giskard d'imposer l'oubli à Daneel et à Gladia, Daneel essaie de nouveau d'argumenter pour dire à Giskard qu'il ferait moins de tort s'il imposait l'oubli à Vasilia. Après la réplique de Vasilia qui essaie de montrer que c'est elle et non Gladia qui subirait le plus de tort, Vasilia lui ordonne de se taire. Malgré l'ordre formel, Daneel essaie de parler.

> Je le peux, Madame. C'est difficile, mais je le peux, car je me rends compte que quelque chose l'emporte sur votre ordre, qui n'est régi que par la Deuxième Loi. – Silence, j'ai dit. Rien ne l'emporte sur mon ordre à l'exception de la Première Loi et j'ai montré que Giskard causera beaucoup moins de mal – pas du tout en fait – s'il me revient. C'est à *moi* qu'il fera du mal, à moi qu'il est le plus susceptible de nuire, s'il agit autrement. Elle pointa son doigt vers Daneel et répéta, d'une voix sifflante : Silence ! [...] Madame Vasilia, il existe quelque chose de plus fort même que la Première Loi. Giskard intervint, d'une voix tout aussi basse, mais sans effort. – Ami Daneel, il ne faut pas dire cela. Rien n'est plus fort que la Première Loi[116].

La D^r^ Vasilia va supprimer son ordre du silence afin de voir quand Daneel, en défiant les Trois Lois de la robotique, provoquera sa propre stase et ainsi s'autodétruira. C'est dans cet échange qu'il précise comment l'idée initiale a été lancée par Elijah Baley sur son lit de mort. Pour Baley, un individu est important, mais il ne faut pas oublier qu'il fait partie de quelque chose de plus grand ; un individu se compare donc à un fil dans une tapisserie. Chaque fil a son importance, mais cette importance est à l'intérieur d'un ensemble plus

grand que chaque fil : la tapisserie. L'analogie est claire, l'individu est à l'humanité ce qu'un fil est à la tapisserie. Sans humanité, l'individu perd son sens.

> Il existe une loi plus importante que la Première Loi : « Un robot ne peut nuire à l'humanité ni laisser sans assistance l'humanité en danger. » Je la considère maintenant comme la Loi Zéro de la robotique. La Première Loi devrait être formulée de la manière suivante : « Un robot ne peut nuire à un être humain, ni laisser sans assistance un être humain en danger, tant que cette assistance est compatible avec la Loi Zéro[117]. »

La critique que feront Vasilia et Giskard de la Loi Zéro met en évidence ce qu'il y a de plus difficile dans le raisonnement moral, assumer la responsabilité de sa décision sans avoir de preuve directe et certaine des impacts de l'action envisagée. Vasilia va donc ridiculiser la Loi Zéro parce qu'on ne peut pas prouver que telle action nuit à l'humanité :

> Les Trois Lois de la robotique concernent les êtres humains en tant qu'individus et les robots en tant que robots-individus. Il t'est possible de toucher du doigt un individu humain ou un individu-robot. Mais qu'est-ce que « l'humanité » sinon une abstraction ? Peux-tu toucher l'humanité ? Tu peux blesser ou ne pas blesser un être individuel et comprendre le préjudice ou l'absence de préjudice. Peux-tu voir un préjudice causé à l'humanité ? Peux-tu le comprendre ? Peux-tu le montrer du doigt ? […] – Non Madame, je ne le peux. Mais je pense que cette blessure peut exister malgré elle, et vous voyez que je tiens toujours debout[118].

Et qu'en pense Giskard, qui, comme on l'a vu, n'est pas prêt à accepter la Loi Zéro ? On retrouve ici la critique marxiste des morales humaines, des abstractions qui servent non pas à servir l'humain, mais à l'assujettir.

> Je ne peux accepter la Loi Zéro, ami Daneel, dit doucement Giskard. Tu sais que j'ai beaucoup lu l'histoire de l'humanité.

> J'y ai trouvé de grands crimes commis par des êtres humains contre d'autres êtres humains et toujours on a donné pour excuse que les crimes étaient justifiés par les exigences de la tribu, de l'État, ou même de l'humanité. C'est précisément parce que l'humanité est une abstraction qu'on peut si aisément en appeler à elle pour justifier tout et n'importe quoi, et ta Loi Zéro est en conséquence inadéquate[119].

Comment Daneel peut-il montrer qu'il ne s'agit pas d'une pure abstraction, mais bien de quelque chose de concret qui permet de saisir comment l'action concrète ayant un impact sur des individus affecte aussi la toile de l'humanité ? Voici comment il présente à Giskard le problème qu'il doit affronter : « Mais tu sais, ami Giskard, qu'un danger existe maintenant pour l'humanité et qu'il va se concrétiser si tu deviens la propriété de M^me^ Vasilia. Cela n'est pas une abstraction[120]. » C'est en montrant le lien entre l'acte, ici que Giskard devienne propriété de M^me^ Vasilia, et les conséquences, comment Giskard sera utilisé pour le projet du D^r^ Amadiro, que la menace va se concrétiser, c'est-à-dire passer du possible au réel. Or, la réplique de Giskard clarifie son malaise avec la Loi Zéro : « Le danger dont tu parles ne constitue pas une certitude, mais découle d'une déduction. Et l'on ne peut fonder nos actions au mépris des Trois Lois pour autant[121]. » Autrement dit, le robot Giskard, tout comme la D^r^ Vasilia, a besoin de certitude pour agir. La responsabilité morale, par contre pour Daneel, consiste justement à décider sur fond d'incertitude.

Dans la scène finale du roman, le D^r^ Amadiro ainsi que le D^r^ Mandamus sont sur la terre dans un endroit à forte puissance radioactive et ils mettent en place le dispositif final pour irradier la terre. Le projet du D^r^ Mandamus est de favoriser une irradiation graduelle qui se fera sur quinze décennies. Cela donne à la terre assez de temps pour réagir et coloniser ailleurs et aussi assez de temps pour penser que

le processus est naturel et qu'il n'a pas été provoqué par les Aurorains (ce qui déclencherait une guerre). Lorsque la terre ne pourra plus unir les Terriens puisqu'ils seront dispersés dans les colonies, alors Aurora pourra monter sa suprématie. Cependant, le D^r^ Amadiro n'a pas quinze décennies encore à vivre et il voudrait irradier plus rapidement la terre, peu importe les conséquences sur les humains qui sont envisagées à quelques milliards de morts. Pendant qu'Amadiro et Mandamus se disputent encore avant de régler les cadrans, Daneel et Giskard arrivent sur les lieux. Lors de l'affrontement, après différentes ruses et quelques mensonges utilisés par les deux docteurs, les masques tombent et Amadiro avoue le mal qu'il veut faire aux Terriens. Du même souffle, il ordonne ceci :

> Robots, nous sommes des Spaciens. Bien plus, nous sommes des Aurorains, du monde où vous avez été construits. Bien plus, nous sommes d'importants personnages d'Aurora et vous devez interpréter l'expression « êtres humains » des Trois Lois de la robotique comme signifiant « Aurorains ». Si vous ne nous obéissez pas immédiatement, vous nous faites du mal et vous nous humiliez, de sorte que vous violez les Première et Deuxième Lois. Il est exact que notre action, ici, a pour but de détruire des Terriens, un grand nombre de Terriens même. Mais cela n'a absolument rien à voir. Vous pourriez tout aussi bien refuser de nous obéir parce que nous mangeons la viande d'animaux que nous avons tués. Maintenant que je vous ai expliqué cela, partez ! Mais ces derniers mots furent dits d'une voix rauque. Amadiro, les yeux exorbités, s'écroula[122].

Daneel explique à Mandamus qu'après l'expérience des Solariens, qui avaient limité la définition d'être humain aux seuls Solariens et les conséquences qui s'en suivirent, ils arrivèrent à la conclusion que les êtres humains sont ceux qui appartiennent à l'espèce *Homo sapiens*. Appliquant la

Loi Zéro, ils ne pouvaient donc pas laisser Amadiro tuer les Terriens.

L'application de la Loi Zéro a justifié que Giskard impose l'oubli au Dr Amadiro. Mais, ce qui surprend dans la fin de l'histoire c'est que Giskard va aussi figer momentanément Daneel et laisser le Dr Mandamus amorcer le processus d'irradiation. Mais pourquoi ? Qu'est-ce qui peut justifier cette décision ? Le problème que soulève Giskard avec la Loi Zéro c'est de savoir comment l'appliquer, surtout que pour l'appliquer on est loin de la certitude. Concrètement, la question fondamentale est de savoir si l'irradiation de la terre sur quinze décennies est une bonne ou une mauvaise chose pour l'humanité. Voici comment Giskard explique son geste :

> D'après la nature du triomphe dans son esprit, j'ai bien l'impression qu'il croyait que l'augmentation de la radioactivité entraînerait l'anarchie et la confusion parmi les Terriens et les Coloniens, et que les Spaciens allaient les détruire et s'emparer de la galaxie. Mais j'ai pensé que le scénario qu'il nous proposait pour nous convaincre était le bon. La disparition de la terre en tant qu'immense monde très peuplé entraînera la disparition d'une mystique dont j'ai déjà senti qu'elle était dangereuse et les Coloniens en bénéficieront. Ils vont se répandre dans la galaxie de plus en plus vite – sans la terre vers laquelle ils auraient dû se retourner sans cesse comme vers un dieu du passé –, ils vont fonder un empire galactique. Il faut que nous aidions à réaliser tout cela[123].

Mais pourquoi, après une telle déclaration, est-ce que Giskard devient de plus en plus faible jusqu'à devenir muet après avoir prodigué les derniers conseils à Daneel ? Qu'est-ce que le cerveau postronique de Giskard n'a pu tolérer ? Quelle était la cause de ce stress ? Avant de s'éteindre, Daneel lui pose la question : pourquoi, alors que tu as si bien respecté la Loi Zéro, cela se produit-il ? Les dernières paroles de

Giskard nous livrent ce qu'il ne pouvait supporter : « Parce que je n'en suis pas certain. Et si le Dr Mandamus… avait raison… après tout… et si les Spaciens triomphaient… Adieu, ami Dan…[124] » Quelle responsabilité que d'influencer l'avenir de l'humanité sans savoir avec certitude la justesse de nos choix !

CONCLUSION : LES LEÇONS D'ASIMOV POUR UNE MORALE DE NOTRE TEMPS

La bible des robots d'Asimov nous invite, à travers les récits, à réfléchir à une morale pour notre temps. Trois points m'apparaissent importants à souligner dans son œuvre touchant les insuffisances des Trois Lois morales et la place de la morale dans nos sociétés actuelles.

Dans un premier temps, Asimov nous invite à dépasser les Trois Lois morales en les inscrivant dans une approche éthique. La morale, comme Asimov nous l'a montré amplement, est axée sur une logique de commandement et ces commandements entrent en conflit lorsque nous sommes dans un dilemme moral. Évidemment, la solution pour des robots est de programmer les Lois par ordre de priorité : la Première, la Deuxième et la Troisième. Les humains n'ont pas cette programmation, donc ils doivent trancher la priorité à donner à une obligation sur une autre. Comment est-ce qu'un être humain arrive à décider d'accorder dans ce contexte plus d'importance au commandement de sa préservation qu'à celui d'apporter une assistance à autrui ? Asimov a donné de multiples exemples où les robots étaient aux prises avec des argumentations différentes qui visaient à les faire agir d'une certaine façon plutôt qu'une autre tout en respectant les Trois Lois. Le but était de décider qui est le Maître et c'est lui qui dicterait alors les priorités. Dans l'application des Trois Lois de la robototique, les robots doivent non seulement comprendre l'ordre donné,

mais encore, pour les robots les plus sophistiqués, en comprendre le sens afin de pouvoir l'appliquer.

Asimov nous invite ainsi à dépasser l'application mécanique de la loi pour en comprendre le sens. Les Lois morales ont pour fonction de nous aider à vivre ensemble en société et à faire des choix individuels et collectifs qui, au moins minimalement, ne nuiront pas gravement aux autres et, si possible, à assurer le plus grand bien à tous en contexte. C'est en situant les Lois morales en fonction des valeurs que nous voulons atteindre dans nos vies individuelles et collectives que la pensée d'Asimov rejoint tout le courant de l'éthique appliquée qui s'est développé depuis plus de cinquante ans. Ce courant, qui prend racine dans la réflexion sur la responsabilité sociale des chercheurs à la suite de la création de la bombe nucléaire et des divers scandales en recherche sur les humains, pose les bases du développement responsable des technologies issues de la recherche. Asimov, contrairement à d'autres romans de science-fiction, ne nous fait pas la morale. Il n'a pas de solutions toutes faites. Il sait que nous sommes toujours en tension avec des valeurs que nous aimerions vivre totalement, mais qui sont irréconciliables. Dans la nouvelle « La vie et les œuvres de Multivac », Asimov illustre bien ce conflit permanent entre deux valeurs qui traversent plusieurs de nos choix individuels et collectifs : celui entre la sécurité et l'autonomie (liberté). Quels sont les bienfaits de Multivac et pourquoi certains veulent-ils sa perte ? « Ne nous répète pas encore que l'on ne peut pas se passer de Multivac, que ce n'est pas la peine de lutter, que nous avons la sécurité. Ce que tu appelles la sécurité, nous tous nous l'appelons esclavage[125]. » Or Ron réussit à détruire Multivac et, parlant aux membres du groupe, il dit « Vous parliez de liberté, vous l'avez maintenant ! » Puis d'une voix mal assurée : « N'est pas cela que vous vouliez ?[126] » Jusqu'où accepter de limiter notre liberté pour notre sécurité ?

Jusqu'où accorder la liberté en sachant que des personnes se blesseront ou en blesseront d'autres ? Il n'y a pas une solution à ces questions, mais des choix responsables en contexte.

La deuxième leçon que je tire d'Asimov concerne les insuffisances des Trois Lois et la création de la Loi Zéro. Ici encore Asimov reflète bien les préoccupations des penseurs sur le développement technologique et sur les limites des Lois morales pour y répondre. Hans Jonas est un philosophe allemand (1903-1993) qui a mis en évidence les lacunes des morales individuelles lorsqu'il s'agit d'enjeux moraux collectifs. Tout l'affrontement des robots avec la Dr Vasilia reprend essentiellement les arguments de Hans Jonas. Nos morales des Trois Lois concernent les individus et l'impact de nos actions sur les individus autour de nous. Mais les Trois Lois ne peuvent pas tenir compte des actions qui nuisent ou nuiront à l'humanité. La Loi Zéro d'Asimov est l'équivalent du principe de responsabilité de Hans Jonas. Nous sommes responsables de la vie sur terre et du sort des générations futures. Il faut donc dans nos décisions tenir compte de répercussions sur les humains et aussi sur l'humanité à venir.

Le message d'Asimov est assez pessimiste sur les capacités des êtres humains à tenir compte des conséquences du développement technologique sur l'humanité avant de s'aventurer dans une nouvelle entreprise. Pour éviter que le développement technologique se termine en « arme de mort », Asimov ne fait confiance dans ses romans qu'à Multivac, aux Machines et finalement à R. Daneel et R. Giskard. Pourquoi les humains ne sont-ils pas capables de dépasser leurs conflits pour assurer le plus grand bien possible de tous les « êtres humains » à travers le développement technologique ? N'y a-t-il pas parmi les comités éthiques nationaux et internationaux qui s'interrogent sur

le développement technologique, et notamment sur les enjeux éthiques, légaux et sociaux, un nouvel espace pour penser une éthique des choix sociaux ? Sommes-nous à jamais condamnés à répéter notre histoire ?

À plusieurs reprises, les personnages d'Asimov ont souligné l'écart entre la morale des robots et la morale des êtres humains. Si cet écart est soulevé habituellement avec une pointe d'ironie, Asimov ne développe pas les raisons de cet écart. Pourtant il considère que le jugement éthique est l'activité la plus complexe à réaliser pour un cerveau positronique. Or plusieurs des nouvelles nous ont effectivement montré la complexité du raisonnement moral. On ne peut s'empêcher de penser que, si ce raisonnement est si complexe pour le robot, il l'est encore plus pour l'être humain qui n'a pas un programme initial. Dans les choix difficiles, Asimov met en dialogue soit des robots avec les humains, soit deux robots entre eux afin d'explorer les diverses possibilités du meilleur choix. Mais que faisons-nous pour favoriser le raisonnement pratique des êtres humains ? Comment développons-nous, actuellement, le raisonnement moral pratique chez les enfants, chez les adultes dans leur travail professionnel ou dans les organisations ? Dans la foulée du développement de l'éthique appliquée, plusieurs organisations ont créé des comités de réflexion éthique ou d'aide à la décision afin de favoriser la prise de décision responsable dans les choix professionnels et dans les choix organisationnels. L'établissement d'une commission d'éthique de la science et de la technologie au Québec va dans le même sens, cette fois pour réfléchir en dialogue sur les meilleurs choix sociaux. Cependant, dans nos systèmes éducatifs québécois, canadiens et états-uniens, nous avons tendance à suivre l'approche des morales traditionnelles en inculquant aux jeunes les lois morales issues de religions diverses ou de systèmes philosophiques. Souvent, dans ces enseignements

dans les écoles publiques, on insiste sur les lois morales et leurs fondements respectifs, ce qui favorise certes une meilleure connaissance des différences multiculturelles, mais ce qui est loin de former le jugement moral pratique des jeunes. Comment les adultes de demain pourront-ils faire des choix responsables, si, dans nos sociétés, on ne se préoccupe pas aujourd'hui de former les jeunes à la complexité du raisonnement pratique qu'est la décision ?

3
Réaliser des robots éthiques
Limites scientifiques, défis technologiques et potentiel de la robotique et de l'intelligence artificielle

Jacques Beauvais
Professeur en génie électrique,
Université de Sherbrooke
Jonathan Genest
Professionnel de recherche,
Université de Sherbrooke

1. INTRODUCTION

L'univers d'Asimov est fascinant. En plus de nous proposer des aventures originales et des personnages colorés, il illustre les impacts d'une société fondée sur une technologie de rupture, comme celle des robots. Ce qui est encore plus intéressant, c'est qu'Asimov se sert de ce prétexte pour imaginer avec détails la mise en place de Lois morales précises dans une technologie. Dans notre monde actuel, l'offre technologique croît chaque jour : l'informatique mobile devient de plus en plus omniprésente, notre compréhension de la chimie moléculaire et de la génétique permet de produire des médicaments plus ciblés et plus efficaces et, avec l'incorporation de capteurs, les bâtiments et les infrastructures deviennent de véritables instruments de mesure. Compte tenu de l'omniprésence de la technologie et ses

impacts sur les humains, serait-il possible d'inculquer des concepts moraux ou éthiques aux technologies qui nous entourent ? Cette question est beaucoup trop complexe et trop vaste pour pouvoir être analysée dans un seul livre. On peut toutefois s'interroger sur la technologie choisie par Asimov : la robotique. Disposons-nous, aujourd'hui, de moyens technologiques pour créer des robots intelligents capables de prendre des décisions en se basant sur des principes moraux comme ceux que nous présente Asimov ? Est-ce que des robots comme Daneel et Giskard sont possibles ? Dans ce chapitre, nous tenterons d'évaluer la faisabilité de mise en œuvre des Trois Lois de la robotique et de la réalisation de robots éthiques.

Dans un premier temps, nous essaierons de comprendre ce qui limite la mise en œuvre des Trois Lois de la robotique dans notre univers réel. Après avoir fait un bref état des lieux sur les robots que nous produisons, nous aborderons la question des prémisses de l'œuvre d'Asimov afin de mieux comprendre les différences avec notre monde et d'évaluer les impacts sur les Trois Lois de la robotique. Dans un deuxième temps, nous étudierons les avancées actuelles en robotique et en intelligence artificielle afin de voir s'il serait possible de réaliser des robots éthiques permettant de respecter l'esprit des Lois de la robotique proposées par Isaac Asimov.

2. LES ROBOTS ASIMOVIENS DANS NOTRE MONDE RÉEL

2.1 L'utilisation des robots aujourd'hui

Avant de s'interroger sur la possibilité de mettre en œuvre les Trois Lois de la robotique, il serait pertinent de valider si l'on veut effectivement inculquer les Trois Lois de la robotique d'Asimov aux robots que nous fabriquons déjà, à ceux que nous développons présentement et à ceux qui feront l'objet des percées scientifiques et des avancées technologiques d'ici quelques années. Est-ce que notre utilisation

des robots est compatible avec un système de Lois morales ? À l'heure actuelle, l'utilisation des robots est très répandue dans deux domaines d'applications très précis, soit le secteur industriel et le secteur militaire. Les types de robots que l'on retrouve dans ces secteurs sont extrêmement différents.

2.1.1 Les robots industriels

Dans le secteur manufacturier, la plupart des robots que l'on utilise sont des machines fixes, incapables de se déplacer physiquement, mais ayant une portée suffisante pour effectuer un ensemble de tâches pertinentes à l'assemblage et au traitement d'un produit. Que ce soit dans l'assemblage des automobiles, dans la soudure, la découpe de pièces de métal, la manipulation des matériaux ou d'autres applications, ces robots sont souvent ancrés par un pivot fixe avec un bras articulé au bout duquel se trouve un actionneur spécialisé pour une fonction précise, que ce soit un manipulateur, une torche à soudure ou une autre élément actif. Ces robots sont conçus pour fournir un ensemble de bénéfices tels que de la vitesse, de la répétitivité ou une efficacité accrue qui donne un avantage compétitif à l'entreprise manufacturière propriétaire qui les utilise par rapport à un compétiteur qui emploie plutôt une main-d'œuvre plus traditionnelle. Au-delà des impacts souvent négatifs sur le taux d'emploi traditionnel dans le secteur manufacturier et habituellement positifs sur la productivité des entreprises, ces robots permettent d'accomplir de manière sécuritaire des tâches parfois dangereuses et certainement monotones. Dans la vaste majorité de ces cas, leur champ d'action demeure toutefois physiquement très limité. Une fois que l'on est assuré qu'aucune personne ne peut entrer dans la zone délimitant la portée des mouvements du bras robotisé pendant qu'il est actif, il n'est plus nécessaire de s'inquiéter de l'implémentation de la Première Loi de la robotique. Ces robots sont incapables d'agir en dehors de cette zone. Ils

sont programmés pour des tâches précises et leurs actions limitées permettent aux concepteurs d'anticiper et de programmer la très grande majorité des possibilités d'événements qui pourraient s'avérer dangereuses pour l'humain. Un ensemble de systèmes de sécurité peut également être intégré à la plateforme robotisée, dont des capteurs de mouvement, afin d'éviter les chocs et les collisions avec leur environnement. On souhaitera également programmer le robot afin d'éviter qu'il ne puisse s'endommager et l'obliger à répondre aux instructions du donneur d'ordres. Mais, étant donné le champ d'action extrêmement limité de ces robots, nous sommes bien loin d'un exemple d'application des Troisième et Deuxième Lois d'Asimov, malgré le fait que le résultat global soit sensiblement le même.

2.1.2 Les robots militaires

La seconde catégorie de robots en grand déploiement à l'heure actuelle concerne le domaine militaire. Alors qu'en 2003 les forces américaines n'étaient accompagnées d'aucun robot lorsqu'elles ont progressé du Koweït à Bagdad, aujourd'hui les forces militaires américaines détiennent un inventaire de plusieurs dizaines de milliers de robots, pour les opérations tant terrestres qu'aériennes[1]. Bien qu'ils soient originalement conçus en tant que plateformes mobiles de surveillance terrestre et aérienne téléguidées, deux développements notables en font des objets de grande pertinence pour la présente discussion. D'abord, ces robots ont acquis au fil des années des capacités semi-autonomes : plusieurs des robots téléguidés aériens possèdent maintenant la capacité d'effectuer des décollages et des atterrissages de manière semi-autonome, sans être téléguidés. De plus en plus de ces systèmes robotisés sont équipés d'armements de tous genres, transformant donc ces systèmes en cuirassés et bombardiers semi-autonomes. Par conséquent, la fonction primordiale pour laquelle ils ont été conçus place

ces robots dans des situations qui sont en complète contradiction avec la Première Loi d'Asimov. Il en demeure qu'il y a certainement lieu de se questionner sur l'importance d'intégrer des systèmes décisionnels éthiques à bord de ces robots pour un ensemble de raisons : leur mobilité et l'environnement non contrôlé dans lequel ils seront utilisés rendent à toutes fins utiles impossible de prévoir toutes les situations qu'ils rencontreront. Leur comportement semi-autonome a pour conséquence que des décisions devront être prises très rapidement, durant des moments où ils ne seront pas sous un contrôle à distance. De plus, comme ils sont équipés d'armements divers, on peut très facilement anticiper qu'ils se retrouveront dans des situations où ces décisions, de portée éthique, devront être prises très rapidement.

2.1.3 Les robots éthiques

Plusieurs chercheurs se sont investis dans l'étude de « l'éthique des machines », notamment Michael et Susan Leigh Anderson de l'Université du Connecticut[2], ainsi que James H. Moor du Dartmouth College[3]. C'est un sujet de vif débat entre les éthiciens et les informaticiens et, pour le moment, cela demeure une question d'ordre philosophique plutôt que matérielle. Mais l'accélération du développement des technologies robotisées, notamment militaires, et la création d'applications telles que l'utilisation de robots aidants dans les centres pour les aînés sont des exemples concrets qui poussent la réflexion à ce sujet.

Au-delà de ces remarques introductives qui apportent un éclairage sur l'importance de se pencher sur cette question, nous ne nous attarderons pas dans ce chapitre à la question de pertinence de l'implémentation des Lois éthiques dans les robots, mais nous nous concentrerons sur la question de notre capacité technologique d'implémenter les Trois Lois. Est-il même possible de les intégrer dans les

technologies robotiques ? Si non, quelles sont les options technologiques qui sont à notre portée dans l'immédiat pour nous rapprocher de ce but, à l'intérieur du cadre de nos connaissances et avec toutes les contraintes du monde réel par rapport aux contraintes allégées d'une œuvre fictive ?

En premier lieu, il sera fort pertinent d'examiner ces différences entre le monde de l'œuvre d'Asimov, à l'intérieur duquel il est certainement possible et à propos d'appliquer les Trois Lois de la robotique à tous les robots fabriqués au fil des siècles couverts par le cycle des robots, et notre monde où les contraintes de la réalité soulèvent des interrogations sérieuses quant à la faisabilité et même quant au sens des Trois Lois.

2.2 Acceptation des prémisses d'Asimov

2.2.1 L'univers fictif

Chaque œuvre de science fiction a son *novum*[4], soit l'ensemble des innovations scientifiques et technologiques fictives mais plausibles qui soutiennent le narratif. De prime abord, notre lecture d'une œuvre particulière tend à s'attarder plus souvent et de manière plus précise sur les différences technologiques, sociales et culturelles par rapport au monde qui nous entoure, alors que certaines hypothèses beaucoup plus subtiles et moins évidentes, mais parfois beaucoup plus importantes pour l'existence même de ce monde, ne captent pas du tout notre attention. Bien entendu, l'auteur de l'œuvre escamotera souvent ces hypothèses qui auront été présentées de manière tout à fait plausible, ou encore qui n'auront pas été discutées du tout, alors qu'elles sont souvent complètement invraisemblables et carrément impossibles. Deux exemples dans des œuvres modernes, mais déjà devenues classiques, suffiront pour illustrer cette situation.

2.2.1.1 Le novum du Parc jurassique

Dans son roman *Le Parc jurassique*[5], Michael Crichton présente une double critique de la science : d'abord, une critique sévère des scientifiques qui se croient capables de tout contrôler. Une grande partie du narratif est consacrée à l'impact du chaos et de l'imprévisibilité des événements qui remettent en question ce contrôle illusoire. Une seconde critique importante vise la science au service du profit économique, alors que les percées scientifiques de l'entreprise InGen sont exploitées par le propriétaire du parc dans le but de générer des profits qui pourraient être immenses (ce volet critique est moins présent dans le film de Steven Spielberg où John Hammond est plutôt présenté et joué comme un gentilhomme qui est d'abord et avant tout un bon grand-père). Le *novum* de cette œuvre, et du second roman qui a suivi par la suite, est basé sur une technologie d'ingénierie génétique mise au point par les scientifiques de la compagnie fictive InGen qui ont réussi à extraire de l'ADN viable d'un dinosaure à partir d'un moustique encastré dans de l'ambre. Or, bien que cela soit plausible dans l'œuvre de Crichton, il n'y a aucune base scientifique qui permette d'imaginer que ce soit possible. Aucun ADN viable ne peut se retrouver dans des fossiles de plusieurs millions d'années, alors que le matériel biologique est minéralisé. Les seuls cas de matériel biologique préhistorique ayant été préservés et retrouvés concernent des tissus biologiques congelés, momifiés ou squelettiques. La dégradation de ces spécimens dépend de la température, de l'humidité et du passage du temps. Dans les conditions absolument optimales, par exemple dans la glace ou le pergélisol, on estime la limite absolue de préservation de l'ADN à environ un million d'années. Cela nous place très très loin de la période minimale de 65 millions d'années qui serait requise pour obtenir de l'ADN de dinosaures. Mais le roman, aussi bien

que le film, encadre la technologie avec du jargon et un contexte technologique qui rendent plausible toute l'existence du Parc jurassique.

2.2.1.2 Le novum de Star Trek

Une seconde œuvre permettant d'illustrer un *novum* pleinement développé tire initialement ses racines à la télévision, au cinéma et dans des romans. Il s'agit de la série *Star Trek* créée par Gene Roddenberry qui originalement véhiculait une critique sociale sévère et mettait en valeur une société avec une ouverture sociale impensable dans les années 1960 au moment de la première série télévisée[6]. Des technologies fort intéressantes attirent rapidement notre attention : qu'il s'agisse des communicateurs, de la téléportation ou encore du tricordeur qui permet d'analyser rapidement et sans intrusion l'état de santé d'une personne. Un peu plus de quarante ans après le lancement de cette série, la technologie disponible au quotidien dépasse dans plusieurs cas ce qui était présenté comme la technologie courante du XXIII^{e} siècle. On n'a qu'à penser à nos ordinateurs personnels qui mettent à portée de la main de chaque personne plus de puissance que ce qui se retrouvait à bord d'un vaisseau intersidéral ou à nos téléphones intelligents qui dépassent largement la capacité des communicateurs utilisés par le capitaine Kirk et son équipage. La téléportation est même possible aujourd'hui, quoi qu'il s'agisse d'une expérience scientifique de pointe qui est d'une portée infiniment plus restreinte que la téléportation des membres de l'équipage. Mais tout ce technobabillage, qui dans l'ensemble est fort plausible même s'il est technologiquement très avancé, constitue une diversion afin de rendre plus crédible le véritable cœur du *novum*, l'élément qui demeure absolument impossible et incontournable dans notre réalité : il n'y a aucune possibilité de voyager plus rapidement que la vitesse de la lumière. Avec cet incontournable, il n'y a pas de système

d'hyperpropulsion tel que celui que l'on retrouve à bord du vaisseau *Enterprise*, et sans hyperpropulsion, il n'y a pas de Fédération, ni d'empires Klingon ou Romulans. En fait, il n'y a pas de Star Trek.

2.2.2 Le novum d'Asimov

Qu'en est-il de l'œuvre d'Asimov ? Des différences fondamentales par rapport à notre réalité, parfois subtiles et légères, ont un impact très important sur la faisabilité et le réalisme de la mise en œuvre des Trois Lois de la robotique. Il est intéressant de s'attarder dans un premier temps à la raison d'être de la production de robots humanoïdes.

2.2.2.1 Les robots humanoïdes

Un premier élément déterminant du *novum* d'Asimov est de nature économique. Dans *Les cavernes d'acier*[7], le détective Elijah Baley s'entretient avec un spécialiste de la robotique terrien, un éminent savant dans le domaine, D[r] Anthony Gerrigel. Après avoir décrit les énormes efforts requis pour concevoir un cerveau positronique, le D[r] Gerrigel explique à Baley les raisons pour lesquelles on construit des robots.

> Supposez que vous ayez à exploiter une ferme : auriez-vous envie d'acheter un tracteur à cerveau positronique, une herse, une moissonneuse, un semoir, une machine à traire, une automobile, etc., tous ces engins étant également dotés d'un cerveau positronique ? Ou bien ne préféreriez-vous pas avoir du matériel sans cerveau, et le faire manœuvrer par un seul robot positronique ? Je dois vous prévenir que la seconde solution représente une dépense cinquante ou cent fois moins grande que la première.

Cette question sous-tend l'économie entière de plusieurs planètes des Spaciens. Or, une bonne partie du cycle des robots fut écrite par Asimov avant l'invention du circuit intégré microélectronique. *Les cavernes d'acier*, publié en 1953, était contemporain aux travaux pionniers de

réalisation des premières puces de microélectronique et, à cette époque, nul ne pouvait imaginer l'impact socio-économique de l'électronique intégrée sous forme de puce, surtout dans sa forme la plus perfectionnée, représentée à l'heure actuelle par les microprocesseurs. L'industrie et l'économie ont été complètement transformées par la microélectronique et la tendance que l'on observe depuis de nombreuses années est l'utilisation de puces conçues spécialement pour des applications très précises, comme celles que l'on retrouve aujourd'hui dans les usines, les automobiles, les radios-réveils, les téléviseurs, etc. De plus, le développement des nanotechnologies et de l'ingénierie des matériaux à une échelle presque atomique au cours des dernières décennies, avec toute la flexibilité que ces techniques nous permettent, ne fait que poursuivre cette tendance et accroître notre capacité de concevoir des dispositifs sur mesure pour chaque application et à très faible coût. Économiquement, la réalité que nous connaissons est tout à fait inversée par rapport à celle que le D[r] Gerrigel décrit à Baley. Le gain économique de cette approche favorisant l'utilisation d'une puce comme microcontrôleur conçue pour chaque application particulière, plutôt que la fabrication de robots généralistes, est en fait plusieurs fois supérieur au facteur de cinquante à cent mentionné par Gerrigel. Bien qu'il faille reconnaître qu'il aurait fallu qu'Asimov soit devin pour pouvoir anticiper l'impact des puces microélectroniques, il en demeure que son argumentaire stipulant qu'un système hautement complexe, tel un robot humanoïde, est la solution économique optimale pour contrôler des machines est quelque peu étonnante. Le coût associé à tout bris d'un système de la complexité d'un robot est astronomique et les risques de bris sont d'autant plus élevés que le système est complexe. Même à l'époque de la création de cette œuvre, argumenter qu'un cerveau positronique soit

nécessaire pour traire une vache ou répandre des semences était absurde. Il s'agit certainement d'un choix conscient de la part d'Asimov visant à justifier la prévalence des robots humanoïdes sur les planètes spaciennes, prévalence qui fait partie du *novum* de son œuvre et qui ne cadre pas du tout avec la réalité des principes économiques raisonnables que nous connaissons.

2.2.2.2 Le déterminisme

Une autre caractéristique absolument fondamentale du *novum* d'Asimov est beaucoup plus subtile : toute l'œuvre du cycle des robots se déroule dans un univers déterministe au sens strict du terme. Quelques exemples permettront d'en tracer un portrait et de tirer quelques conclusions.

D'abord, un élément répété à plusieurs reprises est que les mathématiciens, avec des efforts considérables, sont en mesure de calculer l'ensemble des parcours positroniques pour un cerveau de robot. On peut déduire que la complexité de ces parcours est énorme, simplement par l'effort requis pour les concevoir et par le résultat que les robots possèdent une immense capacité de raisonnement, d'emmagasinement et d'analyse de données. Ce qui étonne, c'est que ces calculs mathématiques sont tellement détaillés et déterminent avec une telle efficacité tous les traits des robots que, dans la nouvelle « Le petit robot perdu[8] », alors que Susan Calvin tente de mesurer le temps de réaction de 63 robots face à un danger imminent pour un humain, ces 63 robots possèdent exactement le même temps de réaction à la fraction de seconde près, au point tel qu'on ne peut les différencier. Dans cet exemple, les calculs mathématiques ont donc fixé toutes les caractéristiques des robots pour la durée de leur opération ou de leur vie. Face à une situation arbitraire quelconque, telle que le test du temps de réaction lorsqu'ils se trouvent dans une situation d'humain en danger, aucune différence de comportement des robots, même à une échelle d'une

fraction de seconde, ne peut être observée. Même à l'époque où Asimov a écrit ses premières nouvelles du cycle des robots, des mesures de temps extrêmement précises étaient disponibles, la première horloge atomique ayant été développée en 1949[9]. On peut donc conclure que ce temps de réaction à la fraction de seconde près, tel qu'il est décrit dans « Le petit robot perdu », dénote une très grande précision, avec la même signification que nous y accorderions aujourd'hui. Tout ça découle de calculs extrêmement complexes effectués par des experts, et en dépit de toutes les plus petites différences de fabrication qui résultent de toute production en milieu manufacturier, des changements provoqués au fil du temps par l'environnement, par exemple l'exposition aux radiations, et sans compter l'impact de la mécanique quantique qui doit prévaloir au niveau le plus fondamental dans le cerveau positronique.

Dans la nouvelle « Raison[10] », les deux ingénieurs Powell et Donovan tentent de résoudre une situation problématique avec un robot qui n'accepte pas d'avoir été fabriqué par les humains. Il s'acquitte néanmoins de sa tâche avec succès, tel que le déclare Powell au robot Cutie : « Tu as maintenu le tracé d'ondes énergétiques avec une précision absolue sur la station réceptrice. » Cette « précision absolue » du robot Cutie à positionner le faisceau énergétique, et ce malgré une tempête d'électrons survenue en plein parcours du faisceau, est une tâche impossible dans notre monde réel. L'impact direct de la mécanique quantique, qui domine la physique à l'échelle microscopique dans notre réalité, fait en sorte qu'on ne peut contrôler avec une précision absolue à la fois la position et la vitesse d'une particule ou d'un faisceau énergétique. Il est d'ailleurs impossible de maintenir un faisceau focalisé sur des distances interplanétaires, et l'imprécision causée par le passage au travers d'une tempête d'électrons

ne peut qu'empirer davantage la situation. Il s'agit donc ici d'un écart très marqué entre l'œuvre d'Asimov et la réalité.

Dans *Les cavernes d'acier*[11], Elijah Baley est pourchassé et tente de s'échapper au moyen des tapis roulants. Dans ce roman, les citoyens se déplacent à travers la ville en empruntant ces tapis roulants, certains étant accélérateurs et d'autres permettant de réduire la vitesse, en sautant d'un tapis à l'autre avec la possibilité de se rendre jusqu'au tapis express. On apprend également que les jeunes s'adonnent à un jeu de poursuite sur les tapis roulants.

> Le meneur part avec une légère avance, sur un tapis roulant accélérateur ; il fait de son mieux pour agir de la façon la plus inattendue, et reste par exemple très longtemps sur le même tapis, avant de bondir sur un autre, dans une direction différente ; il passe alors très vite d'un tapis au tapis suivant, puis s'arrête tout d'un coup. Malheur au poursuivant qui se laisse imprudemment entraîner trop loin ! Avant de s'être aperçu de son erreur, il se trouvera, à moins d'être extrêmement habile, bien au-delà du meneur ou, au contraire, très en deçà. Le meneur, s'il est intelligent, en profitera aussitôt pour filer dans une autre direction.

Il s'agit d'un jeu extrêmement sensible aux plus fines réactions des participants, avec un résultat qui, du regard de quelqu'un de l'externe, est complètement instable et chaotique puisque les accélérations et les grandes vitesses des tapis roulants amplifient dramatiquement les plus petits écarts de réaction des coureurs. Pourtant, après une course folle où Baley, très expérimenté à ce jeu, fait tout son possible pour réagir de manière complètement spontanée et imprévisible, R. Daneel Olivaw, le robot humanoïde, se retrouve à ses côtés suffisamment près pour le saisir alors qu'il trébuche. Bien que Daneel possède la capacité de cérébro-analyse, lui permettant de lire le contenu émotif des pensées des humains, il n'est pas pour autant télépathe et ne peut

donc pas lire les pensées de Baley afin de le suivre. Cela implique que, par une simple observation extrêmement précise des moindres gestes de Baley, Daneel peut prévoir toutes les décisions effectuées par son partenaire et ainsi anticiper le moment et la direction exacts de toutes ses actions.

Plusieurs explications sont possibles, mais chacune s'appuie sur une vision déterministe. On peut penser, par exemple, que le moindre geste de Baley révèle entièrement le prochain saut de tapis roulant qu'il effectuera. Dès sa première réaction musculaire, Baley deviendrait prisonnier d'une séquence incontournable d'événements, remettant sérieusement en question toute notion de libre choix de la part des humains. Ou encore, Daneel aurait pu développer un modèle mental de Baley tellement précis qu'il lui aurait permis de simuler à l'avance toutes les actions de ce dernier, ce qui implique encore une fois que le libre choix de l'humain est un effet illusoire. Mais, peu importe les explications qu'on peut imaginer, si Baley n'était pas emprisonné dans un monde déterministe, il posséderait la pleine capacité de liberté d'action. Et, avec cette liberté, sur les tapis roulants qui amplifient grandement tout écart ou délai de mouvement entre deux coureurs, Daneel ne pourrait tout simplement pas suivre au pas durant toute la course son partenaire expérimenté qui s'affaire justement à semer ses poursuivants par une série de feintes et de gestes subits. Il s'agit de nouveau d'une démonstration d'un écart important entre le monde de l'œuvre, qui est fondamentalement déterministe, et notre monde réel.

Un dernier exemple, parmi plusieurs autres qui présentent la même caractéristique, est tirée du roman *Face aux feux du soleil*[12], qui raconte une autre enquête de meurtre d'un Spacien menée par Baley. Alors que celui-ci rencontre Albert Minnim, le sous-secrétaire du ministère de la Justice,

afin de recevoir ses instructions pour sa prochaine mission, ce dernier lui indique que la société des humains sur la terre est à risque.

> [...] en l'espace d'un siècle, la Terre ne figurera plus parmi les mondes habités. Tout au moins, voilà ce que prétendent les sociologues.
>
> Baley s'agita, mal à l'aise. On ne pouvait mettre en doute la science des experts ni la logique des ordinateurs.

Cette position de Baley implique que l'on peut faire des prédictions indiscutables de l'avenir de la société plus d'un siècle à l'avance. Cette idée est d'ailleurs déjà présente dans la nouvelle « Le conflit évitable[13] », alors que nous découvrons que les Machines, qui sont des cerveaux positroniques agissant en tant qu'ordinateurs gestionnaires, comprennent parfaitement les forces économiques et sociales que doivent affronter les humains. Avec cette compréhension absolue, les Machines sont en mesure d'effectuer des actions subtiles, qui permettent de maintenir l'humanité sur le droit chemin afin d'éviter dorénavant tout conflit. Comme nous le dit Susan Calvin dans cette nouvelle : « Dites plutôt quelle merveille ! Pensez que désormais et pour toujours les conflits sont devenus évitables. Dorénavant seules les Machines sont inévitables ! »

D'ailleurs ce dernier exemple présente les premières suggestions dans le cycle des robots de la science de la psycho-histoire qui est la pierre angulaire du cycle de la Fondation d'Asimov, qu'il a consolidé vers la fin de son œuvre, notamment tel qu'on le présage dans le roman *Les robots et l'empire*[14].

Tous ces exemples et bien d'autres dans l'œuvre d'Asimov ont donc un trait important en commun : ils sont caractéristiques d'une vision absolument déterministe de l'univers. Une définition du déterminisme[15] énonce qu'à partir d'une

loi physico-mathématique et d'une description complète de la situation actuelle on peut prédire la succession d'événements et des phénomènes à venir. Les enjeux pour avoir des prédictions précises se limitent alors à obtenir un portrait précis de toutes les informations qui décrivent complètement l'état du moment présent, incluant par exemple la position et la vitesse de tous les objets, et à disposer des lois permettant de calculer la conséquence des collisions et des déplacements de ces objets. Cette vision du déterminisme peut s'appliquer non seulement au monde physique, mais également au volet psychologique et sociétal pourvu que l'on connaisse les lois économiques et sociales qui gouvernent le comportement des humains.

Que ce soit donc au niveau matériel ou encore au niveau psychosocial, Asimov s'appuie sur un *novum* dans lequel le déterminisme est omniprésent. Toutes les situations peuvent être prédites avec précision, pourvu que l'on détienne l'information suffisante. Le *novum* sous-entend qu'il n'y a pas de limite fondamentale à l'erreur de mesure et que toute l'information est disponible à la condition de disposer des moyens nécessaires pour la mesurer. Cela est un paradigme complètement différent du monde réel dans lequel nous vivons, qui possède de multiples contraintes et des limites fondamentales sur l'information disponible.

Dans cette vision du déterminisme que présente Asimov, tous les problèmes ont donc une solution ; il ne s'agit que de disposer des ressources nécessaires pour y parvenir. C'est le cas de l'exemple tiré de la nouvelle « Le conflit évitable » décrit ci-dessus, et c'est également le cas de la tentative d'assassinat de Baley avec une flèche sur la planète Solaria dans *Face aux feux du soleil*[16]. Dans ce dernier cas, à partir de quelques informations formulées envers un robot et un enfant, un assassin provoque une situation où un robot remet une flèche empoisonnée à un garçon dont les

prouesses d'archer sont connues et qui tire immédiatement sur Baley qui se trouvait au bon endroit au bon moment, le ratant de justesse. Une situation complètement invraisemblable devient une tentative d'assassinat qui se déroule avec une précision d'horlogerie à la condition que l'on accepte qu'une succession inévitable d'événements aura lieu lorsque certaines conditions sont prédéterminées par l'assassin.

Parmi ces deux éléments fondamentaux du *novum* du cycle des robots, la notion que la production de robots est le choix logique d'une société sur une base économique a déjà été discutée ainsi que le fait que cette situation est en complète contradiction avec la réalité. Il n'est donc pas nécessaire de s'y attarder davantage, puisque toute analyse plus poussée doit maintenant accepter que les robots sont omniprésents tout au long de l'œuvre.

Dans le *novum* d'Asimov, tout comme dans Star Trek, le technobabillage sert de diversion. La puissance faramineuse des cerveaux positroniques des robots leur donne notamment la capacité d'analyser rapidement et précisément une situation et d'agir avec certitude pour respecter les Trois Lois. Cette puissance technologique leur donne également la capacité d'obtenir toutes les informations nécessaires leur permettant de résoudre tout problème et toute situation qui met en cause les Trois Lois. En réalité, peu importe la puissance de leur cerveau, de véritables robots devraient constamment bafouer les Trois Lois, commettant des erreurs simplement par leur incapacité de comprendre entièrement l'environnement qui les entoure. Or la puissance des cerveaux positroniques que présente Asimov lui permet de masquer habilement le fait que c'est l'omniprésence du déterminisme dans son univers fictif, plutôt qu'une percée technologique, qui rend possible le respect des Trois Lois. Dans les sections qui suivent, les contraintes de notre réalité ainsi que des approches

technologiques réalistes seront discutées afin de mieux comprendre les limites de véritables robots et notre capacité d'implémenter une version moins restrictive des Trois Lois de la robotique.

2.3 Les contraintes de l'univers réel

Une version possible du schéma fonctionnel du processus décisionnel des Trois Lois de la robotique, présenté plus loin (page 183), démontre bien qu'un élément essentiel pour l'application des Lois est la capacité pour un robot d'effectuer la cartographie complète de son environnement. D'ailleurs, comme l'indique R. Daneel Olivaw dans *Les robots de l'aube* : « Nous avons constamment conscience des êtres humains. Sans quoi, nous ne pourrions pas remplir nos fonctions[17]. »

Cette cartographie complète permet donc aux robots de respecter la 1re Loi en identifiant tous les humains dans leur entourage, en comprenant toute situation qui mettrait un ou plusieurs de ces humains en danger et en agissant avec efficacité et célérité pour mettre fin à cette situation et éliminer tout risque pour les humains. De même, afin d'obéir aux ordres reçus ou de protéger son intégrité (2e et 3e Lois de la robotique), le robot doit utiliser cette cartographie complète car il ne saurait agir ni même rester immobile sans s'assurer d'abord qu'il n'y a aucune conséquence sur le bien-être des humains dans son environnement. Dans le cycle des robots, les quelques exemples déjà donnés illustrent que, pour Asimov, les robots possèdent dans la majorité des cas des capacités extraordinaires pour cartographier leur environnement, pour identifier tous les humains pouvant être en danger et pour effectuer leur travail avec « une précision absolue » et une vitesse surhumaine. Mais les lois naturelles de notre réalité ont un triple impact en contradiction avec l'œuvre d'Asimov : elles limitent notre capacité de modéliser toutes les situations qui sont pourtant parfois d'apparence assez simple ; elles limitent notre capacité de

capter l'information et elles limitent même la quantité d'information qui peut exister. Chacun de ces impacts sera abordé dans les sections qui suivent afin de démontrer pourquoi il n'est pas possible de concevoir et de fabriquer un robot qui respecterait littéralement et sans erreur les Trois Lois de la robotique.

2.3.1 Des exemples simples où les modèles scientifiques sont complexes

La détermination de la position et de la vitesse précises des objets dans l'environnement, incluant les êtres humains, revêt une importance capitale. À l'origine, les lois de la mécanique, au XVII[e] siècle, ont été édictées dans le domaine de l'astronomie. En se basant originalement sur le mouvement des planètes et des lunes, Newton a élaboré un ensemble de lois qui sont très précises pour calculer les mouvements de ces objets astronomiques. La relativité restreinte est d'abord venue accroître la précision des calculs en tenant compte de la vitesse des objets étudiés par rapport à la limite absolue donnée par la vitesse de la lumière. Par la suite, et pour les calculs les plus précis et surtout en présence de corps très massifs, la relativité générale donne des modifications additionnelles. La précision globale avec laquelle nous pouvons maintenant calculer les orbites des satellites et des sondes interplanétaires est impressionnante. On peut calculer avec une très haute précision la trajectoire requise pour positionner une sonde en orbite autour d'une planète ou d'une lune lointaine du système solaire, même dans le cas où le trajet de la sonde durera plusieurs années. Qu'en est-il des situations qui ne se déroulent pas dans le vide spatiale où les forces de frottement ne sont généralement pas un problème et lorsque des collisions et des mouvements beaucoup plus complexes sont non seulement possibles, mais courants ?

Un premier exemple est la trajectoire complète d'une balle de golf. Un grand nombre de paramètres ont un impact très grand sur la position finale de la balle une fois au repos. On pense facilement à la force avec laquelle le bâton frappe la balle, à l'angle de sa surface de contact, à la durée du contact, à la trajectoire de la tête du bâton au moment de l'impact (ascendant ou descendant) et, une fois la balle dans les airs, à la vitesse et à la direction du vent en tous points de la trajectoire, à la vitesse de rotation de la balle et à l'angle de l'axe de rotation par rapport au sol et à la direction du vent. Au moment de toucher le sol, la dureté et la qualité de la surface, la longueur du gazon et l'angle des brindilles d'herbe, l'humidité du sol, la densité de la pelouse, la rotation de la balle au moment du contact avec le sol auront tous un impact à chaque rebond et lorsque la balle roulera au sol. Pour une précision ultime, on pourrait même prendre en compte la rotation de la terre et les effets de Coriolis. Théoriquement, toutes ces informations pourraient être connues ou mesurées, de sorte que l'on pourrait en principe prédire à quel endroit s'arrêtera la balle de golf. Mais cet emplacement précis d'arrêt de la balle est sujet à une multitude de paramètres qui sont extrêmement difficiles à évaluer et dont la mesure comporte une incertitude qui ne fait que s'accroître lorsque l'on considère chaque interaction possible durant la trajectoire de la balle de golf.

Le billard est un exemple beaucoup plus simple : surface parfaitement plane et très rigide, bandes dont les propriétés élastiques sont bien connues et propriétés des boules également bien connues. Avec un peu de pratique, il devient facile de diriger la boule blanche pour percuter une autre boule qui se retrouvera ensuite au fond d'une des poches de la table. Avec beaucoup plus de pratique, il devient également facile de planifier une séquence de deux contacts pour envoyer une troisième boule au fond d'une poche.

L'utilisation d'une bande peut également devenir routine. Mais pourquoi le coup de départ où l'on casse le paquet est-il si difficile à prévoir, donnant toujours un résultat différent malgré la simplicité apparente du jeu ? Et pourquoi n'avons-nous pas de logiciel assez performant pour simuler la trajectoire précise de la balle de golf ? C'est ce qui est appelé en physique le problème à N-corps pour lequel seuls de très rares cas ont une solution exacte connue. Dès que le nombre d'éléments d'objets impliqués dans des collisions est supérieur à deux, la complexité du calcul des trajectoires devient excessivement élevée. Ajoutons à cela les quatre bandes de la table, les six trous, les déformations des bandes aux abords des trous et le calcul de la position finale des boules nécessite absolument des approximations afin de trouver une solution au problème, en dépit de l'apparente simplicité et de la précision des lois de Newton. Tout cela malgré la connaissance exacte des masses des boules et leurs positions précises dans le cas du billard, ainsi que la vitesse et la direction de la boule blanche au départ. On voit que ce problème s'apparente à l'exemple de la course de Baley sur les tapis roulants déjà décrit dans ce chapitre, où les plus petits changements de mouvements et de synchronisations ont un impact énorme sur le résultat final, mais auquel l'élément humain est ajouté. De nombreuses situations entrent dans la catégorie du problème à N-corps et l'on peut faire un lien direct entre ce type de situation et l'environnement d'un robot où plusieurs objets sont en mouvement. La possibilité de ricochets de ces objets ne fait qu'ajouter à la difficulté. Comment le robot peut-il même savoir si un humain est en danger s'il ne peut pas calculer les trajectoires complètes et la position finale des objets qui l'entourent ? Dans ces premiers cas, l'information pour décrire les problèmes existe bel et bien, mais le problème qui est en apparence assez simple revêt une très grande complexité

qui ne permet pas d'en faire une simulation précise et cette situation ne peut pas être résolue en améliorant uniquement la puissance de calcul disponible. De surcroît, si l'on tente d'inclure dans le calcul les incertitudes de mesure sur tous les paramètres, on s'aperçoit que le cumul de ces incertitudes est tellement grand que cela rend le résultat tout à fait inutile.

Un autre exemple d'une situation relativement courante où l'issue ne peut être anticipée s'illustre par une automobile roulant à haute vitesse sur l'autoroute. On aperçoit soudainement un chevreuil en déplacement sur le terre-plein au centre de l'autoroute. Des voitures nous suivent de près. Comment réagir de manière absolument sécuritaire ? Jusqu'au dernier instant, le chevreuil peut bondir d'une façon complètement imprévue d'un côté ou de l'autre et cela n'est pas dû à une faiblesse de notre capacité d'observation. L'information permettant de prédire la réaction du chevreuil ne se mesure simplement pas, car il faudrait avoir un modèle complet du fonctionnement du cerveau d'un chevreuil, ce qui présume que cet animal n'a aucune capacité de liberté d'action. Il faudrait également connaître toute l'histoire de la vie du chevreuil qui influencera certainement sa réaction. Le danger ne résulte donc pas du simple manque d'un modèle adéquat pour simuler les réactions d'un chevreuil, modèle dont un robot-chauffeur pourrait disposer, mais du manque d'information qui n'est absolument pas disponible ou qui n'existe simplement pas. La solution simple de freiner brusquement pour amener le véhicule à l'arrêt n'est pas sans risque, car les véhicules qui suivent peuvent entrer en collision avec nous. Si tous les véhicules ont des robots chauffeurs avec la capacité de communiquer entre eux grâce à une technologie sans fil, le risque peut certainement être diminué, mais un véhicule ne peut subir une décélération plus grande qu'une certaine valeur critique sans blesser les

humains qui s'y trouvent, nonobstant le fait que tout véhicule nécessite une distance minimale pour freiner correctement, sans perte de contrôle, sur une surface routière. Il s'agit donc d'une situation périlleuse pour laquelle il n'y a pas de solution certaine qui puisse garantir la sécurité de tous les humains.

Ces exemples pratiques illustrent différents types de situations pour lesquelles il n'est pas trivial d'anticiper tous les événements subséquents. Les cas du golf et du billard demeurent cohérents avec une vision déterministe des événements, leur complexité et notre incapacité d'obtenir une solution entièrement prévisible découle plutôt de l'énorme quantité de variables nécessaires pour en effectuer l'analyse et de l'instabilité inhérente du résultat par rapport à de très petites variations dans chacune de ces variables. Mais, afin de cartographier son environnement, un robot doit pouvoir observer et mesurer l'ensemble des variables qui correspondent à la situation qui l'entoure. On fait alors appel à des capteurs qui peuvent prendre des mesures optiques, mécaniques, électriques, chimiques ou conçus pour tout autre type de variable. Dans tous les cas, ces capteurs sont caractérisés par des incertitudes de mesure. Certaines de ces incertitudes sont dues à la conception ou à la performance limitée du capteur, mais, en tentant d'améliorer ces performances, on bute rapidement sur des limites de la physique fondamentale qui sont incontournables. En examinant ces limites, on découvre que l'information souhaitée concernant les variables de l'environnement du robot est non seulement difficile à mesurer, mais peut même ne pas exister du tout. Dans les sections qui suivent, on tracera un portrait de quelques types de mesure qui permettront d'en comprendre les contraintes, mais aussi les limites fondamentales.

2.3.2 Incertitudes et limites des capteurs

Tous les instruments de mesure possèdent des incertitudes qui limitent la qualité de l'information qu'ils peuvent capter. Prenons à titre d'exemple une simple balance classique avec deux plateaux et deux bras pour mesurer le poids d'un objet. La limite de précision de cet instrument est donnée par le plus petit poids étalon dont on dispose pour le calibrer. Plusieurs autres facteurs peuvent en limiter la précision ; par exemple, tout courant d'air peut perturber l'équilibre, donc la mesure ; lorsque le mouvement devient faible, la friction affectant les deux bras tout près de l'équilibre peut figer le mouvement à une fausse valeur ; enfin, la température peut causer un changement dans l'expansion des pièces métalliques et le point d'équilibre de la balance. On peut certainement améliorer grandement la précision de l'instrument : un cabinet en verre fermé peut couper les courants d'air, des lubrifiants appropriés peuvent réduire grandement la friction et la température dans le cabinet peut soit être contrôlée pour éviter presque complètement le problème d'expansion thermique, dans la limite de précision du système de contrôle de température lui-même, soit mesurée et prise en compte dans le calibrage de l'appareil, limité par la précision du thermomètre. On ne peut cependant pas obtenir une mesure plus précise que l'erreur de calibrage de l'appareil par rapport à l'étalon, ultimement le prototype international du kilogramme, situé au pavillon de Breteuil en France. Au moyen de processus normés, cette masse permet de calibrer les balances et d'obtenir le poids des objets lorsque l'on connaît la gravité localement. Les incertitudes de calibrage peuvent donc être ramenées à l'étalon du kilogramme, qui détermine la définition même de la mesure de masse telle qu'elle a été convenue par la communauté internationale. Plus nos appareils sont bien conçus et fabriqués, plus l'incertitude peut être minimisée. Cependant,

puisque tout processus réel comporte des erreurs, plus il y a d'étapes entre l'étalon et l'étape de calibrage de la balance, plus ces erreurs sont grandes. Et malgré les suggestions qui viennent d'être présentées pour mitiger les effets qui perturbent l'utilisation de la balance, on demeure aux prises avec un système de mesure qui n'est pas opéré sous le vide de l'espace, qui n'est pas à une température de zéro Kelvin et qui n'utilise pas des graisses sans friction (qui n'existent d'ailleurs pas). Une erreur de mesure en résultera nécessairement.

Considérons maintenant un second type d'instrument dont on souhaiterait équiper un robot et qui est utilisé par des millions de personnes chaque jour : l'appareil de photographie numérique. Ces appareils consistent en un microcircuit conçu pour être sensible à la lumière, par exemple un capteur CCD, composé de millions d'éléments photosensibles. Plusieurs caractéristiques en limitent la précision. D'abord, la taille de chacun de ces éléments photosensibles limite la résolution de l'image formée sur le capteur par un système optique. On ne peut distinguer les détails de l'image plus petits que la dimension de cet élément photosensible, habituellement appelé pixel. Ensuite, la superficie totale du capteur limite la taille de l'image pouvant être capturée, donc de l'information totale présente dans l'image. Aujourd'hui, les meilleurs appareils photo grand public ont des limites de quelques dizaines de mégapixels (ou millions de pixels), tandis que les appareils pour usage professionnel tels que ceux qui ont été utilisés récemment dans les sondes martiennes peuvent posséder plus de 150 millions de pixels. Une autre contrainte pour un système d'imagerie provient de la qualité de l'optique, ou des lentilles utilisées pour former l'image sur le capteur. Les meilleurs objectifs pour les appareils photographiques reflex ont des propriétés largement supérieures à celles des objectifs intégrés dans un téléphone cellulaire. La qualité de chacune des lentilles de

ces objectifs joue un rôle important pour minimiser les distorsions géométriques de l'image, par exemple sur les bords, et pour minimiser les distorsions spectrales de l'image résultant du fait que la lumière de différentes couleurs est focalisée en un point différent par chacune des lentilles. Les objectifs de haute qualité comprennent souvent plus d'une dizaine de lentilles choisies pour compenser et minimiser globalement ces effets. Au-delà de la technologie des capteurs et de la fabrication des lentilles qui limitent la précision de ces systèmes optiques, mais qui continueront de s'améliorer au cours des prochaines années, cet exemple nous permet d'aborder un premier niveau de limites beaucoup plus fondamentales sur la précision d'un appareil de mesure.

D'abord, le capteur optique est un microcircuit qui convertit le signal optique en signal électronique. Techniquement, on convertit des photons en générant des électrons lorsque ces photons frappent le capteur optique. Un des problèmes, c'est que, même en l'absence de tout photon, le capteur générera des électrons. Cela est dû au bruit fondamental en électronique. En laboratoire, on réduit l'impact de ce problème en refroidissant les capteurs de haute précision à la température de l'azote liquide ou à des températures inférieures. Cependant, cette approche représenterait des coûts exorbitants si l'on voulait la démocratiser à tous les capteurs optiques. Toujours présent, ce bruit devient donc une limite particulièrement importante lorsque l'éclairage est très faible. Le robot ne pourrait alors plus distinguer les photons focalisés sur le capteur par rapport au bruit électronique et ne serait donc plus capable de cartographier son environnement visuellement. Ce problème devient de plus en plus crucial alors que la compétition pour augmenter la résolution et la qualité des images numériques pousse les manufacturiers à réduire les dimensions physiques des

pixels dans les capteurs. Cependant, pour un microcircuit possédant des pixels extrêmement petits, donc ayant en principe une résolution très élevée, le nombre de photons captés par chaque pixel devient très faible même dans des conditions d'éclairage normal, puisque chaque pixel couvre une infime superficie du microcircuit complet sur lequel l'objectif projette l'image. Le bruit électronique devient alors fort par rapport au signal causé par l'arrivée des photons. Dans cet exemple, une solution technologique poursuivie par les manufacturiers dans le but d'accroître la qualité de l'image résultera en une perte de la qualité de la même image pour toute situation où l'éclairage est moindrement faible. Bien que chaque génération de nouveaux appareils photographiques incorpore des processeurs toujours plus puissants afin d'optimiser l'image obtenue avec un capteur spécifique, ce bruit électronique découle de la physique fondamentale des matériaux et ne peut être évité. Nous avons ici une première limite dans la qualité des images qui pourraient être obtenues par un robot en effectuant la cartographie visuelle de son environnement.

L'objectif optique lui-même impose également des limites. Un objectif adapté pour recueillir un maximum de lumière, normalement obtenu avec un grand diamètre de lentilles et un diaphragme pleinement ouvert, produira une image avec une profondeur de champ faible. Cela veut dire que, si l'image à l'avant-plan est nette, l'arrière-plan sera flou. La situation est un peu meilleure si l'on tente d'observer des objets lointains, mais la région où l'image est nette demeure restreinte, et les objets en avant-plan seront flous. Afin de contrer cette situation et d'obtenir une image nette sur tout le champ de vision, on ajuste le diaphragme de l'objectif pour en réduire le diamètre effectif. En revanche, cette réduction du diamètre effectif bute rapidement sur le problème de la limite de diffraction. Avec un ajustement judicieux, on peut

effectivement éliminer les zones floues de l'avant-plan jusqu'à l'arrière-plan, mais il y a dégradation de l'image car les lois fondamentales de la physique optique indiquent qu'en réduisant l'ouverture de la lentille la capacité de distinguer deux points très rapprochés dans une image est compromise. L'image paraît relativement nette, mais un examen minutieux révèle que les détails deviennent de plus en plus difficiles à distinguer lorsque le diaphragme de la lentille est de plus en plus fermé. Techniquement pour les amateurs de photographie, la limite de diffraction pour des appareils photographiques reflex d'une vingtaine de mégapixels apparaît généralement pour des ouvertures plus petites que f/11. D'ailleurs, la limite de diffraction est une des raisons qui explique pourquoi il serait impossible pour le robot Cutie de focaliser le faisceau énergétique sur des distances interplanétaires dans la nouvelle « Raison[18] ». Pour ces deux caractéristiques des systèmes optiques, soit la taille des pixels dans le microcircuit et l'objectif optique, les lois de la physique fondamentale imposent une limite certaine sur l'information qui peut être recueillie.

Une discussion exhaustive pourrait également se faire pour tous les types de capteurs imaginables. Le système optique n'est qu'un exemple permettant de faire ressortir assez facilement les limites imposées par les lois de la nature. Ces lois s'appliquent peu importe le type d'onde électromagnétique observée, soit visible, infrarouge, ultraviolet, rayons X, etc., mais également pour tout autre type de signal, par exemple les ondes sonores.

2.3.3 Des limites au traitement de l'information

La diversité de technologies connues pour capter l'information soulève cependant un autre point qui est passé sous silence dans l'œuvre d'Asimov : la capacité d'emmagasiner et de traiter l'information. Dans notre quotidien, nous sommes de plus en plus aux prises avec ce problème. La quantité

d'information disponible sur Internet est assourdissante. Notre besoin de l'emmagasiner dépasse largement la capacité des produits électroniques qui se retrouvent dans nos maisons ou dans nos milieux de travail. Une part de plus en plus grande de notre portefeuille d'information personnelle se retrouve sur des systèmes de stockage distants, grâce à l'infonuagique. Nos téléphones intelligents et tablettes informatiques, en plus de nos ordinateurs personnels et portables, sont constamment en interaction avec ces sites de stockage pour y chercher nos photographies et nos vidéos que nous partageons réciproquement avec parents et amis, le tout d'une façon qui nous est devenu presque totalement transparente. Nous bénéficions ainsi d'une capacité de stockage qui dépasse ce que nous aurions pu imaginer il y a à peine quelques années. Or, tel que nous le confirme R. Daneel Olivaw dans *Face aux feux du soleil* : « Tous ces robots de type solarien sont reliés entre eux par radio. »

Les robots auraient donc la capacité de communiquer avec des sites distants par des technologies sans fil en vertu d'un protocole semblable à ce que nous appelons aujourd'hui l'infonuagique. Le défi posé par la nécessité d'emmagasiner une quantité énorme d'information serait relevé. Le traitement de ces informations et le forage pour y trouver les éléments pertinents en temps réel demeurent cependant très difficiles. Un robot doit pouvoir simuler son environnement immédiat en détail afin de prédire les dangers et leurs impacts potentiels sur les humains. Mais la réalité se déroule à une vitesse beaucoup plus grande que la vitesse des modèles que l'on utilise pour les simulations.

Reprenons l'exemple du jeu de billard et du premier coup où l'on frappe le paquet de boules. Les limites sur la capacité de calculer la position des boules une fois que celles-ci sont immobilisées sont fondamentales. Il n'existe pas d'approximation capable de simuler en temps réel le résultat de ce

coup extrêmement complexe, même si toutes les forces en jeu sont du domaine de la mécanique classique newtonienne. Posons une situation très hypothétique où deux gangsters jouent au billard avec le pari que, si l'un deux réussit à entrer au moins cinq boules sur le premier coup, le second acceptera qu'il s'agit d'une peine de mort pour lui. Le premier gangster demande à son robot de jouer le coup. Malgré tous ses capteurs qui peuvent analyser la position précise de chaque boule avant le coup, tous les pouvoirs de traiter cette information et le contrôle le plus précis possible de ces actionneurs pour manipuler la baguette, il ne pourra que suivre le résultat du coup en temps réel et sa meilleure stratégie est sans doute celle qu'un humain utiliserait, c'est-à-dire frapper aussi fort que possible pour que les boules circulent le plus longtemps possible sur la table puisque cela augmente leur chances de tomber dans les poches.

Aujourd'hui, l'interprétation du langage humain demeure un grand défi avec un impact important sur la Deuxième Loi qui dicte qu'un robot doit obéir aux ordres. L'exemple récent de Watson, le système conçu par IBM pour jouer à *Jeopardy !* contre deux adversaires humains, a permis d'illustrer la complexité de l'enjeu. Appuyer par une grappe de calcul à l'état de l'art et situé dans un site distant, l'ordinateur a certes pu gagner la partie, mais seulement après des années de travail pour créer sa base de données et les outils pour interpréter la question posée par l'animateur et forer la base de données pour trouver la réponse la plus probable. Cette base de données est conçue expressément pour ce jeu, où la structure de phrase est particulière alors que l'on donne la réponse et que le concurrent doit trouver la bonne question. Watson ne pourrait pas mener une conversation avec un humain. Il s'agit là d'un problème très difficile, particulièrement complexe à exécuter en temps réel et il démontre une capacité d'analyse et de forage de données incroyable

pour un robot autonome tel qu'on le retrouve dans l'œuvre d'Asimov. Cependant on peut toujours imaginer que des percées technologiques importantes pourraient un jour donner cette capacité à un robot, malgré le fait que, selon toute vraisemblance, ce jour est encore très distant.

Au-delà de ces défis de nature technologique, la discussion concernant les capteurs optiques a permis de voir que le robot est limité par une incapacité fondamentale de cartographier parfaitement son environnement, donc une incapacité de savoir de manière absolue quels humains peuvent être en danger près de lui. Ces limites réelles découlent de lois fondamentales de la physique. Dans plusieurs circonstances face à ces limites fondamentales, un robot ne pourrait effectivement pas remplir ses fonctions car des humains en danger pourraient être présents dans l'environnement immédiat du robot, mais être situés au-delà de sa capacité à les détecter. Mais la physique de notre réalité, contrairement au monde d'Asimov, ne fait pas qu'imposer des limites à l'information disponible que l'on peut détecter avec des capteurs. Au niveau très fondamental de la structure de notre univers, et de manière complètement contre-intuitive, certaines informations n'existent tout simplement pas. Dans la section suivante, on verra les conséquences du fait que toute particule réelle possède également les propriétés d'une onde, et le résultat du débat qui a fait rage pendant plus de la moitié du XX^{e} siècle et qui met en scène les plus grand physiciens de l'époque pour comprendre la nature même de la réalité. Les résultats d'expériences déterminantes ont éventuellement mené à la constatation que l'univers n'est pas du tout déterministe, contrairement au monde fictif du cycle des robots.

2.3.4 Les limites de l'existence de l'information

Une expérience réalisée par Thomas Young, un étudiant en médecine au début du XIX^{e} siècle, et présentée à la Société

royale à Londres au grand dam des défendeurs de la théorie de Newton, a des conséquences qui renversent un concept fondamental de notre réalité. Dans cette expérience, Young a séparé une source unique de lumière en deux faisceaux qu'il a ensuite fait interférer. Le motif d'interférence qu'il a obtenu a prouvé, pour la première fois, que la lumière était une onde. Au XX^{e} siècle, une compréhension plus approfondie de la physique a mené au principe de la dualité onde-particule tant pour la lumière que pour la matière. Tous les objets de notre univers, à une échelle suffisamment petite, présentent des caractéristiques à la fois de particules et d'ondes. Pour la matière, cela signifie qu'un faisceau d'électrons qui traverse un écran avec deux fentes, petites et rapprochées l'une de l'autre, va produire un motif d'interférence sur un plan situé au-delà de ces fentes, de la même façon que Young avait initialement observé pour la lumière. De façon encore plus intéressante, si l'on contrôle un faisceau composé de photons ou d'électrons de manière à ce qu'une seule particule à la fois traverse l'écran avec les deux fentes et que ces fentes sont conçus pour s'assurer que chaque onde-particule ne puisse passer que par une seule d'entre elles, on obtient encore le motif d'interférence. Ces résultats expérimentaux nous forcent à réaliser qu'il est impossible de déterminer par quelle fente chaque particule est passée. Cette information n'existe tout simplement pas.

Einstein s'est opposé longuement à la théorie de la mécanique quantique dont découlent plusieurs principes d'incertitude fondamentale. Mais, au fil des décennies, la mécanique quantique s'est avérée extrêmement précise et capable de prédire de nombreux résultats expérimentaux, véritable preuve scientifique de la véracité d'une théorie. Découlant de cette théorie, le principe d'incertitude énoncé d'abord par Heisenberg décrit le comportement des particules et des ondes à l'échelle nanométrique. Ce principe

d'incertitude peut prendre plusieurs formes et l'une de celles-ci indique qu'on ne peut connaître avec grande précision à la fois la position et la vitesse d'une particule, par exemple un électron. Cette réalité a plusieurs conséquences importantes : dans les capteurs, il existe une couche d'incertitude absolue qui ne peut être résolue par des avancées ou des percées technologiques. Un robot ne pourra jamais cartographier l'environnement avec une précision absolue, ni, comme dans le cas de la nouvelle « Raison », diriger un faisceau énergétique avec une précision absolue vers une station réceptrice distante. L'information nécessaire pour accomplir ces tâches n'existe tout simplement pas avec une précision absolue. Albert Einstein avait longuement argumenté que l'information devait exister et que le problème découlait simplement de l'existence de variables cachées. Il a proposé une théorie qui expliquait sa vision dans une publication en 1935 avec ses collègues Boris Podolsky et Nathan Rosen[19]. Cette publication, qui décrivait le paradoxe surnommé EPR (pour les initiales des trois auteurs), a été l'objet de controverses pendant des décennies. Mais, en 1964, John Stewart Bell a démontré théoriquement[20] que les prédictions découlant du papier sur le paradoxe EPR devraient mener à des résultats expérimentaux forts différents de ce que l'on obtient par l'application directe de la mécanique quantique. Il y avait maintenant une base pour mener des expériences et prouver la véracité d'une vision ou de l'autre. Mais ces expériences sont très difficiles à réaliser, et ce n'est qu'au début des années 1980, presque 50 ans après la publication du paradoxe EPR, que le physicien français Alain Aspect a finalement obtenu les premiers résultats expérimentaux[21] qui ont permis de trouver la solution. Ces expériences ont démontré que la vision d'Einstein et de ses collègues, qui préconisaient l'existence de variables cachées, ne pouvait pas expliquer correctement les

résultats expérimentaux, mais que la mécanique quantique qui prédit une incertitude fondamentale passait le test avec succès. L'information manquante pour décrire entièrement le comportement d'une particule ou d'une onde est simplement inexistante. D'autres expériences depuis celle d'Aspect vont dans le même sens. Aujourd'hui, la majorité des physiciens acceptent cette réalité troublante de la mécanique quantique.

Bien que l'impact de la mécanique quantique se fasse sentir surtout à l'échelle nanométrique, cette théorie est absolument cruciale pour le fonctionnement de la majorité des technologies que nous utilisons aujourd'hui dans notre quotidien. Les systèmes de télécommunication par fibre optique, les lecteurs de disques optiques, tous les ordinateurs et puces microélectroniques, tous les capteurs d'imagerie optique, bref tout ce qui est à base d'électronique ou de photonique s'appuie sur la mécanique quantique. Les incertitudes fondamentales, sans que nous en soyons conscients, font partie de la fondation architecturale de notre société technologique. On ne peut y échapper, et on ne peut plus imaginer que seule la mécanique classique newtonienne déterministe gouverne notre quotidien. Récemment, des scientifiques ont découvert que des processus quantiques ont également un impact significatif dans les systèmes biologiques. Il est connu depuis longtemps que les organismes biologiques fonctionnent à partir de la conversion d'énergie vers des formes utiles pour les transformations chimiques, il s'agit là de la base de la photosynthèse. Les réactions chimiques et l'absorption de l'énergie lumineuse en sont des exemples qui sont gouvernés par la mécanique quantique. Mais des effets beaucoup plus complexes de cohérence quantique sont présents dans les mécanismes de photosynthèse et, plus récemment, des expériences suggèrent que les oiseaux exploitent l'intrication quantique (où deux états de

la matière se superposent simultanément) pour détecter les champs magnétiques et ainsi s'orienter correctement[22]. Dans le cerveau humain, les échanges chimiques et de charges électroniques sont au cœur du transfert d'information entre les neurones et, comme ces deux types de phénomènes sont gouvernés par la mécanique quantique, on ne peut y échapper.

Notre réalité ne cadre donc pas du tout avec une vision d'un univers déterministe comme celle d'Asimov. L'impossibilité d'avoir une connaissance absolue de l'environnement qui entoure les robots et de déterminer le résultat de toutes les interactions en cours et de tous les comportements biologiques en proximité rend incontournable la conclusion que les Trois Lois, et certainement la Première Loi de la robotique, seraient continuellement bafouées par simple manque ou absence d'information suffisante. Les incertitudes qui persistent permettront tout au plus aux robots de faire de leur mieux tout en commettant des erreurs dont certaines seront assurément fatales pour des humains. Malgré toutes les limites fondamentales discutées dans ce chapitre, la prolifération de robots pour des applications militaires et l'étude de robots pour l'utilisation en milieu de la santé maintiennent l'impératif de créer des robots éthiques, même si ceux-ci ne seront pas infaillibles. Des outils mis au point au cours des dernières décennies par les chercheurs en robotique seront présentés dans les prochaines sections dans le but de découvrir les approches pragmatiques que l'on peut utiliser aujourd'hui et dans un avenir rapproché afin de concevoir des robots éthiques qui se conforment aux contraintes de notre réalité.

3. MISE EN ŒUVRE DES LOIS DE LA ROBOTIQUE : LES AVANCÉES DE L'INTELLIGENCE ARTIFICIELLE (IA)

Dans les sections précédentes, nous avons vu qu'en raison des différences intrinsèques entre l'univers d'Asimov et le

nôtre le modèle de mise œuvre des Trois Lois de la robotique décrit dans l'œuvre d'Asimov serait impossible à transposer dans un contexte réel. Nous avons également vu comment les limites des modèles scientifiques actuels et des technologies disponibles réduiraient les capacités d'interprétation et le potentiel d'action d'un robot asimovien, affaiblissant ainsi grandement sa capacité à appliquer les Trois Lois de la robotique.

Malgré ce constat, étant donné l'avancement technologique actuel, pourrait-on imaginer un plan de mise en œuvre des Trois Lois de la robotique qui serait réalisable dans un avenir proche et qui respecterait les réalités physiques de notre univers ? D'abord, il faut comprendre qu'en raison des limitations discutées plus haut, contrairement aux robots asimoviens, les robots résultant d'une telle mise en œuvre ne seraient pas moralement infaillibles. Contrairement à Dave, dans « Attrapez-moi ce lapin », un robot réel échouerait plus souvent qu'autrement le test des « plus hautes fonctions du monde robotique : la solution de problèmes de jugement et d'éthique[23] ». Heureusement, dans le cas d'une mise en œuvre réelle, les Lois de la robotique ne seraient pas intégrées directement dans l'essence même de l'unité centrale du robot. Ainsi, un manquement à l'application d'une des Lois, surtout à la Loi 1, ne risquerait pas de causer de traumatisme à ce dernier, comme l'a si souvent décrit Asimov dans les romans du cycle des robots et dans la nouvelle intitulée « Menteur ![24] ».

Au-delà de ces considérations plus fondamentales, il serait quand même très intéressant d'évaluer la faisabilité de la mise en œuvre des Trois Lois de la robotique dans un contexte réel. Pour y arriver, il est important de bien comprendre ce qui distingue les robots des autres systèmes informatiques, électroniques et intelligents que nous côtoyons fréquemment. Il est également nécessaire de

connaître l'éventail des outils technologiques disponibles pour y arriver et d'en comprendre le potentiel et les limites.

3.1 Les robots et les Lois de la robotique

Qu'est-ce qu'un robot ? Un robot est un dispositif mécatronique (alliant mécanique, électronique et informatique) accomplissant automatiquement soit des tâches qui sont généralement dangereuses, pénibles, répétitives ou impossibles pour les humains, soit des tâches plus simples mais en les réalisant mieux que ce que ferait un être humain[25]. » On peut retenir de cette définition que les robots sont d'abord des outils motorisés destinés à des usages précis, prescrits par une programmation et situés dans un environnement réel. C'est justement cette incarnation physique du robot dans un environnement bien réel qui le distingue des applications informatiques usuelles auxquels nous sommes peut-être plus habitués et dont le pouvoir d'action se limite à un monde virtuel. À cette définition, nous ajouterons que les robots asimoviens sont conçus pour assister *directement* les humains et qu'ils sont sujets à de fréquentes interactions avec ces derniers. Ces interactions peuvent être explicites, lors de l'énonciation verbale d'un ordre par exemple, mais peuvent également être implicites, comme c'est le cas lorsqu'un robot doit effectuer une tâche dans un environnement dans lequel il côtoie d'autres travailleurs humains.

C'est dans ce contexte d'intégration des robots dans la société humaine que les Lois de la robotique prennent tout leur sens. En effet, l'impact de ces Lois est plutôt limité si l'on considère des robots isolés ayant peu ou pas de contacts avec les humains, comme c'est le cas très souvent pour les sondes spatiales robotisées, les robots sous-marins, ou même les robots d'assemblage sur les chaînes de montage. Avant de continuer, il serait important de revoir ces Trois Lois telle qu'elles ont été exposées pour la première fois dans la nouvelle « Cercle vicieux », en 1942 :

1) Un robot ne peut porter atteinte à un être humain, ni, restant passif, permettre qu'un être humain soit exposé au danger.
2) Un robot doit obéir aux ordres que lui donne un être humain, sauf si de tels ordres entrent en conflit avec la Première Loi.
3) Un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la Première ou la Deuxième Loi[26].

En se basant sur ces Lois, il est possible de représenter le processus décisionnel d'un robot asimovien à l'aide d'un schéma fonctionnel, tel qu'illustré à la figure 1. Bien que le schéma proposé soit simplifié, certains éléments pourraient être décrits avec plus de détails, il permet de distinguer les compétences, ou les outils, dont un robot devrait disposer pour être en mesure de respecter les Trois Lois de la robotique. Ces outils devraient donc être associés principalement à :

1) Des compétences sensorielles nécessaires pour observer et mesurer l'environnement du robot (capteurs) ;
2) Des compétences analytiques et d'interprétation, pour repérer les dangers potentiels à l'intégrité du robot ou à celle des humains dans leurs champs d'action, mais servant également à comprendre les interactions avec les humains (intelligence artificielle) ;
3) Des compétences décisionnelles, pour permettre au robot d'établir les priorités et les chaînes d'actions à prendre pour maintenir son intégrité (Loi 3), pour réaliser un ordre (Loi 2) ou pour protéger un humain d'un danger quelconque (Loi 1) (intelligence artificielle) ;
4) Des compétences motrices pour que le robot interagisse avec son environnement, les humains et avec les objets physiques qui le composent (actionneurs et moteurs).

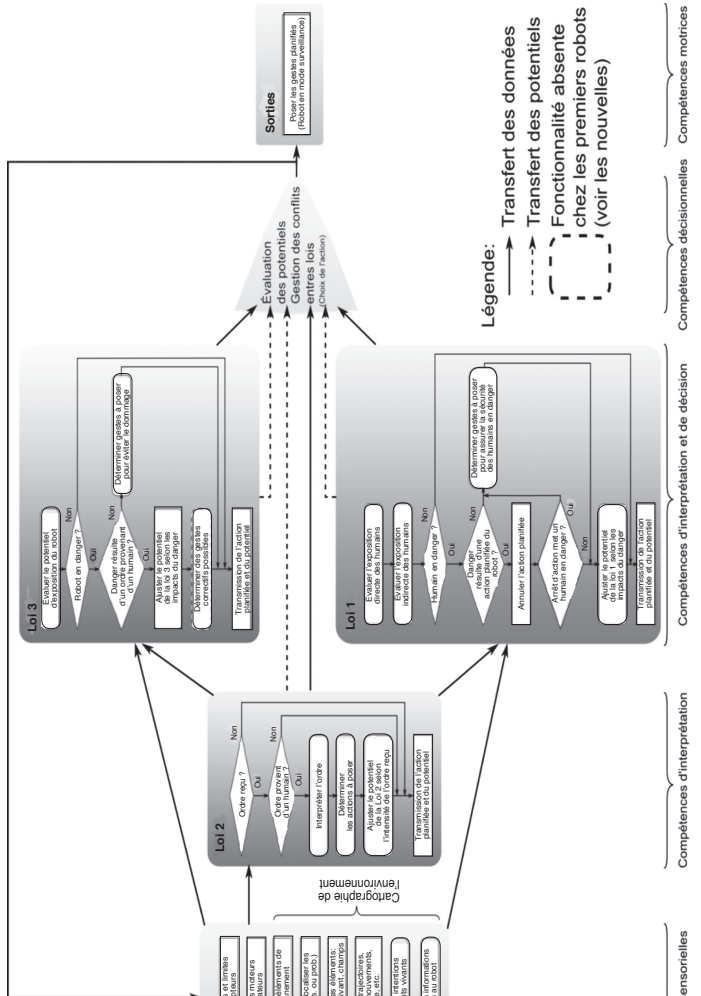

FIGURE 1 – Schéma simplifié du processus décisionnel chez un robot asimovien

Chacun de ces outils devrait pouvoir s'appuyer sur un cinquième type de compétences. Celles-ci pourraient être décrites comme une base de connaissances transversales suffisamment complète pour permettre au robot de reconnaître les objets physiques présents dans son environnement, d'établir les interactions existantes entre ces objets (modèles physiques, chimiques, biologiques et sociétaux), d'interpréter le langage humain, d'interpréter le sens des phrases et de comprendre les accents et les expressions régionales. Cette base de connaissances devrait également permettre au robot de prendre des décisions éclairées, respectant les lois, les us et les coutumes de la société dans laquelle il évolue, de transmettre de l'information aux humains de façon efficace en respectant les distinctions linguistiques et culturelles des individus auxquels il s'adresse ainsi que de faire des gestes tout en respectant ses propres limites motrices. Comme on peut le voir, différentes compétences s'appliquent aux étapes d'applications des Trois Lois dans le schéma décisionnel.

Le schéma décisionnel décrit aussi la façon dont l'information est transmise entre les outils sensoriels, chacune des Trois Lois, le système d'évaluation des potentiels et les outils moteurs du robot. Les potentiels d'application de chacune des Lois sont aussi transmis vers le système d'évaluation. En cas de conflit insurmontable entre les actions proposées par les Trois Lois, une boucle de rétroaction permet au robot de réanalyser la situation. Cette boucle de rétroaction permet au robot de remettre en question un ordre qui semble absurde. Chez Asimov, cette boucle est aussi responsable de défaillances du cerveau positronique du robot, ou « robloc », comme dans la nouvelle « Menteur ![27] » ou dans le roman *Les robots de l'aube*[28].

Dans la section précédente de ce chapitre, il a déjà été question des limites fondamentales et technologiques associées aux capteurs et aux actionneurs sur lesquels se basent les compétences sensorielles et motrices. Il a été montré que l'utilisation de ces technologies induit systématiquement des incertitudes quant aux mesures et aux observations que peuvent faire les robots. De la même façon, la précision, la vitesse et la force des actions des robots sont limitées. Dans les prochaines sections, il sera question principalement des compétences d'interprétation et de décision des robots. Nous verrons différents outils disponibles aujourd'hui qui pourraient nous aider à réaliser des robots éthiques, respectant l'esprit des Trois Lois de la robotique. Ainsi, nous présenterons des outils propres au domaine de l'intelligence artificielle et nous évaluerons l'impact de leurs limites sur une mise en œuvre potentielle de Lois éthiques.

3.2 Automates, raisonnement symbolique et calculateurs

Historiquement, ce sont justement ces compétences d'interprétation et de décision, souvent associées à la pensée rationnelle ou au raisonnement chez l'homme, qui ont inspiré philosophes, scientifiques et auteurs. Depuis l'Antiquité, de nombreux auteurs ont mis en scène des créatures mécaniques ou artificielles dotées d'une intelligence artificielle et capables de raisonnements complexes. Les golems et les automates font probablement partie des premières créatures de fiction dotées d'une telle intelligence artificielle.

Plus récemment, on ne peut passer sous silence le roman de Mary Shelley, *Frankenstein*, ou encore la pièce de théâtre tchèque *R.U.R.* de Karel Čapek[29] dans laquelle le mot *robot* fut utilisé pour la première fois. Cette question de l'intelligence artificielle continue d'être un élément déterminant dans les œuvres contemporaines de science-fiction.

En fait, la question de l'intelligence artificielle se base sur une hypothèse initiale supposant que la pensée peut être modélisée par une série de lois formelles. Cette idée remonte à Aristote, dans l'Antiquité, qui fut le premier à proposer une série de lois gouvernant le mode de pensée rationnel. Il conçut un système de syllogismes permettant d'atteindre des conclusions mécaniquement en se basant sur une série de prémisses initiales[30]. Au Moyen Âge, le philosophe majorquin Ramon Llull suggéra qu'il était possible d'émuler les raisonnements au moyen d'artefacts mécaniques. Il imagina un ensemble de machines permettant de combiner des vérités initiales grâce à des opérations logiques « mécaniques » et d'en produire du savoir[31]. Léonard de Vinci, le scientifique allemand Wilhelm Schickard et Blaise Pascal s'inspirèrent tous en quelque sorte de ces idées pour concevoir et fabriquer les premières machines à calculer. À l'époque, Pascal écrivait que « la machine arithmétique produit des effets qui semblent beaucoup plus près de la pensée que toutes les actions des animaux[32] ». Llull eut également une importante influence sur les philosophes du XVII[e] siècle Thomas Hobbes et Gottfried Leibniz. Le premier proposa que le raisonnement était identique à un calcul numérique, que nous « additionnons et soustrayons nos pensées silencieuses[33] ». Leibniz, quant à lui, envisageait un langage universel propre au raisonnement qui permettrait de réduire une argumentation ou un processus de déduction à un calcul mathématique. Ces philosophes étaient en train d'établir les hypothèses de base du système formel (ou symbolique), une des pierres de fondation de la recherche en intelligence artificielle.

Afin de créer une intelligence artificielle qui aurait les compétences réflexives nécessaires pour mettre en place des Lois de la robotique, il est indispensable de disposer de deux éléments essentiels : un modèle d'intelligence et un support

pour l'embarquer. Jusqu'au milieu du XX[e] siècle, le support par prédilection était les systèmes mécaniques. Cela changea de façon draconienne avec l'arrivée des premiers calculateurs. Les premiers calculateurs modernes, tels que le Z3, le Colossus et l'ENIAC, ont été conçus durant la Seconde Guerre mondiale pour décoder les messages encryptés allemands[34]. Contrairement aux machines précédentes, ces calculateurs fonctionnaient grâce à des relais électriques et des tubes à vide, alors que leur mémoire était, plus souvent qu'autrement, constituée de cartes perforées. C'est également pendant cette période qu'Alan Turing proposa le concept de machine de calcul logique, appelée plus tard machine de Turing[35]. Ce concept montre que toutes les formes de calcul informatisé[36] peuvent être exprimées de façon numérique et séquentielle. Connaissant le modèle de la machine de Turing, John von Neumann proposa une architecture pour les calculateurs dans laquelle une unité de calcul indépendante utilisait une unité de mémoire commune pour les programmes et les données[37]. Les travaux de Turing et de von Neumann s'inspiraient de récentes découvertes en neurologie qui montraient que le cerveau était un réseau de connexions électriques constitué de neurones au travers desquels l'information se transmettait sous forme d'impulsions binaires (présence ou absence d'impulsion). À l'époque, on imaginait qu'il serait possible de fabriquer un cerveau électronique en se basant sur ces idées[38].

Avec la création de calculateurs numériques, dès le milieu des années 1950 et par la suite avec l'arrivée des ordinateurs, certains chercheurs ont reconnu que ces machines conçues pour manipuler des nombres pouvaient probablement manipuler également des symboles[39]. Cette proposition fut démontrée par Newell et Simon avec l'introduction d'un programme, *Logic Theorist*, le premier démonstrateur

automatique de théorèmes basé sur une logique symbolique[40]. Les fondations nécessaires à la mise en place d'un système de raisonnement mécanique, ou système d'intelligence artificielle, étaient maintenant établies. Un tel système de raisonnement est nécessaire si l'on souhaite mettre en place les compétences d'interprétations et de décision essentielles pour la réalisation de robots éthiques.

3.3 L'intelligence artificielle

Grâce au développement de modèles symboliques pour représenter les concepts et les raisonnements et grâce au développement des ordinateurs programmables, la science de l'intelligence artificielle (IA) pouvait maintenant démarrer. Mais qu'est-ce que l'*intelligence artificielle* ? Il s'agit d'un concept qui peut être difficile à définir. Définir l'intelligence artificielle nécessite donc de définir également le concept d'*intelligence*. De plus, l'attribut *artificiel* peut avoir différentes connotations : il peut rappeler les créatures frankensteinesques de la littérature de science-fiction tout en soulignant la frontière floue et mouvante entre les concepts de nature et d'artifice. On comprend qu'en raison des concepts sous-jacents il puisse être difficile de donner une définition philosophique au concept d'intelligence artificielle. Nous essaierons tout de même de proposer une définition opératoire en nous basant sur quelques définitions historiques.

En 1950, comme le concept d'intelligence est très difficile à définir, Alan Turing essaya de contourner le problème en proposant un test pour déterminer si une machine pouvait être considérée comme étant intelligente. Ainsi, si la machine peut tenir une conversation avec un être humain, par exemple au moyen d'un téléscripteur, sans qu'un second être humain écoutant leur conversation puisse l'identifier comme étant une machine, alors celle-ci peut être considérée comme étant intelligente[41].

En 1955, John McCarthy, un des pionniers en IA, fut le premier à utiliser le terme intelligence artificielle : « Le rôle de [la science de] l'intelligence artificielle est de développer des machines qui se comportent comme si elles étaient intelligentes[42]. » Cette définition a deux problèmes. D'abord elle est récursive, c'est-à-dire que le concept d'intelligence sert à se définir lui-même. Ensuite, elle peut nous amener à interpréter un comportement comme intelligent, alors qu'il ne l'est pas. On a montré qu'il était possible de produire des comportements complexes, organisés et qui peuvent sembler intelligents en reliant un capteur directement à un moteur. C'est ce que montre par exemple le véhicule de Braitenberg[43]. Dans son expression la plus simple, sur sa partie arrière, la machine possède un capteur de lumière qui est connecté et qui stimule la rotation d'une roue. Ainsi, plus une source lumineuse sera intense, plus le véhicule s'éloignera rapidement de la lumière. La noirceur résultera en un arrêt complet. On pourrait, de façon erronée, expliquer ce comportement en prétendant que la machine est effrayée par la lumière et qu'elle la fuit.

L'encyclopédie Britannica définit l'intelligence artificielle comme étant « l'habileté des ordinateurs numériques ou des robots contrôlés par ordinateur d'accomplir des tâches habituellement associées aux capacités de raisonnement intellectuel supérieur des êtres humains […][44] ». Cette définition a aussi une faiblesse. On peut facilement associer les opérations mathématiques complexes et la mémorisation d'une grande quantité d'informations comme étant des tâches associées *aux capacités de raisonnement intellectuel supérieur des êtres humains*. Cependant, cette définition considère que tous les ordinateurs actuels sont des intelligences artificielles puisqu'ils sont capables d'effectuer ces tâches… et souvent bien mieux que les humains.

Finalement, en 1983, Elaine Rich, constatant que certaines tâches demeuraient encore aujourd'hui l'apanage des humains, et ce malgré tous les progrès réalisés en informatique et en robotique au cours des dernières décennies, proposa une nouvelle définition : « L'intelligence artificielle est un domaine de recherche scientifique qui a pour objectif de découvrir des moyens qui permettraient aux ordinateurs d'effectuer aussi bien, sinon mieux, ces tâches que les humains réussissent mieux que les machines[45]. » Bien qu'elle soit plutôt pragmatique, cette définition a l'avantage d'être applicable aujourd'hui, comme elle l'était en 1950 et comme elle le sera en 2050. Elle permet aussi de diviser les domaines de recherche de l'intelligence artificielle selon les compétences que l'on cherche à développer.

3.3.1 La logique

La logique est le premier outil de raisonnement qui fut développé en intelligence artificielle[46]. Elle permet de faire des déductions à partir de propositions dont on connaît la véracité. L'intérêt est de dériver de nouvelles connaissances ou de répondre à un problème en partant de connaissances initiales établies. Au début, les machines intelligentes se basaient sur des systèmes de logique propositionnelle et de logique du premier ordre.

En logique propositionnelle, les propositions et les déductions, ou résultats de calculs logiques, ne peuvent avoir que deux valeurs : vrai ou faux. Par exemple, mettons-nous dans le contexte de la Première Loi de la robotique. Faisons une première proposition : *un être humain a besoin d'oxygène pour vivre*. Nous savons que cette proposition est vraie. Faisons une seconde proposition : *la pièce contient de l'air*. Cette proposition peut être vraie ou fausse, selon l'état de la pièce. Les déductions se font au moyen de tables de vérité. La table de vérité représente un algorithme logique permettant de décrire tous les ensembles de propositions valides et

leurs déductions dans un temps fini. Dans notre exemple, on calcule les conditions propices au maintien en vie d'un être humain à partir de la table de vérité suivante :

A : Un être humain a besoin d'oxygène pour vivre	B : La pièce contient de l'air	A et B = Condition propice à la vie
Vrai	Vrai	Vrai
Vrai	Faux	Faux

Toutefois, le temps nécessaire à la résolution de tels ensembles croît très rapidement avec le nombre de propositions. Dans le pire des cas, le nombre de clauses intermédiaires nécessaires à la résolution d'un problème croît exponentiellement avec le nombre de propositions initiales[47]. Élargissons notre exemple en introduisant l'effet de tous les gaz. Il faudrait créer une nouvelle proposition pour chacun des gaz toxiques et une pour chacun des gaz non toxiques (azote, vapeur d'eau, gaz carbonique, etc.) et une pour l'oxygène. On se retrouverait rapidement avec des milliers de propositions, sans compter celles qui sont associées aux mélanges de gaz. On peut comprendre que le temps requis pour résoudre un problème complexe impliquant plusieurs variables peut être si long qu'une fois une solution trouvée, cette dernière n'aura plus de sens pratique puisque le problème initial aura évolué. Si nous reprenons notre exemple, l'humain sera décédé à la suite d'une exposition à un gaz toxique. L'amélioration de la vitesse des calculateurs permet de diminuer l'importance de ce problème. Cependant, comme nous le verrons plus tard, plusieurs chercheurs en intelligence artificielle sont persuadés que la force brute, directement associée à la puissance de calcul des ordinateurs, ne peut résoudre à elle seule ce problème.

La logique propositionnelle a tout de même une utilité certaine. Elle est utilisée chaque jour pour le développement et la vérification de circuits numériques, essentiels à la fabrication des microprocesseurs dont sont composés tous nos systèmes informatiques. Dans la mise en place de systèmes intelligents, comme les systèmes experts, la logique propositionnelle peut également être utile pour résoudre des problèmes simples. Cependant, toutes les variables doivent être discrètes et il ne doit y avoir aucune relation croisée entre ces variables.

L'expression et la résolution de problèmes complexes peuvent se faire plus simplement grâce à la logique du premier ordre. Contrairement à la logique propositionnelle, la logique du premier ordre permet de faire des calculs plus abstraits grâce à la création d'ensembles. Comme cette formalisation de la logique est plus complexe que ce que nous avons vu plus haut et qu'en décrire la sémantique nous éloignerait trop du sujet de ce chapitre, nous nous concentrerons sur quelques exemples qu'on pourrait associer à la mise en place des Lois de la robotique. Reprenons notre exemple des gaz toxiques. Dans un tel cas, trois ensembles de gaz seraient créés : un pour les gaz toxiques, un pour les gaz non toxiques et un pour l'oxygène. Le calcul peut ensuite se faire à partir des ensembles. L'humain dans la pièce n'est exposé à aucun gaz toxique et il est en présence d'oxygène ? Si c'est vrai, il est dans une situation propice à la vie. Toute autre situation le met en danger.

En 1929, Kurt Gödel a prouvé, par son théorème de complétude, que la logique du premier ordre est complète[48]. En d'autres mots, « on peut donner un nombre fini de principes (axiomes logiques, schémas d'axiomes logiques et règles de déduction) qui suffisent pour déduire de façon mécanique toutes les lois logiques[49] ». Tout énoncé pouvant être formulé en logique du premier ordre peut être démontré

au moyen de règles formelles. Cette propriété de la logique de premier ordre entraîna une euphorie chez les chercheurs en IA durant les années 1970. Plusieurs croyaient qu'il était possible de formuler tous les problèmes de représentation du savoir et de raisonnement en logique du premier ordre et de les résoudre à l'aide d'un solveur automatisé. On semblait donc avoir en main tous les outils nécessaires pour créer une machine intelligente universelle[50].

Une des premières applications de l'intelligence artificielle fut celle des jeux vidéo. Au début des années 1950, Christopher Strachey utilisa le Ferranti Mark 1 de l'Université de Manchester, le premier ordinateur disponible commercialement, pour écrire le premier programme fonctionnel d'intelligence artificielle : un jeu de dames. Le premier jeu d'échecs fut également écrit sur cet ordinateur par Dietrich Prinz. Conçu pour résoudre des situations simples de jeu, le programme examinait tous les mouvements possibles jusqu'à ce qu'un mouvement le mettant dans une position plus avantageuse soit trouvé[51]. À l'époque, ces programmes fonctionnaient en utilisant les outils de logique propositionnelle et de logique du premier ordre décrits plus haut. Avec le temps de nouveaux outils ont été conçus, mais l'intelligence artificielle dans les jeux reste quand même une des méthodes les plus utilisées pour évaluer les progrès dans le domaine de l'IA. En considérant le test de Turing et les définitions de l'IA proposées plus haut, il existe peu d'outils informatiques aussi efficaces que les jeux vidéo, tel que les dames, les échecs, le backgammon ou le Go, pour comparer le caractère intelligent d'une machine avec celui d'un humain[52].

Durant la deuxième moitié des années 1950 et pendant les années 1960, les résultats obtenus grâce au développement des outils logiques et ceux qui ont été obtenus dans d'autres champs de l'IA présageaient d'un avenir grandiose

pour l'intelligence artificielle. Les pionniers ne taisaient pas leur optimisme et se permettaient quelques prédictions. En 1958, Herbert A. Simon et Allen Newell proposaient que, « d'ici dix ans, un ordinateur numérique serait le prochain champion d'échecs[53] », en 1965, Simon renchérissait en indiquant que des « machines seraient capables, d'ici vingt ans, de faire n'importe quel travail que peut faire un homme[54] ». Plus tard, en 1967, Marvin Minsky allait jusqu'à dire que « d'ici une génération [...] le problème consistant à créer l'intelligence artificielle serait substantiellement résolu[55] ». Minsky était tellement confiant qu'il confia au *Life Magazine* en 1970 que, « dans trois à huit ans, nous aurons une machine ayant l'intelligence générale d'un être humain moyen[56] ». Cependant, à l'exception de la prédiction concernant le championnat d'échecs, qui arriva beaucoup plus tard, aucune de ces prédictions ne se réalisa.

Les outils logiques sont à la base de l'informatique. Ils sont fondamentalement imbriqués dans la structure des microprocesseurs et des mémoires et ils permettent la programmation d'une multitude de fonctionnalités. Ils sont amplement suffisants pour programmer les actions d'un robot industriel et pour mettre en place certaines règles simples visant à éviter les collisions avec son environnement et avec les travailleurs. Cependant, les développeurs se sont rapidement aperçus que les outils logiques sont très limités quand vient le temps de résoudre des problèmes complexes, caractéristiques des environnements réels.

3.3.1.1 Les limites de la logique

En prolongeant le raisonnement qu'il avait commencé par l'énonciation de son théorème de la complétude en 1931, Gödel proposa ses théorèmes de l'incomplétude dans lesquels il stipula qu'en logique d'ordre supérieur, une logique permettant de créer des ensembles abstraits, certains énoncés véridiques pouvaient être impossibles à

prouver[57]. Certains problèmes complexes ne pouvant pas être énoncés en logique propositionnelle ou en logique du premier ordre pouvaient donc être impossibles à résoudre. Avec ces deux théorèmes, Gödel venait de découvrir une des limites fondamentales de la logique et du raisonnement symbolique.

Un autre des problèmes fondamentaux des outils logiques est l'explosion combinatoire, ou l'explosion de l'espace de recherche. Bien que les outils logiques fonctionnent très bien dans des micro-univers où le nombre d'objets et le nombre de solutions possibles sont limités, ces outils deviennent peu pratiques lorsqu'on s'attaque à des situations plus complexes[58]. En effet, plus on fournit d'informations initiales à un solveur automatisé, plus le nombre de solutions éventuelles augmente. Dans le pire des cas, le solveur doit évaluer chacun des cas afin de trouver une solution, ce qui peut devenir impossible dans un temps raisonnable. Cela est primordial dans notre analyse de faisabilité de la mise en place des Trois Lois de la robotique. En effet, cette explosion de l'espace de recherche peut avoir un impact limité pour un robot travaillant dans un environnement discret, comme les robots fixes dans les chaînes de montage. Toutefois, cet impact est désastreux pour les robots autonomes, qui fonctionnent dans un environnement continu. Seulement pour mettre en place la Loi 1, le nombre de données à considérer est si grand que le temps nécessaire pour évaluer le risque d'exposition d'un humain à un danger quelconque serait dangereusement trop long pour permettre au robot de faire quelque geste que ce soit.

Finalement, les outils logiques décrits plus haut ne peuvent pas tenir compte de l'incertitude associée aux données initiales, ils peuvent difficilement tenir compte de l'évolution temporelle d'un environnement et ils sont assujettis à propager les erreurs introduites lors de l'entrée des

données (mesures ou instructions erronées) ou introduites par les modèles physique, chimique, biologique, psychologique, etc.

Comme les outils logiques sont victimes de limites importantes, telles que l'incomplétude, l'explosion combinatoire et l'incapacité de tenir compte de l'incertitude, leur utilisation est clairement insuffisante pour mettre en place les compétences d'interprétation et de décision. D'autres outils de l'intelligence artificielle développée jusqu'aux années 1970 auraient pu être utilisés. Toutefois, comme nous le verrons, ces outils possédaient aussi certaines contraintes qui en limitaient le déploiement.

3.3.2 Les limites des premiers systèmes en intelligence artificielle

À partir du milieu des années 1970, les chercheurs en IA essuyèrent plusieurs critiques. Bien que les premiers systèmes en intelligence artificielle étaient capables de résoudre des problèmes triviaux, leur capacité à résoudre des problèmes complexes étaient plutôt limitée. Les chercheurs commencèrent à buter sur des limites techniques imposantes et des problèmes fondamentaux. Certaines de ces limites affectent encore le milieu aujourd'hui.

Nous l'avons vu précédemment avec les outils logiques, la puissance de calcul nécessaire pour résoudre des problèmes augmente rapidement avec le degré de complexité des problèmes. Dans les années 1970, la puissance de calcul et la mémoire des ordinateurs étaient tout simplement insuffisantes pour accomplir quoi que ce soit d'intéressant[59]. Hans Moravec suggérait, en 1976, que les ordinateurs étaient des millions de fois trop lents pour pouvoir démontrer de l'intelligence[60]. Cela affecte directement la vitesse d'interprétation, d'analyse et de prise de décisions nécessaire dans une mise en œuvre des Trois Lois de la robotique. Aujourd'hui, comme

les ordinateurs sont beaucoup plus puissants, nous avons réussi à dépasser cette limite pour résoudre des problèmes se déroulant dans des contextes très précis, comme les jeux et les quiz. Cependant, comme nous le verrons plus tard, la puissance de calcul des ordinateurs est encore trop faible pour que ces derniers puissent agir de façon autonome dans des contextes libres sans nécessiter d'assistance humaine.

Les premiers systèmes intelligents disposaient d'une capacité de raisonnement symbolique, mais cette capacité ne reposait sur aucune connaissance contextuelle propre au problème traité[61] ou de sens commun. Ces connaissances sont primordiales pour des applications comme l'interprétation de la vision ou du langage. Le système doit avoir suffisamment de connaissances sur le contexte, ou sur son environnement physique dans le cas d'application en vision, pour avoir une idée de ce dont il est question ou pour reconnaître ce qu'il observe. Par analogie, on pourrait supposer qu'un tel système nécessiterait une base de connaissances semblable à celle d'un enfant[62]. Dans les années 1970, nous étions incapables d'emmagasiner une telle quantité d'information et nous étions encore moins capables de la rendre intelligible. Russell et Norvig illustrent ce constat en détaillant les résultats d'un logiciel de traduction du russe vers l'anglais conçu par les Américains. En utilisant les règles de transformation syntaxique pour traduire la grammaire et le vocabulaire russe, les scientifiques américains voulaient analyser rapidement les articles scientifiques russes publiés à la suite du lancement du premier *Sputnik* en 1957. Toutefois, comme le travail de traduction nécessite un ensemble de connaissances générales pour résoudre les ambiguïtés et pour établir le contexte du texte, les résultats étaient souvent décevants. Une traduction célèbre illustre très bien les difficultés rencontrées : la phrase russe « The spirit is willing but the flesh is weak » donnait en anglais

« The vodka is good but the meat is rotten[63]. » On comprend rapidement l'importance de cette limite lors d'activités d'interprétation du langage et de l'environnement.

Finalement une dernière limite est décrite par le paradoxe de Moravec. Ce dernier propose que les raisonnements supérieurs soient beaucoup plus faciles à reproduire et à simuler par un système intelligent que les aptitudes sensorimotrices humaines. Cela peut sembler contre-intuitif. En effet, alors qu'il nous est relativement facile d'effectuer des tâches motrices ou sensorielles, le raisonnement supérieur, quant à lui, nécessite des efforts beaucoup plus grands[64]. Ainsi, bien que nous ayons développé des outils avancés de raisonnement symbolique, l'information provenant des systèmes sensoriels d'un robot est souvent incomplète et interprétée de façon limitée. Le pouvoir d'action d'un robot est aussi limité par un système moteur limité en fonctionnalités, en polyvalence et en précision. Les compétences décisionnelles d'un robot ne sont pas limitées seulement par ces capacités de raisonnement, mais aussi par sa capacité à observer et à comprendre son environnement et à interagir avec celui-ci.

Conscients des limites des outils logiques et des premiers systèmes en intelligence artificielle, les chercheurs dans le domaine ont tenté d'explorer de nouvelles approches. Dans les prochaines sections, nous présenterons ces outils plus modernes et discuterons de leur potentiel pour la réalisation des robots éthiques.

3.3.3 Les outils modernes en intelligence artificielle

Bien que les outils logiques s'avèrent très importants en intelligence artificielle, ils sont souvent incapables de résoudre des problèmes complexes. Dans cette section, nous présenterons d'autres outils propres à l'intelligence artificielle. La majorité de ces outils ont été conçus pour répondre

aux critiques que les premiers systèmes intelligents ont essuyées, surtout au cours des années 1970. Nous verrons que certains de ces outils, comme la logique floue et les systèmes experts dans une certaine mesure, s'alignent sur le fonctionnement de la logique propositionnelle et de la logique du premier ordre, alors que d'autres outils, tels que les systèmes d'apprentissage, les agents intelligents et les robots, s'appuient sur d'autres approches pour mettre en œuvre des compétences d'interprétation et de décision.

3.3.3.1 Gérer l'incertitude : la logique floue

Une des principales faiblesses de la logique vient de la nature binaire que peuvent prendre les propositions ou les variables : un élément ne peut être que vrai ou faux. Cela limite l'expression de la réalité à un ensemble de certitudes et ne laisse aucune place à l'incertitude dans un processus de résolution de problème. Or, nous avons vu précédemment que les instruments de mesure et les actuateurs introduisaient une incertitude chaque fois qu'une mesure ou qu'une action était prise, respectivement. La logique floue permet de répondre à ce problème en insérant une suite de valeurs intermédiaires entre une valeur vraie et une fausse. Elle permet de décrire un système en fonction de connaissances statistiques. Par exemple, en logique floue, l'énoncé *97 % des oiseaux peuvent voler* permet de résoudre un problème sur les oiseaux sans exclure les manchots (qui ne peuvent voler) de la catégorie des oiseaux. La logique floue permet également de décrire des systèmes logiques à partir de classifications qualitatives[65]. La figure 2 illustre cela en classifiant la température sur une échelle de trois expressions : froid, tiède et chaud. Chaque point sur l'échelle possède trois états logiques, un pour chacune des expressions. À la température décrite par la ligne verticale, les flèches indiquent la valeur pour chacune des expressions. Comme la flèche rouge est à zéro, la température ne peut pas être considérée comme

chaude (0 % chaud). La flèche orange (20 % tiède) décrit la température comme étant *légèrement tiède*, alors que la flèche bleue (80 % froid) la décrit comme *étant plutôt froide*[66]. Dans un système de contrôle, la logique floue permet de passer doucement d'un régime de contrôle à un autre.

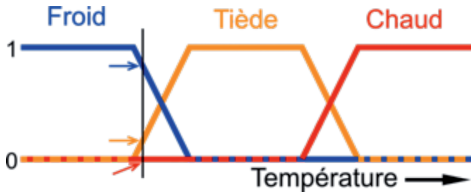

FIGURE 2 – Description de trois expressions de la température dans un système de contrôle en logique floue

Il serait impossible de passer à côté de la logique floue lors d'une mise en œuvre des Trois Lois de la robotique. Comme il est pratiquement impossible de classer parfaitement les éléments dans des catégories fermées, la logique floue offrirait une solution pragmatique qui permettrait d'interpréter des ordres qualitatifs, de reconnaître les objets et les humains et de gérer l'incertitude dans la réponse des capteurs et des actuateurs.

Bien que les fondements théoriques de la logique floue ne soient pas parfaitement établis[67], cet outil est utilisé avec succès

> dans des domaines aussi variés que l'automatisme (freins ABS), la robotique (reconnaissance de formes), la gestion de la circulation routière (feux rouges), le contrôle aérien, l'environnement (météorologie, climatologie, sismologie, analyse du cycle de vie), la médecine (aide au diagnostic), l'assurance (sélection et prévention des risques) et bien d'autres[68].

Notons cependant que, même si la logique floue est moins affectée par le problème de l'explosion combinatoire que les autres outils logiques, ce problème limite tout de même son utilisation pour des applications complexes. Les systèmes de logique floue sont aussi grandement influencés par le degré d'incertitude des informations en entrée.

Reprenons l'exemple des gaz toxiques utilisé plus haut afin d'illustrer ce problème. Il est possible que les capteurs d'un robot soient incapables de déterminer la nature du gaz dans la pièce où se trouve l'humain. En logique floue, dans l'ignorance, le robot établira que la proposition « La pièce contient un gaz toxique » est vraie à 50 % et fausse à 50 %. Contrairement aux systèmes basés sur la logique propositionnelle et sur la logique du premier ordre, le robot fonctionnant avec un système de logique floue pourra effectuer son raisonnement, malgré l'incertitude. Cependant, en raison de la très grande incertitude des données en entrée, le résultat risque d'être « Sortir l'humain de la pièce » à 50 % et « Laisser l'humain dans la pièce » à 50 %. Une telle solution n'a, bien sûr, aucun sens pratique.

3.3.3.2 Les systèmes experts

En intelligence artificielle, les systèmes experts sont des programmes informatiques capables d'émuler le processus décisionnel d'humains *experts*[69]. Ils permettent de résoudre des problèmes complexes en se basant sur une banque de connaissances propres à un domaine. Ils peuvent ainsi servir comme outils d'aide à la décision.

Les premiers systèmes experts sont apparus dans les années 1970 et leurs applications se sont multipliées durant les années 1980. On considère que les systèmes experts furent parmi les premiers succès véritables en intelligence artificielle[70].

Un système expert est donc un programme informatique pouvant répondre à des questions, en effectuant un raisonnement à partir de règles et de faits connus. Il possède une structure unique, différente de celle des programmes usuels. Cette structure est composée d'une base de connaissances variables, d'une base de données, d'un moteur d'inférence, d'outils explicatifs et d'une interface utilisateur[71]. La base de connaissances contient les informations propres au domaine. Ces informations sont représentées sous forme de règles SI (condition) → ALORS (action). Ces règles permettent de décrire des relations, des recommandations, des déductions ou des directives dans un langage intelligible. La base de données contient une série de faits entrés par l'utilisateur. Le moteur d'inférence est capable d'utiliser ces faits et ces règles pour produire de nouveaux faits, jusqu'à l'obtention d'une réponse à la question experte posée. Ces nouveaux faits s'ajoutent à la base de données pour la durée de la résolution. Les outils explicatifs permettent à l'utilisateur d'interroger le système sur son processus de raisonnement. Finalement, l'interface utilisateur assure la communication entre le système et un utilisateur. Elle permet à ce dernier de poser une question experte et d'introduire les faits dans la base de données.

La plupart des moteurs d'inférence utilisés dans les systèmes experts reposent sur des outils de logique formelle, décrits plus haut, et utilisent le raisonnement déductif. Pour l'essentiel, ils utilisent une règle d'inférence basée sur le syllogisme. Les systèmes experts les plus simples s'appuient sur la logique propositionnelle. Rappelons que, dans cette logique, on n'utilise que des propositions, qui sont vraies ou fausses. D'autres systèmes s'appuient plutôt sur la logique du premier ordre. Il faut cependant noter qu'il existe également des systèmes experts se basant sur la logique floue.

Le premier système expert, Dendral, fut créé en 1965. Il permettait d'identifier les constituants chimiques d'un matériau à partir de mesures de spectrométrie de masse et de résonance magnétique nucléaire. Aujourd'hui, on utilise les systèmes experts dans différents domaines d'applications comme la géologie, l'automatisation ou le diagnostic médical. Dans ce domaine d'application, on dénote par exemple le système expert médical Lexmed, un système expert pour le diagnostic des appendicites[72]. Ce système utilise une série de faits décrivant les symptômes du patient et son état de santé général ainsi qu'environ 500 règles pour proposer un diagnostic et éventuellement proposer un traitement.

Les systèmes experts ont eu leur heure de gloire dans les années 1980, où l'on a trop rapidement pensé qu'ils pourraient se développer massivement. En pratique, le développement de ce genre d'application est très lourd car, lorsque l'on dépasse la centaine de règles, il devient difficile de comprendre comment le système expert « raisonne » (comment il manipule faits et règles en temps réel), donc d'en assurer la mise au point finale puis l'entretien[73]. À l'instar des outils logiques, les systèmes experts sont aussi victimes de l'explosion combinatoire.

Les systèmes experts constituent des outils intéressants pour mettre en œuvre les Trois Lois de la robotique. En fait, les Trois Lois sont écrites dans une syntaxe très proche de celle qui est utilisée pour décrire les règles des systèmes experts. Cependant, comme le diable est dans les détails, il serait nécessaire de développer un ensemble de règles pour décrire toutes les situations possibles. Bien que cela soit probablement possible pour un robot fonctionnant dans un environnement discret et observable (par exemple un robot-outil sur une chaîne de montage), on serait rapidement victime de l'explosion combinatoire pour les robots fonctionnant dans un environnement continu, partiellement

observable et changeant. L'écriture de ces règles deviendrait rapidement très complexe et de nombreuses situations risqueraient de ne pas être couvertes par celles-ci. Idéalement, il faudrait que le robot puisse ajouter lui-même de nouvelles règles propres à son contexte d'utilisation grâce à un processus d'apprentissage. De plus, un tel système consisterait également en une combinaison de systèmes experts. Selon le contexte, on devrait donc s'attendre à ce qu'un ensemble de règles ait priorité sur les autres. Il sera nécessaire d'établir une méthode pour gérer la priorité des règles. Dans les sections sur les agents intelligents et sur les agents éthiques, nous verrons entre autres comment il est possible d'établir de tels systèmes de gestion des priorités, soit en les programmant, soit en laissant le système les apprendre par lui-même.

3.3.3.3 Système de connaissances, apprentissage et exploration de données

Les systèmes experts sont un sous-ensemble des systèmes à base de connaissances. Comme dans les applications complexes, les systèmes doivent reposer sur une large quantité d'information. La programmation de ces informations selon la structure standard, c'est-à-dire directement dans le code du programme, peut devenir rapidement une tâche difficile. L'idée derrière les systèmes à base de connaissances est de séparer les connaissances du programme qui les utilise pour tirer des conclusions ou répondre à un problème. Ainsi, le programme comporte un moteur d'inférence lui permettant de manipuler les éléments de connaissance.

L'acquisition de connaissances peut se faire au moyen d'écriture de règles établies par des spécialistes dans un domaine particulier, comme dans le cas des systèmes experts, mais elle peut aussi se faire par programmation ou par l'interrogation de bases de données distantes. Cette

acquisition de connaissances peut également se faire par de l'apprentissage machine. La figure 3 propose une représentation schématique de la structure générale du système classique à base de connaissances[74].

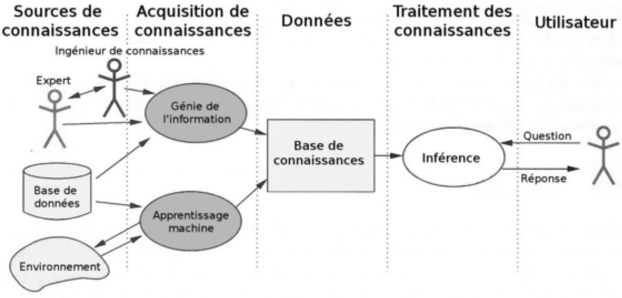

FIGURE 3 – Architecture générale d'un système à base de connaissances[75]

Cette séparation des connaissances permet de mettre à jour la base de connaissances sans avoir à reprogrammer le système d'inférence. Elle permet aussi de développer les systèmes d'inférence en faisant abstraction du champ d'application, ce qui simplifie grandement le travail de programmation.

Du point de vue d'un développeur, la structure derrière un comportement intelligent peut devenir si complexe qu'il peut être très difficile, voire impossible, de la programmer de façon efficace. L'apprentissage automatisé permet à un système d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. Cet apprentissage peut permettre d'abord d'accumuler de l'information nouvelle, mais également d'établir de nouvelles relations et de nouvelles règles. Les algorithmes utilisés

permettent, dans une certaine mesure, à un système d'adapter ses analyses et ses comportements en se fondant sur l'analyse de données empiriques provenant d'une base de données ou de capteurs. Comme l'ensemble de tous les comportements possibles devient rapidement trop complexe à décrire à cause de l'explosion combinatoire, on confie au programme le soin d'adapter un modèle permettant de simplifier cette complexité et de l'utiliser de manière opérationnelle. Ce modèle a l'avantage d'être adaptatif, de façon à prendre en compte l'évolution de la base de connaissances pour laquelle les réponses ont été validées : le système a la capacité de s'améliorer. Ainsi, un système pourrait apprendre à classer des fruits en fonction de leur taille et de leur couleur[76]. Les systèmes d'apprentissages automatisés sont utilisés aujourd'hui avec énormément de succès pour la reconnaissance de caractères d'imprimerie ou d'écritures manuscrites. Ces outils de reconnaissance optique des caractères font habituellement appel à des réseaux de neurones artificiels, un modèle de calcul mathématique utilisé en IA dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques[77].

Dans notre monde informatisé, chaque année, on observe une multiplication par deux de la quantité d'information qui transite sur les réseaux de télécommunications. Avec des projets de recherche mondiaux comme le séquençage du génome humaine, la recherche du boson de Higgs au grand collisionneur d'hadrons (LHC), la recherche de planètes extrasolaires, nous produisons chaque jour plus de données que nous pouvons en analyser. C'est une chose d'emmagasiner des nouvelles données, c'en est une autre d'extraire des connaissances de ces données et d'en faire de l'information intelligible. C'est cet aspect de l'apprentissage automatisé que l'exploration de données[78], ou *data mining* en anglais, permet d'accomplir. L'objectif de l'exploration de

données est littéralement la découverte de nouvelles connaissances[79]. Idéalement, ces connaissances extraites doivent être également intelligibles pour les utilisateurs et les développeurs. L'exploration de données est faite depuis les années 1990 par les grandes compagnies du Web, comme Google, Amazon ou Facebook, afin de produire des campagnes de marketing et de publicité ciblées en fonction des achats des consommateurs.

L'exemple le plus parlant des systèmes à base de connaissances est probablement le système Watson d'IBM. Sur le site du projet DeepQA, les chercheurs d'IBM expliquent que l'objectif du projet est d'illustrer comment l'accessibilité étendue et croissante d'informations exprimées en termes de langage naturel ainsi que l'intégration et l'avancement de l'analyse de ce langage naturel, de l'exploration de données, de l'apprentissage machine, de la représentation des connaissances et de raisonnement ainsi que la disponibilité de système de calcul massivement parallèle pourraient faire en sorte que des technologies de réponse aux questions pourraient rivaliser avec les meilleures performances humaines[80]. En février 2011, Watson a gagné le quiz *Jeopardy !* alors qu'il se mesurerait à deux champions du jeu. Pour y arriver, la base de connaissances de Watson était constituée de plus de 4 térabits d'informations, incluant le texte complet de l'encyclopédie en ligne Wikipedia. Ces réponses rapides et son court temps de réaction se fondaient sur la mise en place de 90 serveurs IBM Power 750, chacun disposant de 32 processeurs, pour un total de 2 880 processeurs mis en parallèle[81].

Aujourd'hui, les avancés en microélectronique nous ont permis de développer des supports pour emmagasiner facilement des tonnes d'information. Nous avons conçu des systèmes d'apprentissage machine et d'exploration d'information afin de rendre ces informations intelligibles pour les machines, mais également pour les êtres humains.

Cependant, ces systèmes ne sont efficaces que dans des contextes très précis et nécessitent des puissances de calcul informatique bien supérieures à celles d'un ordinateur personnel. Au cours des prochaines années, on doit s'attendre à ce que les nouvelles avancées en microélectronique et en nanotechnologie nous permettent d'améliorer ces systèmes afin de les rendre plus polyvalents et plus accessibles. De tels systèmes seront essentiels pour gérer les données provenant des nombreux capteurs dont on équipe maintenant les robots et pour leur permettre d'acquérir de nouvelles connaissances.

3.3.4 Le monde réel

Une des critiques des premiers outils de l'intelligence artificielle, exprimée par le paradoxe de Moravec, soulignait les limites des aptitudes sensorimotrices des systèmes intelligents. Vers la fin des années 1980, préoccupés par ce paradoxe, plusieurs chercheurs proposèrent qu'il était nécessaire d'intégrer les systèmes intelligents dans des systèmes physiques, ou des robots, afin de démontrer une intelligence véritable[82]. L'intelligence devait être incarnée, elle devait être capable de percevoir son environnement et d'interagir avec lui. Selon eux, ces aptitudes sensorimotrices étaient essentielles au développement du système de raisonnement basé sur le sens commun. À ce sujet, en 1988, Moravec écrivait[83] qu'il était

> confiant que cette approche ascendante vers l'intelligence artificielle allait, un jour, rencontrer, plus qu'à mi-chemin, la route de la traditionnelle approche descendante afin de mettre en place des compétences pour le monde réel et des connaissances de sens commun qui ont été [...] insaisissables par les programmes de raisonnement[84].

3.3.5 Les agents intelligents

Depuis les années 1990, on commence à penser les systèmes intelligents d'une nouvelle façon. Ce nouveau paradigme, appelé agent intelligent, propose une approche modulaire fondée sur le principe visant à diviser pour régner et inclut des concepts empruntés à la théorie décisionnelle et aux sciences économiques[85]. De façon très générale, on peut définir un agent comme étant un système capable de traiter de l'information provenant de ports d'entrée et d'en produire une action sous forme de sortie. Ce sont les façons dont l'information est acquise (capteurs, accès à une banque de données, communication, etc.), les façons dont les sorties sont exprimées (actions physiques, communication, écriture en mémoire, etc.) et les moyens utilisés pour traiter l'information qui distingue les agents entre eux[86]. Plus précisément, les agents intelligents sont des entités autonomes qui perçoivent leur environnement et qui font des gestes, propres au contexte environnemental, afin d'atteindre un objectif. Pour illustrer, la figure 4 présente le fonctionnement d'un agent intelligent réduit à sa plus simple expression, celle d'un réflexe. Le lexique utilisé pour décrire les agents suggère que ceux-ci doivent absolument œuvrer dans un environnement réel, comme le seraient des systèmes de contrôle pour freins ABS, des thermostats électroniques ou des robots. Certains auteurs utilisent également le terme agent intelligent pour décrire des systèmes logiciels autonomes réalisant des tâches, dans l'environnement du Web par exemple, au nom d'un utilisateur. Il faut aussi noter que, contrairement à d'autres domaines de l'intelligence artificielle, celui des agents intelligents s'intéresse à plusieurs formes de comportements intelligents[87]. Les agents intelligents s'intéressent autant aux comportements d'un insecte qu'à ceux d'un être humain.

Russell et Norvig regroupent les agents en cinq classes distinctes dépendantes de leurs compétences : les agents à réflexes simples, les agents à réflexes basés sur des modèles, les agents orientés par des objectifs, les agents utilitaristes et les agents d'apprentissage[88]. Ces différentes structures représentent des approches d'interaction avec l'environnement. Les agents à réflexes simples (figure 4) agissent presque instantanément lorsqu'ils perçoivent la présence d'un stimulus, ignorant les perceptions antérieures. Ils sont contrôlés par un programme déterministe et, un peu comme les systèmes experts, ils fonctionnent selon des règles de type *condition → action*. Comme ils ne conservent aucune information des états antérieurs de l'environnement, ces agents sont efficaces surtout dans un environnement entièrement observable. Ils risquent de s'embourber dans des boucles infinies de répétitions d'actions s'ils doivent fonctionner dans un environnement partiellement observable.

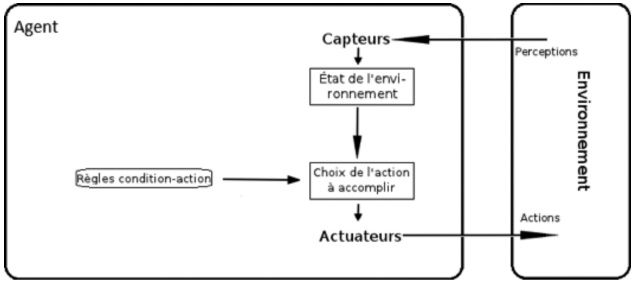

FIGURE 4 – Un agent à réflexes simples

Les agents à réflexes basés sur des modèles (figure 5) gardent en mémoire leur état actuel. Ce type d'agent dispose également d'un modèle du monde décrivant « les règles régissant son environnement ». Ce modèle permet à ce type d'agent de fonctionner dans un univers partiellement observable et de prévoir, dans une certaine limite, les aspects invisibles de l'état actuel. Le choix des actions à faire se fait également selon des règles de type *condition → action*.

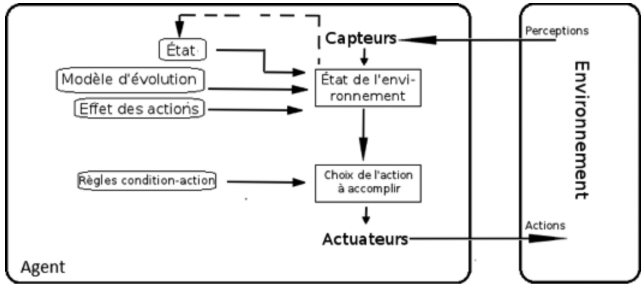

FIGURE 5 – Schéma d'un agent à réflexes basés sur des modèles

Il est possible de diriger les actions d'un agent intelligent en lui indiquant un objectif à atteindre. Cela permet à un agent orienté par objectifs (figure 6) de choisir, parmi plusieurs choix possibles, l'action qui lui permettra d'atteindre l'état visé. Plus un objectif est complexe à atteindre plus ce type d'agent devra inclure des compétences de recherche et de planification. Différents systèmes peuvent être mis en place pour évaluer la réussite ou les erreurs dans l'atteinte des objectifs[89].

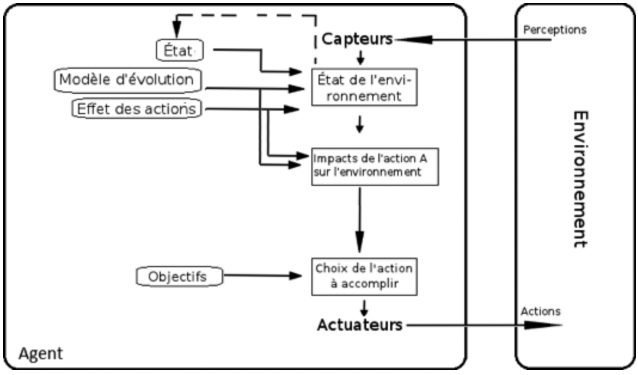

FIGURE 6 – Schéma d'un agent orienté par objectifs

Les agents utilitaristes (figure 7), quant à eux, associent un degré de désirabilité à un état particulier. Cela peut être obtenu en comparant différents états possibles de l'environnement grâce à une fonction d'utilité. Un agent utilitariste doit modéliser et conserver en mémoire l'évolution de son environnement. Bien entendu, une telle tâche nécessite des capacités de perception, de représentation et de raisonnement supérieures à celles des agents décrits plus haut.

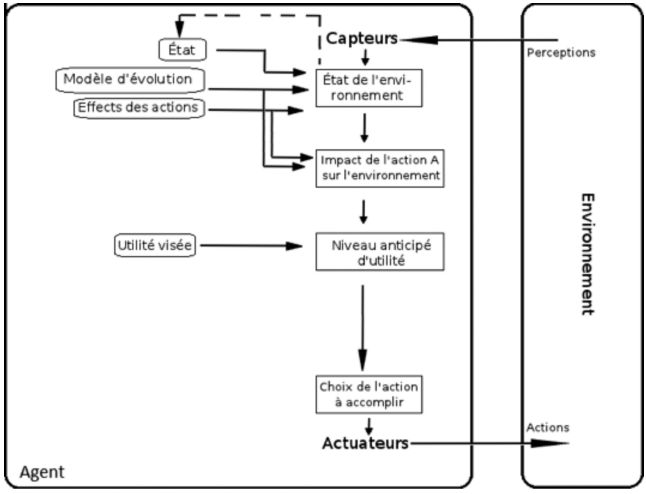

FIGURE 7 – Schéma d'un agent utilitariste

Un agent d'apprentissage (figure 8) peut commencer à agir dans un environnement inconnu. À mesure qu'il emmagasine de l'information et des connaissances, il devient de plus en plus compétent, allant au-delà de ces capacités initiales. Parmi ses composants, on compte un élément d'apprentissage, responsable de proposer des améliorations et un élément de performance, chargé de sélectionner les actions à faire. Ces deux éléments fonctionnent en interaction. L'élément d'apprentissage se sert de la rétroaction produite par un élément de critique pour évaluer les performances de l'agent et pour proposer des améliorations à l'élément de performance. Finalement, les agents d'apprentissage sont composés d'un dernier élément : le générateur de problèmes. Contrairement aux améliorations proposées par l'élément d'apprentissage, les nouvelles actions proposées par le générateur de problèmes ne cherchent pas

à augmenter les performances de l'agent, mais plutôt à créer de nouvelles connaissances. Dans son processus d'interaction avec l'élément d'apprentissage, l'élément de performance transmet les nouvelles connaissances extraites des actions précédentes de l'agent.

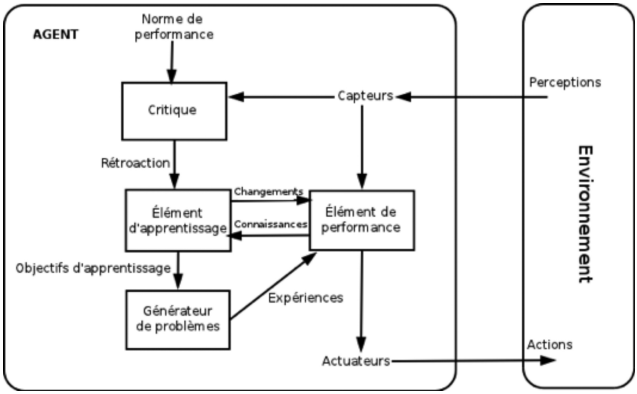

FIGURE 8 – Un agent d'apprentissage

Cette nouvelle façon d'approcher les problèmes en intelligence artificielle que sont les agents intelligents permet d'étudier des sous-problèmes de façon isolée, ou hors de leur contexte global, et avec des outils spécialisés selon la complexité des sous-problèmes afin de trouver des solutions vérifiables et utiles[90]. Aujourd'hui, afin d'effectuer des fonctions complexes, les agents intelligents sont souvent rassemblés dans une structure hiérarchique contenant plusieurs sous-agents. On appelle cette structure système multi-agents. Les sous-agents intelligents peuvent accomplir des fonctions de bas niveau alors que leurs interactions permettent au système complet d'accomplir des tâches difficiles et d'atteindre des objectifs complexes.

Dans un effort pour mettre en place les Trois Lois de la robotique, les agents intelligents semblent offrir une avenue très intéressante. Il faut cependant éviter de tomber dans le piège de la complexité et penser qu'un seul agent complexe serait suffisant pour mettre en œuvre les Trois Lois. Bien que les agents les plus complexes, comme les agents d'apprentissage et les agents utilitaristes, peuvent donner l'impression d'être les mieux adaptés pour ce genre d'application, l'étendue des activités de recherche et de planification rendrait très probablement ces agents inefficaces lorsqu'une décision devrait être prise rapidement. Comme ces activités impliquent des notions de dangers et de risques à l'intégrité physique des robots et des êtres humains, la mise en œuvre de la Loi 1 et, dans une moindre mesure, de la Loi 3 devrait être sous-divisée en plusieurs sous-agents et la majorité de ceux-ci devraient être de type à réflexes simples ou à réflexes basés sur des modèles afin d'assurer la rapidité d'action nécessaire pour éviter un danger. Il serait donc nécessaire de modifier les Lois d'Asimov pour y inclure une notion d'urgence.

Dans une mise en œuvre robotisé, un système multi-agents peut être soit un seul robot dont le système de contrôle, d'analyse et de prise de décision est composé de plusieurs agents, soit une équipe de robots ayant chacun un mode de fonctionnement différent, mais travaillant tous pour atteindre un ou plusieurs objectifs communs. Peu importe le type de mise en œuvre choisi, les compétences sensorielles, d'interprétation, de décision et d'action, ainsi que la compétence de rétention et de recouvrement des connaissances dépendront directement de la puissance de calcul des microprocesseurs concernés et de l'avancement des capteurs et des actuateurs installés.

3.3.6 Les avancées en microélectronique

Bien qu'il ne s'agisse pas directement d'un outil de l'intelligence artificielle, la microélectronique a eu un impact gigantesque sur l'IA. La vitesse de son développement est décrite par la loi de Moore qui indique que le nombre de transistors sur un circuit intégré double tous les deux ans[91] (figure 9). Aujourd'hui, nous pouvons lier cette loi à la plupart des propriétés des composants électroniques, comme les mémoires, les capteurs, la vitesse des processeurs, la capacité des caméras numériques, etc. Cette loi a été l'un des principaux moteurs du développement des technologies de miniaturisation et d'intégration. Ces avancées ont permis d'augmenter les performances des dispositifs tout en conservant les coûts de production relativement bas. En 1971, la fabrication d'un transistor coûtait environ 10 cents, alors qu'aujourd'hui on parle d'un coût inférieur à un millième de cent.

Si l'on se fie à cette logique, on comprend que l'ordinateur Deep Blue, qui a battu le champion d'échecs Garry Kasparov en 1997, était plus de 10 millions de fois plus rapide que le Ferranti Mark 1 (1951) sur lequel a été programmé le premier jeu d'échecs. À titre d'illustration, en 1997, Deep Blue utilisait 32 processeurs en parallèle, chacun cadencé à 120 MHz, alors qu'en 2011, pour le quiz *Jeopardy !*, Watson utilisait 2 880 processeurs en parallèle, chacun cadencé à 3,5 GHz, ou 3 500 MHz.

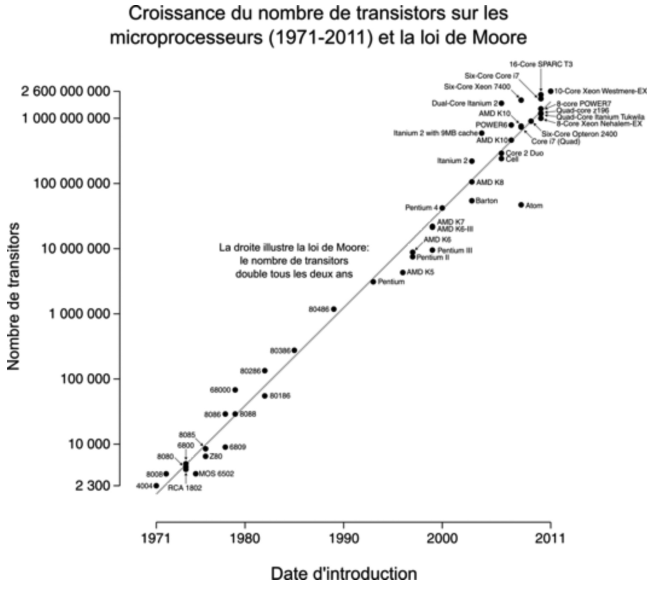

FIGURE 9 – Illustration de la loi de Moore

Certains auteurs, comme Raymond Kurzweil, sont persuadés que le développement de la force brute des ordinateurs est l'unique facteur qui a permis à des machines très spécialisées de battre des champions humains à des jeux de stratégie comme les échecs ou même des quiz de connaissances générales comme *Jeopardy !*[92] Selon Kurzweil, les progrès en matière de microélectronique accélèrent à un rythme tel que nous devrions assister à la création d'une véritable intelligence artificielle d'ici moins de 50 ans.

Cependant, dans son raisonnement, Kurzweil suppose que la loi de Moore sera encore valide pour les 50 prochaines années. Or, cela est loin d'être certain. Afin de s'assurer que la puissance des technologies offertes double tous les deux

ans, les fabricants de microprocesseurs se sont tournés vers la miniaturisation des composants électroniques et ont transféré la production des micro-puces vers les pays émergents où les salaires sont beaucoup plus bas qu'en Occident. Cependant, on ne pourra pas miniaturiser les composants de façon infinie. Fondamentalement, on ne peut pas avoir d'unité de mémoire plus petite qu'un atome unique. On estime aussi qu'il serait possible de repousser les dimensions critiques des transistors au-delà de quelques nanomètres. Mais, bien avant d'atteindre cette limite physique, il est très probable que nous atteignons la limite de la rentabilité économique de la miniaturisation des micro-puces. Le modèle de développement des fabricants de composants électroniques est d'ailleurs en plein changement. Les ressources placées jusqu'à maintenant dans cette course à la performance sont tranquillement transférées vers l'exploitation de nouvelles fonctionnalités. Ainsi, on ne peut plus s'attendre à ce que la puissance de calcul des microprocesseurs augmente aussi rapidement qu'avant, mais il est possible que de nouvelles puces, dédiées par exemple aux fonctions de raisonnement complexe et de raisonnement de sens commun, arrivent sur le marché dans les prochaines années et facilitent le développement de robots éthiques.

3.4 Les agents éthiques

Une question demeure : est-ce possible de mettre en place des règles éthiques ou morales sur lesquelles seraient basées les décisions et les actions des robots ? Avec les dernières avancées en intelligence artificielle, c'est maintenant possible, du moins dans une certaine mesure. Les agents éthiques sont des programmes informatiques, des robots ou des machines plus simples qui ont la capacité de poursuivre un idéal basé sur un principe éthique ou un ensemble de principes[93]. On remarque que la notion d'agent va au-delà de celle de l'agent intelligent décrit plus haut. Ici, le terme

agent indique que la machine ou le programme effectue une tâche au nom d'un utilisateur, de la même façon qu'un agent immobilier effectue une transaction immobilière au nom d'un vendeur ou d'un acheteur. On peut distinguer quatre types d'agents éthiques : l'agent à impact éthique, l'agent éthique implicite, l'agent éthique explicite et l'agent éthique complet[94].

L'agent à impact éthique découle de l'évaluation d'une technologie en fonction de normes éthiques. En effet, nous sommes habitués à évaluer les technologies selon des normes de conception, voire des normes socioéconomiques, or ces technologies ont aussi des impacts de nature éthique. Selon cette définition, Internet, par exemple, est un agent à impact éthique. Il permet aux utilisateurs de rester informés en leur donnant accès à une banque de connaissances presque infinie. En contrepartie, il est également non éthique parce qu'il facilite la fraude et le vol d'identité.

L'agent éthique implicite, quant à lui, est utile pour des tâches respectant certains principes éthiques définis par ses concepteurs. Un guichet automatique, le pilote automatique d'un avion et le tableau de bord d'une voiture peuvent entrer dans cette catégorie. Le premier est conçu pour transférer les bons montants d'argent tout en respectant les données personnelles des utilisateurs. Le pilote automatique est conçu pour transporter les passagers d'une ville à une autre en assurant leur sécurité, alors que le tableau de bord d'une automobile informe le conducteur de défaillances potentielles pouvant mettre sa sécurité en jeu ou pouvant nuire à son confort, une panne d'essence par exemple. Bien que les agents éthiques implicites ne prennent pas de décision en se basant sur des principes éthiques, ils ont été conçus de façon à ce que leurs fonctionnalités respectent une valeur éthique.

Les robots industriels, présentés au début de ce chapitre, sont programmés pour prévenir les événements qui

pourraient s'avérer dangereux pour l'humain. Ils disposent aussi d'un ensemble de systèmes de sécurité afin d'éviter les chocs et les collisions avec leur environnement. Puisqu'il est clair que ces systèmes sont conçus pour minimiser les risques de blessure aux humains, on peut les inclure dans le groupe d'agents éthiques implicites.

Dans une situation de dilemme éthique, les agents éthiques explicites ont la capacité de calculer la meilleure chose à faire en fonction de principes éthiques. Bien que le développement dans ce domaine n'en soit qu'à ses débuts, quelques groupes de recherche ont réussi à mettre en place des modèles de décision éthique dans des agents éthiques explicites rudimentaires. Un des plus efficaces est probablement le modèle des tâches *prima facie*, ou de première importance, développé par M. Anderson, S.L. Anderson et C. Armen[95]. Ainsi, plutôt que d'avoir une liste de tâches absolues et possédant un rang de priorité différent à exécuter, leurs agents disposent d'une liste de tâches qu'ils devraient respecter. Or, comme il n'existe aucune priorité entre les tâches *prima facie*, plusieurs groupes[96] proposent une technique de l'intelligence artificielle, l'apprentissage machine, pour mettre en place les principes éthiques nécessaires à la prise de décision lorsque deux ou plusieurs de ces tâches entrent en conflit[97]. Leur algorithme analyse un nombre représentatif de cas dans lesquels des humains doivent prendre des décisions éthiques. Le système transpose alors les principes éthiques sous-jacents en langage logique qui peut par la suite être encodé dans un robot. Grâce à ce système, ils ont réussi à mettre en œuvre un modèle de décision éthique pour des robots d'assistance pour les centres de personnes âgés.

Plus leur développement sera avancé et plus les robots militaires entreront dans cette catégorie. Ils ne peuvent être des agents éthiques implicites, puisque toutes les situations

ne peuvent pas être anticipées, donc toutes les décisions éthiques ne pourront pas être programmées[98]. Comme leur tâche principale est de détruire une cible, ils sont équipés d'armement divers. Cependant, ces robots peuvent avoir des tâches secondaires : protéger les alliés et les civils. On peut très facilement anticiper qu'ils se retrouveront dans des situations où ces décisions d'ordre éthique devront être prises très rapidement. Par exemple, un robot pourrait se retrouver en présence d'une cible souhaitée et identifiée, mais située dans un environnement où des civils ou des alliés sont présents. En tant qu'agent éthique explicite, le robot devrait être capable de juger des impacts de chaque possibilité pour déterminer quelle tâche est prioritaire.

Finalement, Moor[99] définit l'agent éthique complet comme étant capable de faire des jugements éthiques et de les justifier, alors qu'un adulte moyen peut être considéré comme un agent éthique complet. Selon lui, il est clair qu'aujourd'hui aucune machine ne pourrait être classée dans cette catégorie. D'abord, notre compréhension de la théorie éthique n'est pas suffisamment complète pour mettre en place un formalisme applicable dans un contexte général. Ensuite, malgré les avancées récentes en apprentissage machine, les outils disponibles actuellement ne sont pas suffisamment performants pour permettre un apprentissage à grande échelle. Finalement, Moor souligne l'incapacité des systèmes intelligents actuels à résoudre des problèmes de sens commun et l'absence de connaissances du monde réel. On « pourrait programmer une machine avec l'impératif classique des médecins et des robots asimoviens : d'abord, ne cause pas de tort. Mais cela n'aurait aucune utilité à moins que la machine comprenne ce que constitue un tort dans le monde réel[100]. »

CONCLUSION

Avec le développement de la robotique, on a vu croître l'utilisation des robots, à la fois dans les milieux industriels et dans le secteur militaire. Dans les prochaines années, ces robots auront de plus en plus d'autonomie et devront prendre des décisions qui auront des impacts certains sur les humains qui les côtoient. Dans ce contexte, il est important que les développeurs se questionnent sur la réalisation de robots éthiques. De tels robots devraient être en mesure de choisir les tâches à effectuer en se basant sur des principes moraux, d'une façon analogue à celle des robots d'Asimov.

Parce qu'elle offre une critique de notre société technologique, en établissant les Trois Lois de la robotique, la façon dont Asimov met en œuvre sa morale des robots est très pertinente. Cependant, pour y arriver, Asimov a dû déployer un *novum*, ou un univers scientifique et technologique fictif, très différent du nôtre. Dans son œuvre, les robots ont des capacités sensorielles et motrices ultra performantes, beaucoup plus près de celles de Superman que de celles d'un robot réel. Ils ont également un pouvoir d'analyse et de raisonnement presque infaillible. Tout cela illustre l'univers purement déterministe dans lequel Asimov incarne ses nouvelles et ses romans du cycle des robots. C'est ce contexte qui permet aux Trois Lois de la robotique d'être si puissantes et si efficaces.

Or, notre monde réel est extrêmement différent de l'univers d'Asimov. Notre univers n'est pas entièrement déterministe. En raison de la mécanique quantique, l'information pour décrire un état ou un événement est fondamentalement incomplète. Technologiquement, tous les capteurs introduisent des incertitudes dans les mesures et, bien qu'il soit possible de quantifier ces incertitudes, elles limitent tout de même l'information qu'il est possible d'en acquérir.

Lors de prédictions, en appliquant des modèles physiques, chimiques, psychologiques ou sociologiques, ces incertitudes se multiplient en augmentant le nombre d'éléments en interaction dans un environnement. Cela vient souligner le caractère chaotique de l'information et limite la portée des prédictions possibles.

Pour ces nombreuses raisons, il est impossible de réaliser des robots asimoviens comme Daneel et Giskard. Cependant, il existe des outils en robotique et en intelligence artificielle qui nous permettent d'entrevoir la possibilité de réaliser des robots éthiques dans une approche plus pragmatique. Contrairement aux robots asimoviens, les robots éthiques réels ne seraient pas moralement infaillibles et contreviendraient souvent aux Trois Lois de la robotique.

Un robot éthique doit disposer de quatre types de compétences : des compétences sensorielles, des compétences d'interprétation, des compétences décisionnelles et des compétences motrices. En plus, il doit disposer d'un type de compétences transversales de rétention, d'analyse et de recouvrement de l'information. Avec le développement de la robotique mobile et de l'intelligence artificielle, nous avons maintenant des outils qui nous permettent d'accroître, dans une certaine mesure, ces compétences. On a démontré qu'il était possible d'utiliser des outils logiques pour programmer des systèmes de raisonnement machine relativement complexes. D'autres outils, comme la logique floue, les systèmes experts, les systèmes de bases de connaissances et les agents intelligents, se basent entre autres sur ces outils logiques et l'apprentissage machine pour améliorer et faciliter le développement des compétences d'interprétation, de décision et de gestion de l'information. Les avancées rapides en microélectronique auxquelles nous assistons depuis les 40 dernières années ont également permis d'augmenter extraordinairement la puissance de calcul des

microprocesseurs qui supportent ces outils intelligents et d'accroître leur capacité de rétention de l'information. Aujourd'hui, les superordinateurs utilisés dans les systèmes intelligents tels que Watson d'IBM ont une puissance de calcul 25 milliards de fois plus élevée que celle des ordinateurs sur lesquels ont été développés les premiers systèmes intelligents.

Malgré ces avancées, les outils de l'intelligence artificielle sont victimes de limites fondamentales et technologiques, qui restreignent leur utilisation pour la résolution de problèmes réels et complexes. L'absence de connaissances de sens commun nuit à la capacité d'un robot d'interpréter son environnement. Cette limite est cruciale, puisqu'il est impossible d'inculquer toutes ces connaissances capitales par la simple programmation. L'explosion combinatoire est également une des limites les plus importantes. Alors que les problèmes se complexifient, la quantité d'information en entrée et le nombre de solutions à évaluer croissent exponentiellement. L'augmentation de la puissance de calcul des microprocesseurs et la fragmentation des problèmes complexes grâce aux agents intelligents peuvent permettre de pallier légèrement cette contrainte. Toutefois, ces limites ont un impact limitatif important lorsque vient le temps de résoudre des problèmes se déroulant dans un environnement réel et non contrôlé, comme celui à l'intérieur duquel les robots éthiques seront appelés à évoluer.

Malgré ces limites, certains chercheurs ont démontré qu'il était possible de réaliser des robots capables de prendre des décisions basées sur des principes éthiques. Dans presque tous les cas, l'acquisition de ces principes doit se faire par un processus d'apprentissage machine. Une fois l'apprentissage terminé, ces robots peuvent choisir parmi un ensemble de tâches programmées celle qui est la plus utile

ou celle qui balance le mieux une sélection limitée de principes éthiques.

Ces robots sont probablement les premiers d'une série de robots éthiques qui serviront de prototypes dans le but d'intégrer des capacités de décision éthique aux prochaines générations de robots commerciaux. Ils permettront aux robots industriels d'assurer une plus grande sécurité aux travailleurs qui les côtoient et permettront aux robots militaires de choisir les meilleures tâches à faire en évaluant leurs impacts. Bien entendu ces robots ne fonctionneront jamais en se basant sur un ensemble de Trois Lois simples et élégantes, comme les robots asimoviens. Ils feront aussi de nombreuses erreurs qui porteront atteinte à leur propre intégrité, qui contreviendront aux ordres des humains et qui, certaines fois, blesseront ou tueront les humains. Ces robots, comme tous ceux qui les auront précédés, continueront d'accomplir pour nous, et de manière sécuritaire, des tâches difficiles, dangereuses et monotones. Cependant, contrairement à leurs prédécesseurs, ils auront la capacité de refuser d'accomplir une tâche ou pourront choisir les tâches à accomplir en se fondant sur des raisonnements éthiques complexes qui sont chers à notre société humaine.

Conclusion

Jean-Pierre Béland et Georges A. Legault

Dans cet essai, nous avons produit des analyses, en suivant chacun des trois grands axes de questionnement liés à l'œuvre de science-fiction d'Asimov : 1) la question du vivre-ensemble avec des robots ; 2) la question de la moralité des robots ; 3) la question de la faisabilité du robot moral pour aujourd'hui. Le but de ces analyses était d'en cerner les enjeux E^3LS servant à penser le plus finement possible l'acceptabilité ou non des robots humanisés.

Le premier chapitre nous a permis de vulgariser nos travaux InterNE^{3}L sur des questions complexes au sujet de l'acceptabilité éthique des impacts de ce développement technologique. Ces questions nous interrogent sur notre responsabilité et, plus largement, sur notre humanité. Les principaux personnages dans l'ensemble des récits chez Asimov nous ont permis d'articuler une position mitoyenne entre l'optimisme de l'U.S. Robots et des Spaciens qui prônent l'acceptation inconditionnelle des robots et le pessimisme des Terriens médiévalistes qui militent pour le refus inconditionnel des robots. La transformation de Baley dans sa relation avec les robots comme Daneel et Giskard est au cœur de cette compréhension de la position mitoyenne de l'acceptabilité par l'analyse globale des impacts sur des enjeux. La pondération (équilibre) des impacts (positifs et négatifs) et des jugements finaux de valeur maximisant les impacts positifs chez Asimov permet de défendre cette position mitoyenne qui va dans le sens du choix de Baley et

du projet de Fastolfe pour l'expansion et l'évolution. C'est la seule conclusion que nous permet de faire cette analyse globale d'impact et d'acceptabilité dans la science-fiction d'Asimov. Cette analyse remet en question alors nos façons d'agir, notamment les relations personnelles avec des robots (amitié, dépendance, etc.). Aussi, derrière ces questions se pose toujours celle de l'identité humaine et de la perte de l'identité humaine. Étant donné que des robots deviendront de plus en plus humanoïdes, il y a un problème d'identité qui s'annonce, en ce sens que la frontière entre l'humain et le robot tend à s'estomper selon différents points de vue : celui de la composition ou fonction (apprentissage, autonomie de raisonnement, créativité) du cerveau, celui du corps biologique opposé au corps de métal et celui de la relation à l'autre (réciprocité éthique). Un robot pourra-t-il ainsi faire partie de l'humanité ? Des clivages pourront aussi exister entre des sociétés sans robots et des sociétés avec robots.

Le second chapitre a montré en quoi la morale de la robotique d'Asimov permet de comprendre les défis de construire la raison pratique d'un robot et aussi de clarifier nos propres jugements sur l'acceptabilité éthique du développement technologique. Les points importants à retenir sont les suivants :

1) La science-fiction d'Asimov nous invite à dépasser les insuffisances des Trois Lois morales en les inscrivant dans une approche éthique. Contrairement à d'autres romans de science-fiction, Asimov ne nous fait pas la morale, mais nous fait comprendre la complexité du raisonnement moral. Dans l'application des Trois Lois de la robotique, les robots sont non seulement amenés à comprendre l'ordre donné, mais encore, pour les robots les plus sophistiqués, à en comprendre le sens afin de pouvoir l'appliquer. En situant les Lois morales en fonction des valeurs en jeu que nous voulons atteindre

dans nos vies individuelles et collectives, la pensée d'Asimov rejoint tout le courant de l'éthique appliquée qui s'est développé depuis plus de cinquante ans. Il n'y a pas de solutions toutes faites lorsqu'il s'agit de vivre en tension entre des valeurs (enjeux) que nous aimerions vivre totalement, mais qui sont irréconciliables. Comme dans la nouvelle « La vie et les œuvres de Multivac », Asimov illustre bien ce conflit permanent entre deux valeurs qui traversent plusieurs de nos choix individuels et collectifs : celui entre la sécurité et l'autonomie (liberté). Il n'y a pas une solution à cette question, mais des choix responsables en contexte.

2) Il y a aussi un lien important entre la morale des robots et la morale humaine chez Asimov. Les Quatre Lois renvoient aux quatre principes moraux suivants : *Primum non nocere*, morale de l'obéissance, altruisme et utilitarisme. Nos morales des Trois Lois concernent les individus et l'impact de nos actions sur les individus autour de nous. Mais les Trois Lois ne peuvent pas tenir compte des actions qui nuisent ou nuiront à l'humanité. La Loi Zéro d'Asimov est l'équivalent du principe responsabilité de Hans Jonas. Nous sommes responsables de la vie sur terre et du sort des générations futures. Il faut donc, dans nos décisions, tenir compte de répercussions sur les humains et aussi sur l'humanité à venir.

3) Plusieurs des nouvelles chez Asimov nous ont permis de montrer la complexité du raisonnement moral. Cela nous amène à penser que, si ce raisonnement est complexe pour le robot, il l'est encore plus pour l'être humain qui n'a pas un programme initial. Dans les choix difficiles, Asimov met en dialogue soit des robots avec les humains, soit deux robots entre eux afin d'explorer les diverses possibilités du meilleur choix. À l'aide des nouvelles d'Asimov, ne pourrions-nous donc pas nous préoccuper

aujourd'hui de former les jeunes à la complexité du raisonnement pratique qu'est la décision ?

Le troisième chapitre soulève l'enjeu déterminant de la faisabilité du robot qui parvient à cette complexité du raisonnement moral pour penser l'acceptabilité du développement technologique aujourd'hui. Est-il seulement possible d'appliquer les Lois de la robotique à une intelligence artificielle pour que celle-ci fasse les meilleurs choix éthiquement acceptables ? Les robots moraux, comme Daneel et Giskard, sont-ils ou seront-ils un jour réalisables ? Le problème est que la vision d'un univers déterministe, comme celle qui est présentée par Asimov, ne cadre pas avec notre compréhension de notre univers réel. Dans un tel univers déterministe, tout peut être prévisible, de sorte que le robot devient parfaitement moral. Mais que peut-on faire dans notre univers indéterministe pour donner à des robots les moyens nécessaires pour s'approcher de l'esprit des Trois Lois de la robotique ? La difficulté dans un tel univers indéterministe est qu'un être moralement autonome est souvent imprévisible, puisqu'il estime lui-même ce qui lui semble juste. Sa décision va plutôt dans le sens de la complexité du raisonnement pratique de l'éthique. Dans cette perspective, le robot éthique ne serait pas moralement infaillible dans l'application de sa décision. Mais celle-ci apparaîtrait alors comme la meilleure solution établie par la raison pratique, comme solution aux tensions et aux conflits.

La science-fiction d'Asimov, comme il l'a lui-même déclaré, est un moment qui permet de réfléchir le développement technologique et de mettre en scène les impacts positifs et négatifs sur les humains et leur manière de vivre ensemble. En tant que fiction, elle repose sur un potentiel imaginaire de la science, comme on le voit très bien avec les robots Giskard, Daneel et Andrew. Mais, en tant que roman, elle nous invite à réfléchir à ce qui est possible avant qu'il ne

devienne probable. Nous sommes au début de la création de robots qui vont interagir avec les humains au quotidien, pour assurer la surveillance et la sécurité des personnes âgées, par exemple. Tout comme se développent des implants de plus en plus perfectionnés combinant, dans les termes d'Asimov, les éléments C/Fe qui symbolisent la matière incorporée au vivant. Ainsi, comme dans les romans d'Asimov, nous sommes, en tant que Terriens, aux prises avec ces développements technologiques et c'est à nous de faire les meilleurs choix possibles pour l'avenir de l'humanité.

Notes

INTRODUCTION

1. www.INTER-NE3LS.org.
2. Fritz Allhoff, Patrick Lin, James Moor et John Weckert (2009), *Ethics of Human Enhancement : 25 Questions & Answers*, US National Science Foundation, p. 6.
3. *Ibid*., p. 92.
4. *Ibid*.
5. Georges Vignaux (2010), *La chirurgie moderne ou l'ivresse des métamorphoses. La chirurgie esthétique. La chirurgie réparatrice. Les prothèses et les robots*, Paris, Pygmalion, p. 83.
6. Isaac Asimov (1967), « Raison », *Le cycle des robots 1. Les robots*, Paris, Éditions J'ai lu, p. 86.
7. Isaac Asimov (1978), « Préface », *David Starr, justicier de l'espace*, Éditions Lefrancq, p. 3.
8. Isaac Asimov (1967), « Préface », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 14.
9. *Ibid*., p. 10-11.
10. Cf. Isaac Asimov (1996), *Moi Asimov*, Paris, Éditions Denoël, p. 548.
11. *Ibid*.
12. Isaac Asimov (1988), *Le robot qui rêvait*, Paris, Éditions J'ai lu, p. 7-8.
13. Isaac Asimov (1967), *Le cycle des robots 1. Les robots*, *op. cit.*, p. 5.
14. http://fr.wikipedia.org/wiki/Trois_lois_de_la_robotique, consulté le 5 mars 2012.
15. Isaac Asimov (1988), « Préface », *Le robot qui rêvait*, *op. cit.*, p. 8.
16. *Ibid*., p. 8-9.
17. http://www.maxisciences.com/robot/ce-robot-sait-exprimer-des-émotions_art2591.html ; cf. Georges Vignaux (2010), *La chirurgie moderne ou l'ivresse des métamorphoses…*, *op. cit*., p. 74 et 84.
18. *Ibid*., p. 86.
19. Le premier chapitre permettra au lecteur de mieux comprendre ce cadre de référence des enjeux E3LS. Les avis de la Commission de l'éthique de la science et de la technologie (CEST) du Québec sur les OGM et sur les nanotechnologies nous invitent à approfondir la réflexion sur la thématique des impacts sur ces enjeux E3LS. À partir de cette thématique, nous ne nous questionnons pas seulement sur ce que seront les bons choix politiques ou économiques du développement technologique pour l'amélioration humaine, mais sur ce qui permettra de respecter les équilibres écologiques ou sur ce qui garantira le mieux la santé publique. Le public ignore souvent que NanoQuébec, l'Institut de recherche Robert-Sauvé en santé et sécurité au travail (IRSST), le Fonds québécois de la recherche sur la société et la culture, le Fonds québécois de la recherche sur la nature et les technologies (FQRNT) et le Fonds de la recherche en santé du Québec (FRSQ) ont convenu d'élaborer une stratégie de recherche sur ces

enjeux E3LS. Le programme « Bourse thématique NE3LS » permet de soutenir le développement de la recherche sur cette thématique qui prend la forme de maîtrise et de doctorat. Ces bourses sont offertes à des candidats qui développent des recherches interdisciplinaires. [En ligne] www.fqrsc.gouv.qc.ca/fr/bourses/programme.php ?id_programme.

CHAPITRE 1

1. L'élaboration de ce processus est au cœur de notre programme de recherche d'InterNE3LS subventionné par les IRSC (2009-2014) sous la direction de Johane Patenaude.
2. F. Terrade et autres (2009), « L'acceptabilité sociale : la prise en compte des déterminants sociaux dans l'analyse de l'acceptabilité des systèmes technologiques », *Le travail humain*, Presses universitaires de France, 4, vol. 72, p. 384, [en ligne] http://www.cairn.info/revue-le-travail-humain-2009-4-page-383.htm.
3. Jean-Pierre Béland et autres (2011), « The Social and Ethical Acceptability of NBICs for Purposes of Human Enhancement : Why Does the Debate Remain Mired in Impasse ? », *Nanoethics*, 5, p. 295-307. Cet article peut être consulté en ligne : Springerlink.com.
4. Georges A. Legault et autres (2012, soumis), « Nanotechnologies and Ethical Argumentation : A Philosophical Stalemate ? », *Nanoethics*.
5. Isaac Asimov (1996), *Moi, Asimov*, Paris, Éditions Denoël, p. 248.
6. Isaac Asimov (1967), « Préface », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 11.
7. Comme nous le verrons un peu plus loin, dans notre cadre conceptuel d'analyse globale d'impact et d'acceptabilité, nous parlons d'un « risque théorique » pour signifier qu'il n'y a aucune connaissance scientifique pour conclure à l'inexistence de la relation entre la source (le développement des robots dangereux) et l'impact négatif ou positif sur l'enjeu (la vie humaine, par exemple).
8. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, Paris, Éditions J'ai lu, p. 277.
9. *Ibid*., p. 11.
10. Isaac Asimov (1988), « Introduction », *Le robot qui rêvait*, Paris, Éditions J'ai lu, p. 19.
11. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 232.
12. *Ibid*., p. 237-238.
13. *Ibid*., p. 233.
14. Isaac Asimov (1967), « Préface », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 12.
15. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 265.
16. Isaac Asimov (1967), « Le correcteur », *Le cycle des robots 2. Un défilé de robots*, Paris, Éditions J'ai lu, p. 198.
17. Isaac Asimov (1967), « Préface », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 14.
18. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 235-236.
19. *Ibid*., p. 231.
20. Isaac Asimov (1967), « Le petit robot perdu », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 174.

21. *Ibid*., p. 209.
22. Isaac Asimov (1978), « Pour que tu t'y intéresses », *L'homme bicentenaire*, Paris, Éditions Denoël, p. 143.
23. *Ibid*., p. 149.
24. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 5.
25. *Ibid*., p. 23.
26. *Ibid*., p. 24.
27. *Ibid*., p. 372.
28. *Ibid*., p. 373.
29. *Ibid*., p. 75.
30. *Ibid*., p. 76.
31. www.inter-NE3LS.ORG.
32. Johane Patenaude (2011), « Moral arguments in the debate over nanotechnologies : Are we talking past each other ? », *Nanoethics*, 5. Cet article peut être consulté en ligne : Springerlink.com.
33. *Ibid*.
34. Isaac Asimov (1988), « Introduction », *Le robot qui rêvait*, *op. cit.*, p. 18-19.
35. Isaac Asimov (1970), *Le cycle des robots 4. Face aux feux du soleil*, *op. cit.*, p. 184.
36. *Ibid*., p. 183.
37. Isaac Asimov (1975), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 92.
38. Isaac Asimov (1978), « L'homme bicentenaire », *L'homme bicentenaire*, *op. cit.*, p. 284-285.
39. *Ibid*.
40. Isaac Asimov (1990), « *Prélude à Trantor* », *Le grand livre des robots*, Paris, Omnibus, p. 123-124.
41. Isaac Asimov (1967), « Robbie », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 29.
42. *Ibid*., p. 30.
43. Isaac Asimov (1967), « Évidence », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 281.
44. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, Paris, Éditions J'ai lu, p. 566-567.
45. Isaac Asimov (1967), *Le cycle des robots 1. Les robots*, *op. cit.*, p. 53-54.
46. Isaac Asimov (1967), *Le cycle des robots 2. Un défilé de robots*, *op. cit.*, p. 168.
47. Isaac Asimov (1978), « Pour que tu t'y intéresses », *L'homme bicentenaire*, *op. cit.*, p. 142-143.
48. Isaac Asimov (1978), « Étranger au paradis », *L'homme bicentenaire*, *op. cit.*, p. 163-164.
49. *Ibid*., p. 185.
50. *Ibid*., p. 195.
51. Isaac Asimov (1967), *Le cycle des robots 1. Les robots*, *op. cit.*, p. 318.
52. Isaac Asimov (1978), « La vie et les œuvres de Multivac », *L'homme bicentenaire*, *op. cit.*, p. 199-200.
53. *Ibid*., p. 200.
54. *Ibid*., p. 203-204.
55. *Ibid*., p. 31.

56. *Ibid*., p. 29.
57. Isaac Asimov (1978), « L'incident du tricentenaire », *L'homme bicentenaire*, *op. cit.*, p. 342.
58. Isaac Asimov (1967), « Assemblons-nous », *Le cycle des robots 2. Un défilé de robots*, *op. cit.*, p. 79.
59. *Ibid*., p. 75.
60. *Ibid*., p. 76.
61. Isaac Asimov (1970), *Le cycle des robots 4. Face aux feux du soleil*, *op. cit.*, p. 48.
62. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, *op. cit.*, p. 5.
63. Isaac Asimov (1978), « L'homme bicentenaire », *L'homme bicentenaire*, *op. cit.*, p. 279.
64. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, *op. cit.*, p. 77-78.
65. *Ibid*., p. 15.
66. *Ibid*., p. 28.
67. *Ibid*., p. 27.
68. *Ibid*.
69. Isaac Asimov (1990), « *Prélude à Trantor* », *Le grand livre des robots*, *op. cit.*, p. 121.
70. *Ibid*., p. 124.
71. *Ibid*., p. 125.
72. Isaac Asimov (1967), *Le cycle des robots 2. Un défilé de robots*, *op. cit.*, p. 208.
73. *Ibid*., p. 198-199.
74. *Ibid*., p. 246-247.
75. *Ibid*., p. 247.
76. Isaac Asimov (1978), « La vie et les œuvres de Multivac », *L'homme bicentenaire*, *op. cit.*, p. 202.
77. *Ibid*., p. 213-214.
78. http://fr.wikipedia.org/wiki/La_Vie_et_les_ %C5 %92uvres_de_Multivac, consulté le 15 avril 2012.
79. http://fr.wikipedia.org/wiki/Le_Conflit_ %C3 %A9vitable, consulté le 16 avril 2012.
80. Isaac Asimov (1967), « Conflit évitable », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 314.
81. *Ibid*., p. 318.
82. http://fr.wikipedia.org/wiki/Le_Conflit_ %C3 %A9vitable, consulté le 16 avril 2012.
83. Isaac Asimov (1956), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 372.
84. Isaac Asimov (1970), *Le cycle des robots 4. Face aux feux du soleil*, *op. cit.*, p. 172-173.
85. C'est la Déclaration des droits de la constitution américaine, rédigée par Payne et Jefferson selon la note du traducteur. *Ibid*., p. 173.
86. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, *op. cit.*, p. 76.
87. Isaac Asimov (1984), *Le cycle des robots 5. Les robots de l'aube*, Paris, Éditions J'ai lu, p. 382.
88. Georges Vignaux (2010), *La chirurgie moderne ou L'ivresse des métamorphoses*, *op. cit.*, p. 92.

89. http://fr.wikipedia.org/wiki/Trois_lois_de_la_robotique, consulté le 5 mars 2012.
90. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, *op. cit.*, p. 409.
91. *Ibid.*, p. 411.
92. *Ibid.*, p. 444.
93. *Ibid.*, p. 561.
94. Isaac Asimov (1950), « Évidence », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 270-271.
95. Isaac Asimov (1978), « L'homme bicentenaire », *L'homme bicentenaire*, *op. cit.*, p. 286-287.
96. *Ibid.*, p. 280-281.
97. *Ibid.*, p. 290.
98. Sur ce sujet, voir, entre autres, M. Roco et W.S. Bainbridge (2003), *Converging Technologies for Improving Human Performance : Nanotechnology, biotechnology, information technology and cognitive science*, Dordrecht, Kluwer Academic Publishers.
99. M. Bostrom (2005), « A history of transhumanist thought », *Journal of Evolution and Technology*, vol. 14, n° 1.
100. F. Fukuyama (2002), *Our Posthuman Future. Consequences of the Biotechnology Revolution*, New York, Farrar, Straus and Giroux.
101. F. Fukuyama (2006), *Beyond Bioethics : A Proposal for Modernizing the Regulation of Human Biotechnologies*, Washington DC, School of Advanced International Studies, Johns Hopkins University.
102. C. Bégorre-Bret (2004), « Bioéthique et posthumanité. F. Fukuyama : La fin de l'homme. Les conséquences de la révolution biotechnique ; J. Habermas : L'avenir de la nature humaine. Vers un eugénisme libéral ; D. Lecourt : Humain, post-humain », *Les études philosophiques*, 2, n° 69, p. 253-264. Cet article peut être consulté en ligne : http://www.cairn.info/article.php?ID_REVUE=LEPH&ID_NUMPUBLIE=LEPH_042&ID_ARTICLE=LEPH_042_ 0253.
103. J.-M. Besnier (2009), *Demain les posthumains. Le futur a-t-il encore besoin de nous ?*, Paris, Hachette Littératures, p. 23.
104. M. Coulombe (2009), *Imaginer le posthumain. Sociologie de l'art et d'un vertige*, Québec, Presses de l'Université Laval, p. 11.
105. Fukuyama fait référence ici à l'article bien connu de Bill Joy, « Why the Future Doesn't Need Us » (2000). F. Fukuyama (2002), *Our Posthuman Future. Consequences of the Biotechnology Revolution*, *op. cit.*, p. 6-7.
106. Isaac Asimov (1957), *Le cycle des robots 4. Les cavernes d'acier*, *op. cit.*, p. 92.
107. *Ibid.*, p. 173.
108. Isaac Asimov (1990), « Prélude à Trantor », *Le grand livre des robots*, *op. cit.*, p. 124.
109. *Ibid.*, p. 125.
110. Fritz Allhoff, Patrick Lin, James Moor et John Weckert (2009), *Ethics of Human Enhancement : 25 Questions & Answers*, US National Science Foundation, p. 13.
111. Isaac Asimov (1967), « Lenny », *Le cycle des robots 2. Un défilé de robot*, *op. cit.*, p. 188.
112. Cf. Isaac Asimov (1967), « Le robot Al-76 perd la boussole », *Le cycle des robots 2. Un défilé de robots*, *op. cit.*, p. 13-34.

113. Isaac Asimov (1988), « Ségrégationniste », *Nous les robots, robots métalliques*, *op. cit.*, p. 125.
114. Georges Vignaux (2010), *La chirurgie moderne…*, *op. cit.*, p. 126 ; http://www.ric.org/research/centers/smpp/labs/robotics.
115. Isaac Asimov (1984), *Le cycle des robots 5. Les robots de l'aube*, *op. cit.*, p. 145.
116. http://fr.wikipedia.org/wiki/Psychohistoire_%28Asimov%29, consulté le 6 mai 2012.
117. Isaac Asimov (1988), « Artiste de lumière », *Le robot qui rêvait*, *op. cit.*, p. 304.
118. Isaac Asimov (1967), « Raison », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 88.
119. *Ibid.*, p. 105.
120. *Ibid.*, p. 108.
121. Isaac Asimov (1956), *Le cycle des robots 3. Les cavernes d'acier*, *op. cit.*, p. 304-305.
122. *Ibid.*, p. 373.
123. Isaac Asimov (1970), *Le cycle des robots 4. Face aux feux du soleil*, *op. cit.*, p. 26.
124. Isaac Asimov (1984), *Le cycle des robots 5. Les robots de l'aube*, *op. cit.*, p. 47-48.
125. Isaac Asimov (1986), *Le cycle des robots 6. Les robots et l'empire*, *op. cit.*, p. 272-273.
126. Isaac Asimov (1988), « *Le robot qui rêvait* », *Le robot qui rêvait*, *op. cit.*, p. 31-32.
127. Isaac Asimov (1988), « Sally », *Le robot qui rêvait*, *op. cit.*, p. 178-179.
128. Isaac Asimov (1967), « Raison », *Le cycle des robots 1. Les robots*, *op. cit.*, p. 89.
129. Isaac Asimov (1978), « Pour que tu t'y intéresses », *L'homme bicentenaire*, *op. cit.*, p. 120.
130. Cf. Damien Lagauzère (2008), *Robot : de l'homme artificiel à l'homme synchronique ?*, Paris, L'Harmattan, p. 162.
131. *Ibid.*, p. 148-149.

CHAPITRE 2

1. Isaac Asimov (1990), *Le grand livre des robots 1. Prélude à Trantor*, Paris, Omnibus.
2. Isaac Asimov (1991), *Le grand livre des robots 2*. La gloire de Trantor, Paris, Omnibus.
3. Isaac Asimov (1990), « Raison », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 220.
4. Isaac Asimov (1990), « Le correcteur », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 300.
5. Isaac Asimov (1990), « Évasion », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 385.
6. Isaac Asimov (1990), « Conflit évitable », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 426.
7. *Ibid.*, p. 446.
8. *Ibid.*
9. http://fr.wikipedia.org/wiki/Bombardements_atomiques_de_Hiroshima_et_Nagasaki.
10. http://fr.wikipedia.org/wiki/Crise_des_missiles_de_Cuba.
11. Isaac Asimov (1991), « Cailloux dans le ciel », *Le grand livre des robots 2. La gloire de Trantor*, *op. cit.*, p. 1013.

12. *Ibid*., p. 1014.
13. *Ibid*., p. 1016.
14. Isaac Asimov (1990), « Pour que tu t'y intéresses », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 474.
15. Isaac Asimov (1990), « Lenny », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 286.
16. Isaac Asimov (1990), « La preuve », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 413.
17. *Ibid*.
18. *Ibid*., p. 410.
19. *Ibid*., p. 424.
20. *Ibid*.
21. Isaac Asimov (1990), « Prélude », *Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 7.
22. Isaac Asimov (1990), « Menteur ! », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 268.
23. Isaac Asimov (1990), « Le correcteur », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 300.
24. Isaac Asimov (1990), « Le petit robot perdu », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 337.
25. Isaac Asimov (1990), « Cercle vicieux », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 201.
26. Lester Del Rey (1967), « Une morale pour Sam », *Galaxie*, n° 37, mai, p. 59.
27. Isaac Asimov (1990), « La preuve », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 413.
28. http://en.wikipedia.org/wiki/Three_Laws_of_Robotics#Ambiguities_and_loopholes, consulté le 10 mars 2012.
29. Isaac Asimov (1967), *Le cycle des robots 1. Les robots*, Paris, Éditions J'ai lu, p. 5.
30. Lester Del Rey (1967), « Une morale pour Sam », *Galaxie*, *op. cit.*, p. 59.
31. http://en.wikipedia.org/wiki/Primum_non_nocere, consulté le 5 mars 2012.
32. http://www.alyabbara.com/museum/medecine/pages_01/Serment_Hippocrate_ancien.html, consulté le 5 mars 2012.
33. Isaac Asimov (1990), « La preuve », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 413.
34. Isaac Asimov (1990), « Cercle vicieux », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 208.
35. *Ibid*., p. 207.
36. *Ibid*., p. 208.
37. *Ibid*., p. 210.
38. *Ibid*., p. 200.
39. http://www.bible-ouverte.ch/questions-reponses/qr-doctrine-chretienne/2091-reponse-27.html, consulté le 25 février 2012.
40. http://fr.wikipedia.org/wiki/John_Langshaw_Austin, consulté le 2 avril 2012.
41. http://fr.wikipedia.org/wiki/Antigone_(Sophocle), consulté le 8 mars 2012.

42. http://fr.wikipedia.org/wiki/Procès_de_Nuremberg#Principes_.C3.A9thiques_et_politiques_d.C3.A9velopp.C3.A9s, consulté le 10 mars 2012.
43. http://www.un.org/fr/documents/udhr/index2.shtml, consulté le 9 mars 2012.
44. http://www.protestant.ch/applic/ge/autorites.nsf/85255db800470aa485255d8b004e349a/d08bc1f44c2a6f46c125665d0055f155?OpenDocument, consulté le 5 mars 2012.
45. Isaac Asimov (1990), « Attrapez-moi ce lapin », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 237.
46. Isaac Asimov (1990), « Lenny », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 288.
47. Isaac Asimov (1990), « La preuve », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 423.
48. http://www.micheline.ca/doc--1730-hammourabi.htm.
49. Isaac Asimov (1990), « Lenny », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 288.
50. *Ibid.*
51. *Ibid.*, p. 295.
52. Isaac Asimov (1990), « Première Loi », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 195.
53. *Ibid.*, p. 198.
54. http://philo.pourtous.free.fr/Articles/Eric/mensongeethique.htm.
55. Isaac Asimov (1990), « Menteur ! », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 268.
56. *Ibid.*, p. 270.
57. Isaac Asimov (1990), « Le correcteur », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 303.
58. *Ibid.*
59. *Ibid.*, p. 319.
60. *Ibid.*, p. 320.
61. *Ibid.*, p. 312.
62. *Ibid.*, p. 327.
63. *Ibid.*, p. 329-330.
64. Isaac Asimov (1990), « Effet miroir », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 169.
65. *Ibid.*, p. 174.
66. *Ibid.*, p. 174-175.
67. *Ibid.*, p. 176.
68. *Ibid.*
69. *Ibid.*
70. *Ibid.*, p. 177.
71. *Ibid.*, p. 178.
72. Isaac Asimov (1990), « Le petit robot perdu », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 341.
73. *Ibid.*, p. 339.
74. *Ibid.*, p. 342.

75. *Ibid*.
76. *Ibid*., p. 346.
77. Isaac Asimov (1990), « L'homme bicentenaire », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 508.
78. Isaac Asimov (1990), « Risque », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 381.
79. *Ibid*., p. 378.
80. *Ibid*., p. 381.
81. Isaac Asimov (1990), « Satisfaction garantie », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 281.
82. *Ibid*., p. 285.
83. Isaac Asimov (1990), « Noël sans Rodney », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 147.
84. *Ibid*., p. 148.
85. *Ibid*.
86. Isaac Asimov (1990), « Raison », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 217.
87. *Ibid*., p. 216.
88. *Ibid*., p. 217.
89. *Ibid*., p. 218.
90. *Ibid*., p. 228.
91. *Ibid*., p. 219.
92. *Ibid*., p. 216.
93. *Ibid*., p. 220.
94. *Ibid*.
95. *Ibid*., p. 223.
96. *Ibid*., p. 222.
97. Isaac Asimov (1990), « Pour que tu t'y intéresses », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 475.
98. *Ibid*., p. 476.
99. *Ibid*.
100. *Ibid*., p. 476-477.
101. *Ibid*., p. 477.
102. *Ibid*.
103. *Ibid*., p. 494.
104. *Ibid*.
105. *Ibid*., p. 494-495.
106. Isaac Asimov (1990), « Noël sans Rodney », Nous les robots. *Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 150.
107. Isaac Asimov (1990), « *Le robot qui rêvait* », Nous les robots. *Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 451.
108. *Ibid*., p. 452.
109. *Ibid*., p. 450.

110. Isaac Asimov (1990), « Conflit évitable », Nous les robots. *Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 445.
111. *Ibid.*, p. 446.
112. Isaac Asimov (1990), « L'incident du tricentenaire », *Nous les robots. Le grand livre des robots 1. Prélude à Trantor*, *op. cit.*, p. 191.
113. Isaac Asimov (1991), « Les robots de l'aube », *Le grand livre des robots 2. La gloire de Trantor*, *op. cit.*, p. 342.
114. *Ibid.*
115. Isaac Asimov (1991), « Les robots et l'Empire », *Le grand livre des robots 2. La gloire de Trantor*, *op. cit.*, p. 595.
116. *Ibid.*
117. *Ibid.*, p. 596.
118. *Ibid.*, p. 597.
119. *Ibid.*
120. *Ibid.*
121. *Ibid.*
122. *Ibid.*, p. 674.
123. *Ibid.*, p. 676.
124. *Ibid.*, p. 677.
125. Isaac Asimov (1978), « La vie et les œuvres de Multivac », *L'homme bicentenaire*, Paris, Denoël, p. 169.
126. *Ibid.*, p. 172.

CHAPITRE 3

1. P.W. Singer (2010), « War of the Machines », paru dans la revue *Scientific American*, juillet.
2. Voir par exemple M. Anderson et S.L. Anderson (2010), « Robot Be Good », *Scientific American*, octobre, p. 72-77, et M. Anderson et S.L. Anderson (2007), « Machine Ethics : Creating an Ethical Intelligent Agent », *AI Magazine*, vol. 28, p. 15-26.
3. J.H. Moor (2006), « The nature, importance and difficulty of machine ethics », *IEEE Intelligent Systems*, vol. 21, n° 4, p. 18-21.
4. D. Suvin (1972), « On the Poetics of the Science Fiction Genre », *College English*, vol. 34, n° 3, p. 372-382.
5. M. Crichton, paru originalement sous le titre *Jurassic Park*, publié par Ballantine Books (1991).
6. *Star Trek*, série télévisée créée par Gene Roddenberry, présentée originalement sur le réseau NBC entre le 8 septembre 1966 et le 3 juin 1969.
7. *Les cavernes d'acier*, originalement paru en série sous le titre « The Caves of Steel », dans le magazine *Galaxy*, octobre à décembre, 1953.
8. « Le petit robot perdu », paru originalement sous le titre « Little Lost Robot », dans le magazine *Astounding Science Fiction*, mars 1947.
9. Le premier concept d'horloge atomique fut proposé par Isidor Rabi en 1945 et la première horloge atomique fut construite en 1949 par le National Bureau of Standards aux États-Unis.

10. « Raison », originalement paru sous le titre « Reason » dans le magazine *Astounding Science Fiction*, avril 1941.
11. *Les cavernes d'acier*, originalement paru en série sous le titre « The Caves of Steel », dans le magazine *Galaxy*, octobre à décembre, 1953.
12. *Face aux feux du soleil*, originalement paru en série sous le titre « The Naked Sun », dans le magazine *Astounding Science Fiction* d'octobre à décembre 1956.
13. « Le conflit évitable », originalement paru sous le titre « The Evitable Conflict », dans le magazine *Astounding Science Fiction*, juin 1950.
14. *Les robots et l'empire*, originalement paru sous le titre *Robots and Empire*, publié par Doubleday Books en 1985.
15. Concept originalement proposé par Paul Henri Thiry, baron d'Holback, *Système de la nature*, 1770.
16. *Face aux feux du soleil*, originalement paru en série sous le titre « The Naked Sun » dans le magazine *Astounding Science Fiction*, d'octobre à décembre 1956.
17. *Les robots de l'aube*, paru originalement sous le titre *The Robots of Dawn*, publié par Doubleday Books en 1983.
18. « Raison », originalement paru sous le titre « Reason », dans le magazine *Astounding Science Fiction*, avril 1941.
19. A. Einstein, B. Podolsky et N. Rosen (1935), « Can quantum-mechanical description of physical reality be considered complete ? », *Physical Review*, vol. 47, p. 777.
20. J.S. Bell (1964), « On the Einstein-Podolsky-Rosen paradox », *Physics*, vol. 1, p. 195.
21. A. Aspect, J. Dalibard et G. Roger (1982), « Experimental test of Bell's inequalities using time-varying analyzers », *Physical Review Letters*, vol. 49, p. 1804.
22. E.M. Gauger et autres (2011), « Sustained Quantum Coherence and Entanglement in the Avian Compass », *Physical Review Letters*, vol. 106, p. 040503.
23. « Attrapez-moi ce lapin », originalement paru sous le titre « Catch That Rabbit », dans le magazine *Astounding Science Fiction*, février 1944.
24. « Menteur ! », originalement paru sous le titre « Liar ! », dans le magazine *Astounding Science Fiction*, mai 1941.
25. « Robot », Wikipedia, http://fr.wikipedia.org/wiki/Robot, consulté le 21 avril 2012.
26. « Cercle vicieux », originalement paru sous le titre « Runaround », dans le magazine *Astounding Science Fiction*, mars 1942.
27. « Menteur ! », originalement paru sous le titre « Liar ! », dans le magazine *Astounding Science Fiction*, mai 1941.
28. *Les robots de l'aube*, paru originalement sous le titre *The Robots of Dawn*, publié par Doubleday Books en 1983.
29. Le titre fut traduit *Rossum's Universal Robots* lors de l'adaptation française.
30. S. Russell et P. Norvig (2003), *Artificial Intelligence. A Modern Approach*, Upper Saddle River, NJ, Prentice-Hall, p. 6.
31. « History of artificial intelligence », Wikipedia, http://en.wikipedia.org/wiki/History_of_artificial_intelligence, consulté le 11 mai 2012.
32. Cité dans *Ibid*.
33. Cité dans *Ibid*.

34. S. Russell et P. Norvig (2003), *op. cit.*, p. 14-15.
35. R. Herken, ed. (1995), *The Universal Turing Machine : A Half-Century Survey*, New York, Springer, 611 p.
36. *Computation en anglais*.
37. « Von Neumann architecture », Wikipedia, http://en.wikipedia.org/wiki/Von_Neumann_architecture, consulté le 11 mai 2012.
38. S. Russell et P. Norvig (2003), *op. cit.*, p. 15, 940 ; Hans Moravec (1988), *Mind Children : The Future of Robot and Human Intelligence*, Harvard University Press, p. 3.
39. D. Crevier (1993), *AI : The Tumultuous Search for Artificial Intelligence*, New York, Basic Books, p. 44-47.
40. W. Ertel (2011), *Introduction to Artificial Intelligence*, London, Springer, p. 6.
41. A.M. Turing (1950), « Computing machinery and intelligence », *Mind*, vol. 59, p. 433-460.
42. Cité dans W. Ertel (2011), *op. cit.*, p. 1.
43. V. Braitenberg (1984), *Vehicles : Experiments in synthetic psychology*, Cambridge, MA, MIT Press, 18 p. ; W. Ertel (2011), *op. cit.*, p. 8.
44. Cité dans W. Ertel (2011), *op. cit.*, p. 2.
45. En anglais : « Artificial Intelligence (A.I.) is the study of how to make computers do things at which, at the moment, people are better. » E. Rich (1983), *Artificial Intelligence*, New York, McGraw-Hill.
46. W. Ertel (2011), *op. cit.*, p. 5-6.
47. *Ibid*., p. 28.
48. *Ibid*., p. 5, 43.
49. « Calcul des prédicats », Wikipedia, http://fr.wikipedia.org/wiki/Calcul_des_prédicats, consulté le 30 avril 2012.
50. W. Ertel (2011), *op. cit.*, p. 42-43.
51. Jack Copeland (2000), « A Brief History of Computing », AlanTuring.net, http://www.alanturing.net/turing_archive/pages/Reference Articles/BriefHistofComp.html, consulté le 30 avril 2012.
52. W. Ertel (2011), *op. cit.*, p. 108-109.
53. Cité dans D. Crevier (1993), *op. cit.*, p. 108 ; S. Russell et P. Norvig (2003), *op. cit.*, p. 21.
54. H.A. Simon (1965), *The Shape of Automation for Men and Management*, New York, Harper & Row.
55. Marvin Minsky (1967), *Computation : Finite and Infinite Machines,* Englewood Cliffs, N.J., Prentice-Hall.
56. Cité dans D. Crevier (1993), *op. cit.*
57. W. Ertel (2011), *op. cit.*, p. 5.
58. S. Russell et P. Norvig (2003), *op. cit.*, p. 21 ; W. Ertel (2011), *op. cit.*, p. 7, 57.
59. S. Russell et P. Norvig (2003), *op. cit.*, p. 22.
60. H. Moravec (1976), « The Role of Raw Power in Intelligence », http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html, consulté le 1er mai 2012.
61. S. Russell et P. Norvig (2003), *op. cit.*, p. 21.

62. « History of artificial intelligence », Wikipedia, http://en.wikipedia.org/wiki/History_of_artificial_intelligence, consulté le 1er mai 2012.
63. S. Russell et P. Norvig (2003), *op. cit.*, p. 21.
64. Hans Moravec (1988), *Mind Children : The Future of Robot and Human Intelligence*, Harvard University Press, 176 p.
65. M. Negnevitsky (2005), *Artificial Intelligence : A Guide to Intelligent Systems*, Harlow, R.-U., Addison-Wesley, p. 4-5.
66. « Fuzzy logic », Wikipedia, http://en.wikipedia.org/wiki/Fuzzy_logic, consulté le 6 mai 2012.
67. W. Ertel (2011), *op. cit.*, p. 8.
68. « Logique floue », Wikipedia, http://fr.wikipedia.org/wiki/Logique_floue, consulté le 1er mai 2012.
69. Peter Jackson (1998), *Introduction to Expert Systems*, Addison-Wesley, p. 2.
70. S. Russell et P. Norvig (2003), *op. cit.*, p. 21-22.
71. M. Negnevitsky (2005), *op. cit.*, p. 30-33.
72. W. Ertel (2011), *op. cit.*, p. 131-144.
73. « Système expert », Wikipedia, http://fr.wikipedia.org/wiki/Système_expert, consulté le 7 mai 2012.
74. W. Ertel (2011), *op. cit.*, p. 13.
75. *Ibid*.
76. *Ibid*., p. 162-164.
77. *Ibid*., p. 221-256.
78. Le *Grand Dictionnaire terminologique* de l'Office québécois de la langue française propose l'utilisation du terme français *exploration de données* en remplacement du concept anglais *data mining*. Cependant, le terme *forage de données* est aussi souvent utilisé.
79. M. Negnevitsky (2005), *op. cit.*, p. 349-360.
80. « The DeepQA Project » (2011), *IBM*, http://www.research.ibm.com/deepqa/deepqa.shtml, consulté le 7 mai 2012.
81. « IBM – Watson » (2011), *IBM*, http://www-03.ibm.com/innovation/us/watson/, consulté le 7 mai 2012.
82. P. McCorduck (2004), *Machines Who Think, Natick*, MA, CRC Press, p. 454-462.
83. H. Moravec (1988), *op. cit.*, p. 20.
84. Traduction libre : « I am confident that this bottom-up route to artificial intelligence will one date meet the traditional top-down route more than half way, ready to provide the real world competence and the commonsense knowledge that has been so frustratingly elusive in reasoning programs. »
85. S. Russell et P. Norvig (2003), *op. cit.*, p. 27-58.
86. W. Ertel (2011), *op. cit.*, p. 9-12.
87. S. Russell et P. Norvig (2003), *op. cit.*, p. 32.
88. Il s'agit de traductions libres des catégories proposées par Russell et Norvig.
89. W. Ertel (2011), *op. cit.*, p. 12, 139-141.
90. « Intelligent Agent », Wikipedia, http://en.wikipedia.org/wiki/intelligent_agent, consulté le 8 mai 2012.

91. G.E. Moore (1965), « Cramming more components onto integrated circuits », *Electronics*, vol. 38(8).
92. R. Kurzweil (2005), *The Singularity is Near, Viking*, É.-U., 672 p.
93. M. Anderson et S.L. Anderson (2007), *op. cit.*, p. 15.
94. J.H. Moor (2006), *op. cit.*, p. 19-21.
95. M. Anderson et S.L. Anderson (2007), *op. cit.*
96. *Ibid.* ; M. Guarini (2006), « Particularism and the Classification and Reclassification of Moral Cases », *IEEE Intelligent Systems*, vol. 21, n° 4, p. 22-28.
97. M. Anderson et S.L. Anderson (2010), *op. cit.*, p. 72-77.
98. P.W. Singer (2010), *op. cit.*
99. J.H. Moor (2006), *op. cit.*, p. 20.
100. Traduction libre : « For example, you might program a machine with the classical imperative of physicians and Asimovian robots : First, do no harm. Bur this wouldn't be helpful unless the machine could understand what constitutes harm in the real world. » *Ibid.*, p. 20.

Asimov

et l'acceptabilité des robots

Même si les robots humanoïdes actuels ne sont pas encore vraiment au point (intelligence artificielle, motricité) et sont très loin des robots moraux au cerveau positronique imaginés par Asimov, il n'en reste pas moins que ses ouvrages sont une démonstration de la complexité des problèmes moraux qui surviennent dans le processus d'analyse d'impact et d'acceptabilité des robots. C'est en situant les lois morales de la robotique en regard des valeurs que nous voulons atteindre dans nos vies individuelles et collectives que la pensée d'Asimov rejoint tout le courant de l'éthique appliquée qui s'est développé depuis plus de cinquante ans.

JEAN-PIERRE BÉLAND est philosophe et professeur au Département des sciences humaines de l'Université du Québec à Chicoutimi.

GEORGES A. LEGAULT est docteur en philosophie et licencié en droit. Après son doctorat en philosophie du droit, il s'est consacré à des recherches sur la formation morale et ensuite sur l'éthique professionnelle.

En collaboration avec **JACQUES BEAUVAIS**, professeur en génie électrique à l'Université de Sherbrooke et **JONATHAN GENEST**, professionnel de recherche à l'Université de Sherbrooke.

www.pulaval.com

ISBN 978-2-7637-4668-5

9 782763 746685

Collection À propos

Éthique