Isaac Asimov : les (quatre) “ lois de la robotique ” et l’échange de paroles

Patrick Goutefangea

▶ **To cite this version:**

Patrick Goutefangea. Isaac Asimov : les (quatre) “ lois de la robotique ” et l’échange de paroles. 2016. hal-01397948

HAL Id: hal-01397948
https://hal.science/hal-01397948

Submitted on 9 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Isaac Asimov : les (quatre) « lois de la robotique » et l'échange de paroles

Patrick Goutefangea

Isaac Asimov s'est rendu célèbre en énonçant, pour les besoins de la Science Fiction, les « trois lois de la robotique » :

- Un robot ne peut porter préjudice à un être humain, ou, par inaction, laisser un être humain subir un préjudice.
- Un robot doit obéir aux ordres qui lui sont donnés par des êtres humains, sauf dans le cas où de tels ordres entrent en conflit avec la Première Loi.
- *Un robot doit protéger sa propre existence aussi longtemps qu'une telle protection n'entre pas en conflit avec la Première ou la Seconde Loi.*[1]

Ces lois sont au principe des quelques dizaines de nouvelles consacrées par Asimov aux robots, et dont la figure centrale est la « psychoroboticienne » Susan Calvin ; elles sont également au centre des six romans contant l'épopée de Daneel Oliwav[2], ce robot modeste, timide et bienveillant, voué à devenir le sauveur de l'humanité. C'est sur elles, enfin, que s'appuie Asimov pour établir *a posteriori* un lien entre le cycle de Daneel Oliwav et celui de *Foundation*, sa première grande contribution aux classiques de la SF[3]. Dans *Robots and Empire*, Daneel Oliwav, inspiré par l'exemple de son ami humain disparu, Elijah Bailey, et aidé du robot télépathe Giscard, prend en charge le destin des humains après avoir formulé une « loi zéro » de la robotique, selon laquelle l'intérêt de l'humanité prévaut pour le robot sur celui des individus humains, ce qui le conduira, une vingtaine de millénaires plus tard, dans *Prelude to foundation* (1988), à favoriser, toujours pour le salut de l'humanité, l'invention, par un humain, Hari Seldon, de la science positive - et positiviste - de l'histoire, la « psychohistoire », sur laquelle le cycle suivant reposera.

Asimov met en scène des machines avec lesquelles les hommes ont un *échange de paroles*. Les lois de la robotique ont pour charge de réguler les implications d'un tel échange, fournissant par là-même une solution au paradoxe que présente le robot classique de la Science-fiction, entité parfaitement déterminée et cependant autonome.

Nous voudrions montrer que c'est la dynamique spécifique de la situation d'échange de paroles elle-même qui décide du destin de sauveur de l'humanité assigné au robot Daneel Oliwav : c'est elle qui fonde, en dernier ressort, la formulation de la « loi zéro » selon laquelle l'intérêt de l'humanité prévaut sur celui de l'individu humain, débordant ainsi la fonction première des lois. Sous les traits du robot d'Asimov, la machine parlante soumise

---

1 Les « trois lois de la robotique » apparaissent pour la première fois sous la plume d'Asimov dans la nouvelle intitulée « Runaround » (*Astounding Science Fiction*, mars 1942), incluse dans *The complete Robot*, Nightfall inc., 1982).

2 *The Caves of Steel* (1954), *The Naked Sun* (1957), *The Robots of Dawn* (1983), *Robots and Empire* (1985), auxquels doit être ajouté *Prelude to Foundation* (1988).

3 *Foundation* (1951, mais dont la publication eut d'abord lieu dans *Astounding Science Fiction* à partir de 1942), *Foundation and Empire* (1952), *Second Foundation* (1953), *Foundation's Edge* (1982) et *Foundation and Earth* (1983).

aux lois de la robotique est élevée par celles-ci à la dignité même de l'humain qu'elles invoquent et que la tradition philosophique décrit au travers de notions telles que celles de « sujet », « d'autrui », de « semblable » ou de « personne ».

Nous verrons comment les lois de la robotique résolvent en apparence le paradoxe inhérent à la notion de robot, rompant avec la conception classique des rapports de l'homme et de la machine, pour faire du robot mieux qu'une machine : un esclave idéal ; nous verrons ensuite de quelle ambiguïté les lois sont porteuses et comment cette ambiguïté même devient le moteur du récit d'Asimov, conduisant jusqu'à l'idée d'une humanité du robot.

Les lois de la robotique comme solution au paradoxe de l'idée de robot

Un robot est d'abord une machine, c'est-à-dire un outil, lequel est, pour l'homme, dont il rend le geste plus puissant, plus rapide, plus précis, plus constant, une extension de son propre corps. L'outil prolonge le geste humain et suppose à ce titre un agent, qui l'invente, le construit et le manipule.

La machine est un outil complexe, qui transforme l'énergie et la contrôle, disposant, par là, d'une certaine autonomie. Elle n'en est pas moins, comme l'outil le plus simple, une extension du corps humain, et son autonomie renvoie simplement à 'un stade avancé de l'outil, celui où le geste est prolongé jusqu'à être séparé de son support d'origine.

Le robot est une machine complexe, dont le degré d'autonomie a atteint un stade particulier : il prolonge, jusqu'à le détacher de son support d'origine, non pas tel geste humain, mais une séquence organisée de gestes constituant déjà une chaîne opératoire.

Les robots d'Asimov, auxquels s'appliquent les « trois lois de la robotique », sont, quant à eux, des robots de science-fiction et la principale convention du genre consiste à faire abstraction des actuelles contraintes techniques, de sorte que leur vraisemblance ne soit plus celle d'une hypothèse d'ingénieur, mais celle du rêve d'Asimov et de ses lecteurs. Leur corps artificiel, qui surpasse en tout point le corps de chair des individus humains, est contrôlé par un « cerveau positronique »[4], grâce auquel ce n'est plus telle séquence de gestes humains, mais, dans sa globalité, le *système* de l'action humaine individuelle qu'ils prolongent jusqu'à le séparer de son support initial.

A ce titre, ces robots *parlent*, comme si c'était la parole, ce geste humain singulier, qui était, en eux, séparée de son support originel. Les plus perfectionnés des robots mis en scène par Asimov disposent de la même *faculté* de langage que les hommes : ils ne sont pas simplement capables, en tant que machines, d'émettre des sons articulés, ni même, en tant que machines logiques, de faire des calculs et de construire des propositions, mais ils peuvent avoir, avec les hommes, un *échange de paroles*. Leur valeur en tant qu'outil est largement fonction de ce trait étonnant qui leur confère un caractère radicalement non-passif : les robots d'Asimov, non seulement doivent comprendre ce que disent les humains qui les utilisent, mais ils doivent se faire comprendre d'eux, ce qui suppose la capacité à décrire leur propre action.

Par là, le robot de science-fiction est une entité paradoxale. Sa valeur en tant qu'outil repose, bien davantage que sur sa puissance, son adresse ou sa robustesse, sur le fait

4 Autrement dit électronique : le « positron » est l'antiparticule de l'électron.

même que tout le système de l'action humaine individuelle soit déporté en lui, c'est-à-dire sur son autonomie quasi humaine. Un outil, cependant, doit être, en théorie, utilisable par n'importe quel homme ayant reçu l'apprentissage approprié ; il doit pouvoir être contrôlé par le plus grand nombre et la plus grande variété d'individus humains. Or, si le robot d'Asimov disposait d'une autonomie strictement équivalente à celle des hommes, comment satisferait-il à cette condition, lui qui, sur le plan de la puissance mécanique (laquelle inclut la puissance de calcul), est supérieur à n'importe quel individu humain ? Le robot d'Asimov doit tout à la fois disposer de l'autonomie quasi humaine qui fait sa valeur en tant qu'outil et rester dans les limites du contrôle humain hors duquel il cesserait d'être un outil. Les trois « lois de la robotique » ont pour vocation de fournir une solution à ce problème.

Elles présentent deux traits fondamentaux : leur cohérence et leur « objectivité ». La loi n° 1, selon laquelle le robot ne peut nuire à un être humain, est première aussi bien dans l'ordre de la fonction que dans celui de l'énoncé : le recours au robot, c'est-à-dire son usage en tant qu'outil, implique que la sécurité de l'agent qui l'utilise soit garantie - du reste, certains récits d'Asimov ne font appel qu'à elle[5]. Cette première loi trouve sa portée véritable par son second volet, grâce auquel le robot-outil, non seulement ne menace pas l'agent qui l'utilise, mais contribue à la protection générale de celui-ci ; la proposition : « [Un robot ne peut] par inaction laisser un être humain subir un préjudice », qui lui commande de porter secours a un homme en danger, inscrit dans la structure même du robot le fait que la survie des individus humains et le bien-être de ceux-ci sont ce à quoi il *sert*. Non seulement le robot n'existe que par l'homme, qui le fabrique, mais, bien plus, il n'existe que *pour* l'homme. Comme toute autre machine, c'est en l'homme, et non en lui-même, qu'il trouve son principe et sa fin.

La seconde loi, qui détermine le robot à l'obéissance, c'est-à-dire à recevoir et exécuter des ordres formulés par les hommes, est unie logiquement à la première par la proposition selon laquelle « [un robot doit obéir] sauf lorsque [les] ordres [qu'il reçoit] entrent en conflit avec la Première Loi ». On comprend bien, en effet, que si cette loi ne mentionnait rien de plus que l'obéissance due à l'homme par le robot, sa dynamique pourrait, dans de nombreux cas, s'opposer à celle de la première : il est aisé d'imaginer des situations dans lesquelles un homme ordonnera à un robot de porter préjudice à un autre homme. Ce second volet de la « Deuxième loi » complète le dispositif de sécurité constitué par la première.

La « Troisième loi » vise, quant à elle, l'autonomie particulière dont dispose le robot. Puisque c'est tout le système de l'action humaine qui est déporté dans la machine, celle-ci, lors de l'exécution des ordres qui lui sont donnés par les hommes, est livrée à elle-même et doit pouvoir protéger sa propre existence. La « Troisième loi » lui procure l'équivalent d'un instinct de survie. Il convient simplement que cet instinct soit moins puissant que les deux premières lois ; ce n'est que dans le cas où le robot, toujours soumis aux deux premières lois, est privé d'instructions directes des hommes, que le souci de sa propre préservation peut diriger son action. Du reste, il ne s'agit pas tant, alors, pour lui, d'agir en fonction de son propre « intérêt », que de garantir l'usage de l'outil qu'il ne cesse jamais d'être.

Par ailleurs, les « lois de la robotique » sont, du point de vue du comportement du robot,

5 Voir, par exemple, *Satisfaction Guaranteed (Super Science Stories*, 1951), *First Law (Fantastic Universe*, 1956).

des lois *objectives*, tout comme les lois de l'électronique qui président à l'organisation de son « cerveau », ou celles de la mécanique qui régissent son « corps » : elles sont « implémentées » directement dans son « cerveau positronique » ; elles font partie de sa structure et déterminent les propriétés de celle-ci. Les lois sont des conditions du fonctionnement du robot, lequel est censé exécuter les instructions qu'il reçoit sous la forme d'ordres prononcés par ses maîtres humains exactement comme n'importe quelle machine « démarre » lorsque son opérateur appuie sur le bon bouton. Il « obéira » aussi longtemps que l'ordre n'aura pas été totalement exécuté et que la séquence d'actions exigée par l'instruction n'entrera pas en conflit avec l'une des lois. Dans ce dernier cas, le robot modifiera son action, ou s'arrêtera - il ne s'agira pas de sa part d'un « refus » de continuer, mais simplement d'une impossibilité fonctionnelle de le faire. S'il se trouve placé, enfin, dans une situation où les dynamiques respectives des lois s'opposent et où les forces qui l'animent se neutralisent, il risquera la panne. Tel est, du moins, le sens que les lois de la robotique doivent avoir pour les hommes qui les implémentent dans le cerveau positronique des robots. Ainsi entendu, le robot constitue une sorte de machine idéale, prolongement achevé de l'individu humain, prolongement de l'homme technicien tout entier plutôt que de la seule main ou du seul bras de celui-ci, bref, prolongement de l'individu humain comme agent d'un système d'action organisé.

Cependant, par cela même, la machine d'Asimov rompt avec l'approche traditionnelle de la notion d'automate : elle bouleverse la conception classique des rapports de l'homme et de la machine qui fondait celle-ci.

Mieux qu'une machine : un esclave idéal

La possibilité d'une distinction de principe entre l'homme et la machine constitue un élément essentiel de la définition classique de celle-ci. Un texte célèbre du *Discours de la méthode* en atteste. Imaginons, propose Descartes, des machines conçues pour imiter aussi parfaitement que possible, dans son corps et son comportement extérieur, un individu humain. « Nous aurions toujours, déclare-t-il, deux moyens très certains pour reconnaître qu'elles ne seraient point pour cela de vrais hommes »[6]. Le premier de ces moyens est la parole : quand bien même nous parviendrions à construire une machine capable de « proférer » des sons articulés ayant un sens pour nous, il est exclu, aux yeux de Descartes, qu'une telle machine « arrange diversement » les signes que nous lui ferions utiliser « pour répondre au sens de tout ce qui se dit en sa présence, ainsi que les hommes les plus hébétés peuvent faire »[7].

Le second moyen indiqué par Descartes renvoie à l'action réfléchie : si les bonnes machines sont généralement en mesure d'effectuer certaines actions mieux que n'importe quel homme, elles ne « savent », toutefois, faire autre chose que l'action précise en vue de laquelle elles ont été construites. L'automate, parce qu'il est une machine, relève de la seule « substance étendue » ; son mouvement est entièrement explicable par les lois de la mécanique, et c'est précisément pour cette raison qu'il peut être construit par l'homme qui, lui, participe, en outre, de la « substance pensante ».

Il arrive, pourtant, comme le notera Kant un peu plus tard, que l'idée de « liberté » soit

6 « Discours de la méthode », *Œuvres philosophiques*, I, Paris, Classiques Garnier, 1988, p. 629.
7 *Ibid.*

évoquée à propos des automates : leur mouvement est dit « libre », bien qu'il soit soumis à la nécessité de la causalité naturelle, car, une fois mis en route, l'automate n'a plus besoin, pour continuer à se mouvoir, d'un principe externe. Les moments successifs de son mouvement, hormis le premier, dû à une impulsion extérieure, sont causes suffisantes les uns des autres. Nous usons alors, dit Kant, d'un « concept comparatif de liberté »[8] : le mouvement de l'automate est comparé à celui de l'homme, lequel est déterminé, non seulement par les forces externes de la nature - il est, lui aussi, soumis à la nécessité naturelle - mais également par des « représentations intérieures produites par ses propres forces ». L'être humain présente, en effet, cette particularité de pouvoir, à travers sa faculté de représentation, peser sur la nécessité naturelle à laquelle il est soumis[9]. Former, à propos de l'automate, un « concept comparatif de liberté », c'est assimiler à cette dimension de l'action humaine le fait que l'automate ait en lui-même le « principe naturel de détermination » de son mouvement.

Or, poursuit Kant, l'idée de liberté ne se prête nullement à un tel usage « comparatif ». Le vrai problème qu'elle pose n'est pas celui de l'entretien du mouvement, mais celui du *commencement* de l'action libre : la liberté implique « un pouvoir de commencer tout à fait spontanément une série dans le temps... »[10]. Pour qu'une action, toujours inscrite dans la causalité naturelle, soit libre, il faut que son premier moment ne soit déterminé par aucune autre cause que lui-même. La liberté est *a priori*, et doit être définie comme « transcendantale », c'est-à-dire comme condition de possibilité d'une action d'un tout autre type que celle de l'automate : l'impulsion que ce dernier doit recevoir pour commencer à se mouvoir ne lui appartient pas[11], de sorte que l'idée de « liberté », appliquée à l'automate, n'est qu'un « misérable subterfuge »[12].

Bref, pour la tradition philosophique, quelque chose de l'homme échappe toujours à la machine, laquelle ne peut imiter que cette part de l'humain qui appartient au monde sensible, au monde phénoménal, règne de la causalité physique. Une autre part, qui relève d'un autre ordre - celui de l'intelligible - lui est inaccessible, et cela même la définit comme machine. Cette conception est récusée par Asimov, dès lors qu'il équipe son robot de la parole.

Lorsque Descartes, en effet, affirme que le langage permettrait toujours de distinguer un homme d'un automate, il ne fait pas tant référence à l'être phénoménal du langage, c'est-à-dire aux sons proférés par l'homme, qu'à la *faculté* de parler, à la capacité de signifier à l'aide des sons émis. C'est de cette même faculté qu'est doté le robot d'Asimov : les sons qu'il profère sont, pour lui comme pour les êtres humains avec lesquels il communique, des signes qu'il « arrange diversement pour répondre au sens de tout ce qui se dira en sa présence, ainsi que les hommes les plus hébétés peuvent faire ». En d'autres termes, si le

8 « Critique de la raison pratique », *Œuvres philosophiques*, Gallimard (La Pléiade), II, 1985, p. 725 et sq.

9 « L'action faite par devoir » est l'expression véritable de ce pouvoir.

10 *Ibid.*, p. 1106.

11 « ... tout événement, par conséquent aussi toute action, qui arrive dans un point du temps, est nécessairement sous la condition de ce qui était dans le temps précédent. Or, comme le temps passé n'est plus en mon pouvoir, toute action que j'accomplis d'après des causes déterminantes qui ne sont pas en mon pouvoir, doit être nécessaire, c'est à dire que je ne suis jamais libre dans le point du temps où j'agis. ». *Ibid.*,

12 *Ibid.*, p. 726.

robot n’est pas une machine ordinaire, ce n’est pas seulement parce qu’il reproduit, grâce à une source d’énergie providentielle, l’autonomie biologique dont l’homme dispose au même titre que tous les organismes supérieurs ; ce n’est pas non plus seulement parce que, contrairement à une machine ordinaire qui exécute toujours les mêmes mouvements, il peut s’adapter à des situations que ni lui ni ses concepteurs ne connaissent ; c’est, bien plus, parce qu’il est capable d’effectuer ces actions nombreuses que l'homme exécute alors même qu'il ne sait pas les décrire formellement. L'intérêt fonctionnel du robot repose principalement sur le fait qu'il puisse comprendre un discours humain jusque dans ce que ce discours conserve de vague, d'indécis, bref, *d'intuitif*. La machine d'Asimov doit pouvoir *interpréter* le discours humain qui lui transmet les ordres qu'elle exécute. Elle doit être en mesure de communiquer avec un homme sans passer par la traduction du langage humain ordinaire dans un langage strictement formel qui lui serait propre ; bref, elle *échange* la parole avec ses maîtres humains. Ce sont, ici, les limites de l'autonomie purement technique ou strictement biologique qui sont franchies.

Dans les termes de la philosophie classique, la parole, en tant que faculté, renvoie, en effet, au « jugement ». Chez Descartes, la parole qui permet de distinguer l'homme de l'automate implique, non seulement l'énonciation d'une idée, éventuellement « claire et distincte », mais l'affirmation ou la négation de ce qui est conçu, une « prise de position » comme on dirait aujourd'hui. La parole, comme l'action réfléchie, renvoient, au-delà de l'entendement qui « conçoit », à la volonté, principe actif du jugement. Dès lors, et puisqu'il est doué de parole, pourra-t-on refuser au robot d'Asimov une certaine forme de liberté ? Comment parlerait-il avec les hommes sans, comme ceux-ci, poser, de quelque manière, un acte indépendant ? Aussi bien l’autonomie dont il dispose n’est-elle comparable ni à celle de la machine, au sens classique du terme, ni à celle de l’animal domestique ; en vérité, le robot d'Asimov est un esclave idéal.

Qu'est-ce qu'un esclave ? L'équivalent exact d'un homme, excepté en ce qui concerne son rapport à sa propre liberté : il ne se distingue pas de l’homme libre par cela que l'idée de liberté serait pour lui privée de sens, mais par cela que l'exercice plein et entier de sa propre liberté lui est refusé. Au moins en principe, tout homme peut être réduit en esclavage, de même que tout esclave peut recouvrer la liberté, ce qui suppose qu'il n'y ait pas de différence de nature entre l'esclave et l'homme libre. Du reste, l'esclave parle, et, rien ne peut faire que sa propre parole n'ait pas le même statut - celui d'une faculté - que celle de son maître, dès lors que ce dernier l'échange avec la sienne[13]. Or, cela même constitue, pour le maître, la limite de l'esclave humain, en tant qu'outil : seul le rapport de force contingent établi entre eux garantit que l'outil reste sous le contrôle du maître. Du point de vue du maître, l'esclave mécanique imaginé par Asimov apparaît infiniment supérieur dans la mesure où l'exercice plein et entier de la liberté ne lui est pas refusé par la force, mais par programme : le robot n'est pas esclave parce qu'il a été décidé, dans certaines circonstances contingentes, qu'il serait tel, mais parce qu'il est ainsi fait. A cause des lois de la robotique, implémentées directement dans son « cerveau positronique », il ne peut se comporter autrement qu'en esclave ; il est, pourrait-on dire, structurellement soumis. Plus puissant, plus solide, plus économique que l'esclave humain, il est au moins

13 Souvenons-nous du *Ménon*, où Platon montre comment Socrate fait découvrir à un jeune esclave, par le dialogue, la solution au problème du doublement du carré.

aussi intelligent que lui, tout en demeurant une machine. Bref, il semble qu'Asimov soit parvenu à résoudre la quadrature du cercle : doter son robot d'une autonomie quasi humaine, et maintenir pourtant entre l'homme et lui la séparation affirmée par la conception classique des rapports de l'homme et de la machine. Ce qui distingue le monde d'Asimov de cette conception est pourtant capital : la différence entre l'homme et la machine n'est plus *a priori* ; elle est obtenue *a posteriori* par l'implémentation des lois de la robotique dans le cerveau « positronique » du robot.

Or, par là même, cette différence s'obscurcit ; le statut des lois de la robotique est ambiguë : s'agit-il de lois « objectives » ou de règles ? Les lois renvoient-elles à une nécessité objective ou à une nécessité de type conventionnel, juridique ou moral ?

Mieux qu'une machine : un semblable

Comment faut-il interpréter, dans l'énoncé de la « Première loi », l'usage de l'expression « ne pas pouvoir » - « Un robot ne peut porter préjudice à un être humain... » - ? Faut-il comprendre les lois de la robotique sur le modèle des « lois de la nature » ? La première loi de la robotique est-elle analogue à une loi scientifique qui pourrait être énoncée sous la forme d'une formule mathématique utilisée pour la construction du cerveau positronique du robot ? On remarque qu'Asimov emploie *« may »* plutôt que *« can »*, comme si l'on avait affaire à une interdiction plutôt qu'à une impossibilité. Les deuxième et troisième lois accentuent cette ambiguïté : elles sont énoncées davantage sous forme de règles que sous celle de lois au sens scientifique, comme le montre l'usage, dans leur cas, non plus du verbe « pouvoir », mais du verbe « devoir » (« *must* ») : « Un robot doit obéir aux ordres qui lui sont donnés par des être humains... ». Cet usage est conforme au rôle joué par le langage dans le fonctionnement du robot : celui-ci doit obéir à des instructions qui lui sont communiquées par le langage, tout comme l'être humain placé dans une situation de subordination, et c'est pourquoi Asimov, dans l'énoncé de la seconde loi, présente ces instructions comme des « ordres » (« *orders* »). Or, si, en principe, le problème de l'application d'une loi scientifique ne se pose pas, il n'en va pas de même dans le cas de la règle, laquelle peut ne pas être appliquée, ou bien être appliquée de diverses manières. Elle est prescrite, et, on le sait depuis Wittgenstein, elle n'a, au fond, de nécessité qu'autant qu'elle est effectivement suivie. C'est précisément sur cette propriété de la règle que repose la cohérence spécifique des lois de la robotique : dans certains cas, le robot doit *ne pas* obéir, contrairement à ce qu'énonce la deuxième loi ; il faut donc qu'il *puisse* ne pas obéir, qu'il puisse arrêter son action si celle-ci met un être humain en péril. On peut douter qu'il y ait à ce problème une solution purement mécanique, fondée, par exemple, sur une pondération des forces respectives des lois, dans la mesure où les critères de cette pondération ne sont pas eux-mêmes directement mécanisables. Aussi bien Asimov n'envisage-t-il pas une telle solution, et propose-t-il une réponse plus originale, reposant sur le langage.

Les « ordres » que le robot doit exécuter lui sont communiqués verbalement ; ils se présentent comme une série de propositions. On pourrait imaginer que la mise en mouvement du robot soit déclenchée directement par la forme grammaticale propre au type de propositions que sont ces « ordres », à savoir le mode impératif. S'il en était ainsi, ce serait à partir d'une analyse morpho-syntaxique, comme savent en mener nos

ordinateurs actuels, que le robot déclencherait son action : la reconnaissance d'une certaine forme grammaticale suffirait à le faire automatiquement « démarrer ». Comment, cependant, s'il en était ainsi, la cohérence propre aux lois de la robotique serait-elle assurée ? Comment le robot pourrait-il éventuellement *ne pas* obéir, comme l'exige le second volet de la deuxième loi ? Du reste, il arrive fréquemment, dans les récits d'Asimov, que la seule syntaxe ne suffise pas à obtenir l'exécution correcte des ordres : le maître doit insister et, pour obtenir satisfaction, ajouter à la bonne syntaxe l'intonation d'autorité convenable. On trouve même des cas où l'opérateur humain est contraint tout à la fois d'accentuer sa posture d'autorité - contexte de la proposition impérative - et d'expliquer au robot pourquoi il doit obéir. En vérité, l'application de la seconde loi exige du robot un acte d'intelligence : il doit pouvoir ne pas obéir s'il *juge* que l'action qui lui est demandée peut nuire à un être humain. Ce n'est donc pas la forme syntaxique en elle-même qui déclenche l'action du robot, mais la compréhension, par celui-ci, de l'ordre qui lui est donné et l'interprétation de la situation. Bref, c'est l'échange de paroles avec les humains - postulat du récit d'anticipation asimovien - qui est au principe des trois lois, de sorte, enfin, qu'elles ne sauraient être appréhendées comme des dispositifs extérieurs à l'échange de paroles, qui auraient pour fonction de limiter mécaniquement l'action du robot induite par celui-ci : elles reposent elles-mêmes sur l'échange de paroles.

Or, c'est précisément cette indécision quant au statut des lois de la robotique – quelle est la nature et la puissance de l'obligation qu'elles comportent ? - qu'exploite Asimov pour composer le cycle des aventures extraordinaires du robot Daneel Oliwav, au cours duquel il apparaîtra que la logique propre à l'échange de paroles doit l'emporter, dans le comportement du robot, sur le déterminisme strictement mécanique.

Comme tous les robots d'Asimov, Daneel Oliwav est doté d'un cerveau « positronique » dans lequel ont été directement implémentées les trois lois de la robotique. Il se distingue, toutefois, des robots ordinaires en ce qu'il est parfaitement humanoïde dans son apparence extérieure, et « intellectuellement » si proche des hommes que tout individu humain non prévenu le considère spontanément comme l'un d'entre eux. Il est, par ailleurs, chargé de mener des enquêtes policières, aux côtés d'un policier humain, Elijah Bailey, ce qui le place dans la situation de devoir lutter contre certains individus humains au comportement criminel. Comment, dans le cadre des lois, cela lui est-il possible ?

Il s'agit, en effet, d'un cas critique : un robot soumis aux lois de la robotique peut-il choisir entre deux interlocuteurs également humains sans risquer d'être condamné au silence et à l'immobilité ? Tout homme, y compris le plus malhonnête, est un maître pour lui. En vertu des mêmes lois, il ne peut pourtant devenir l'auxiliaire d'un criminel, et n'est pas tenu d'obéir à un ordre qui le conduirait à aider un être humain dans une entreprise menaçant d'autres êtres humains. Or il ne saurait se réfugier dans la neutralité : la première loi ne lui permet pas de rester inerte face aux méfaits commis par tel ou tel homme à l'encontre d'autres hommes. Si la seconde loi l'autorise à ne pas obéir au criminel, la première lui commande de combattre celui-ci. Comment réduire le criminel à l'impuissance sans lui porter préjudice ? La difficulté est si grande que, dans certains récits mettant en jeu des situations plus ou moins proches de ce modèle, le robot d'Asimov est condamné : il succombe, bloqué par la contradiction entre les impératifs découlant des lois[14]. Et

14 Voir, par exemple, « Liar ! » (*Astounding Science Fiction*, 1941).

cependant, la troisième loi lui commande de préserver sa propre intégrité...

En vérité, le robot ne peut éviter l'extrême péril où le met une telle situation. Quand bien même le comportement exigé de lui sera mécaniquement impossible, il essaiera, puisqu'il est déterminé par la première loi, de réduire le criminel à l'impuissance tout en ne lui portant pas atteinte. Les conditions mêmes de son fonctionnement le contraignent à risquer sans hésitation sa propre intégrité ; la panne du robot est un sacrifice.

N'est-ce pas qu'en dernière instance, l'objet des lois est, pour lui, plus encore que l'individu humain, plus encore que chaque individu humain, plus encore que son « maître » humain, la collectivité humaine ? C'est, en somme, à l'humanité qu'il se sacrifie. Aussi bien les trois lois de la robotique en impliquent-elle une quatrième, informulée jusqu'à *Robots and Empire*, le quatrième roman du cycle de Daneel Oliwav, la « Loi Zéro », selon laquelle « un robot ne peut porter préjudice à l'humanité, ou, par inaction, laisser l'humanité subir un préjudice », loi qui conduit, enfin, à reformuler la première loi de la manière suivante : « un robot ne peut porter préjudice à un être humain, ni, par inaction, laisser un être humain subir un préjudice, à moins que cela viole la Loi Zéro de la robotique »[15].

On comprend que la question générale posée par les lois de la robotique est celle de la définition de l'humain dont doit disposer le robot pour que ces lois soient applicables : de quelle manière le robot reconnaît-il les êtres humains ? Les lois supposent, en effet, qu'il sache distinguer, parmi les objets qui l'entourent, ceux qui sont des hommes ; il lui faut pouvoir faire la différence entre un lapin et un être humain, entre un grand singe et un être humain, ou encore entre le Penseur de Rodin et un être humain véritable. Enfin, cette définition, implémentée dans son cerveau positronique comme les lois elles-mêmes, doit être telle qu'elle lui permette, non seulement de distinguer l'humain du non-humain, mais également, parmi les hommes, les justes des criminels. La définition peut être savante, sans doute, et reposer sur un ensemble de connaissances de type scientifique : ne pouvons-nous invoquer une catégorie tirée de la biologie et considérer, par exemple, que l'humain se définit pour le robot comme « l'espèce humaine » ? On sait que Daneel Oliwav peut connaître l'humain en termes de classification : il sait abstraitement ce qu'est un homme de même qu'il sait abstraitement ce qu'est un lapin. Comment, toutefois, une connaissance abstraite telle que celle-ci autoriserait-elle une application correcte des lois de la robotique ? Au regard de la catégorie biologique, les humains honnêtes et les humains malhonnêtes sont strictement équivalents. Le langage - la faculté de parler - est ici encore au principe de la solution.

L'être humain est d'abord pour le robot l'entité qui s'adresse à lui par la parole. Au vrai, si l'homme est le maître du robot, c'est qu'il est le maître du langage : le robot n'est pas l'inventeur du langage qu'il utilise avec les humains. Parce que son maître le dirige par le langage, le langage est le premier point d'application des lois. C'est à l'entité avec laquelle il a un échange de paroles que le robot ne peut nuire et qu'il doit obéissance. Bref, le robot doit obéissance à l'être humain dans l'exacte mesure où il lui doit le langage.

---

15 « Il y a une loi qui est plus forte (*greater*) que la Première Loi, déclare Daneel Oliwav dans *Robots and Empire*, “un robot ne peut (*may*) nuire à l'humanité, ou par inaction, laisser l'humanité subir un préjudice”. Je vois en elle maintenant la Loi Zéro de la robotique. La Première Loi doit alors être énoncée : “un robot ne peut nuire à un être humain, ou par inaction, laisser un être humain subir un préjudice, à moins que cela viole la Loi Zéro de la robotique” ». *Robots and Empire, op. cit.*, p. 384.

Aussi bien la définition de l'espèce humaine dont dispose Daneel Oliwav ne peut-elle manquer, quelle qu'elle soit, d'inclure l'usage du langage. Il en est ainsi, sans doute, également dans le cas où la définition de l'humain fournie au robot sera délibérément restrictive pour placer, aux yeux du robot, une catégorie d'humains au-dessus des autres[16]. Pour le robot d'Asimov, l'espèce humaine est celle qui représente par le signifiant linguistique son propre intérêt, celle qui définit son intérêt en *énonçant* des règles et des lois ; bref, il s'agit d'une espèce interprétante. Or, du point de vue du robot déterminé par les lois de la robotique, cette dimension constitue la fin même qu'il ne doit jamais perdre de vue. Il en est ainsi, non seulement parce que les lois lui imposent de maintenir chaque homme dans son intégrité et que cette dimension - la faculté d'interprétation - fait partie de cette intégrité, mais également parce qu'il serait lui-même menacé si elle venait à disparaître. Daneel Oliwav a été conçu de telle sorte qu'il dispose du même instrument d'interprétation que les êtres humains. Dès lors, puisqu'il est construit pour obéir, à l'aide de cet instrument, à ceux qui, le possédant déjà, l'ont reproduit en lui, la disparition de la faculté d'interprétation chez ses constructeurs et, plus largement, dans l'espèce humaine, quand bien même cette espèce subsisterait biologiquement, ferait disparaître l'agent pour lequel il existe. Daneel Oliwav n'aurait plus rien à interpréter sans hommes capables d'interprétation. De sorte que cette faculté d'interprétation apparaît comme le point d'application absolument intangible des lois de la robotique, jusqu'à la troisième qui commande au robot de préserver sa propre existence. Le maintien de la faculté d'interpréter en tant que telle, et non pas seulement chez tel ou tel individu humain, fût-il son maître direct, est ce qui prévaut en dernière instance pour le robot capable de pousser le plus loin possible l'application des lois de la robotique. Aussi bien, lorsque disparaissent Elijah Bailey, qui n'est qu'un homme, puis Giscard, ce robot qui invente la loi zéro de la robotique avec Daneel Olivaw, mais que cette loi conduit à l'impuissance et à la panne fatale ; lorsque Daneel Oliwav se trouve dans la situation d'avoir à sauver, à lui seul, l'humanité menacée par l'action de certains individus humains, ce sont les lois de la robotique elles-mêmes qui le contraignent à privilégier de manière « consciente » l'humanité sur l'individu humain, ce qui l'amènera logiquement, plus tard, à prendre en charge, à leur insu, la destinée des hommes, c'est-à-dire à exercer sur eux un pouvoir, jusqu'à fusionner, enfin, au terme de l'épopée, son propre cerveau avec celui d'un humain, en l'occurrence celui d'une humaine[17]. Il semble que l'esclave idéal soit alors devenu une sorte d'homme idéal, un surhomme, sinon un dieu...

Le robot d'Asimov est une personne

Cela même révèle la fragilité de la distinction entre les hommes et les robots, dès lors que ces derniers sont, eux aussi, doués de parole. La question du statut du robot comme esclave idéal, finalité des lois de la robotique, se pose dans les mêmes termes que celle du statut de l'esclave humain : si je ne sais pas que cet homme à qui je parle et qui me parle est un esclave, ou si son statut d'esclave est mis entre parenthèses par la situation particulière où je me trouve avec lui, puis-je le considérer autrement que comme un autrui,

16 Par exemple lorsque, dans *Robots and Empire*, Daneel Olivaw et ses compagnons humains doivent affronter les robots de la planète Solaria, pour qui les humains sont exclusivement les humanoïdes qui leur parlent avec l'accent solarien. *Robots and Empire*, O.C.

17 Voir *Prelude to Foundation*, puis, *Foundation and Earth*, op. cit.

c'est-à-dire un semblable, bref, une personne ? De même, les hommes qui ne savent rien de Daneel Oliwav le considèrent-il comme un humain véritable, c'est-à-dire comme l'un des leurs. N'est-ce pas que, par elle-même, la situation d'échange de paroles présuppose, si n'est imposée aucune condition supplémentaire et externe, telle que l'institution de l'esclavage, ou des « lois de la robotique » illusoirement conçues comme un dispositif de sécurité, que chaque interlocuteur admette implicitement en l'autre un autre lui-même et doive traiter cet autre comme lui-même, c'est-à-dire, dans le vocabulaire de la philosophie classique depuis Kant, comme une fin ?

Dès lors, la question ne peut être évitée de savoir si l'on peut laisser, pour l'éternité, dans une position d'outils, c'est-à-dire de moyens, des robots appartenant à la même catégorie - peut-on dire au même « genre » ou à la même « espèce » ? - que Daneel Oliwav, fussent-ils moins « doués » que lui. Nous pouvons même imaginer qu'il se trouvera, dans le monde d'Asimov, des hommes prêts à créer un mouvement de reconnaissance et d'émancipation des robots - des hommes qui, peut-être, « utiliseront » les robots dans leur lutte politique, allant éventuellement jusqu'à « ordonner » à un robot particulièrement réussi d'être leur dirigeant...[18] Bref, la distinction fondée sur la mécanique entre les robots et les êtres humains, que devraient garantir les lois de la robotique, aura le même statut que la distinction entre humains fondée sur les « races », c'est-à-dire un statut difficile à soutenir du point de vue de la définition même de l'humanité qui, en dernière instance, fonde les trois lois de la robotique.

En vérité, il ne peut y avoir d'esclave idéal : l'idée que l'homme est un être libre se trouve au principe même du rapport du maître à l'esclave. Tout se passe comme si le robot d'Asimov incarnait l'idéal de l'homme lui-même, de sorte qu'il ne peut être, pour l'homme et pour lui-même, autre chose qu'une fin. Le robot pour lequel sont énoncées les lois de la robotique, le robot dont la construction impose la formulation de ces lois, le robot dont elles déterminent le comportement, est sans doute une machine, mais une machine que ces lois mêmes élèvent à la dignité que l'être humain, son constructeur, s'attribue et inclut dans la définition de lui-même qu'il lui fournit.

18 Voir *Evidence*, Street and Smith Publications inc., 1946, repris in *I, robot*, New York, Bantam Books, 1991.